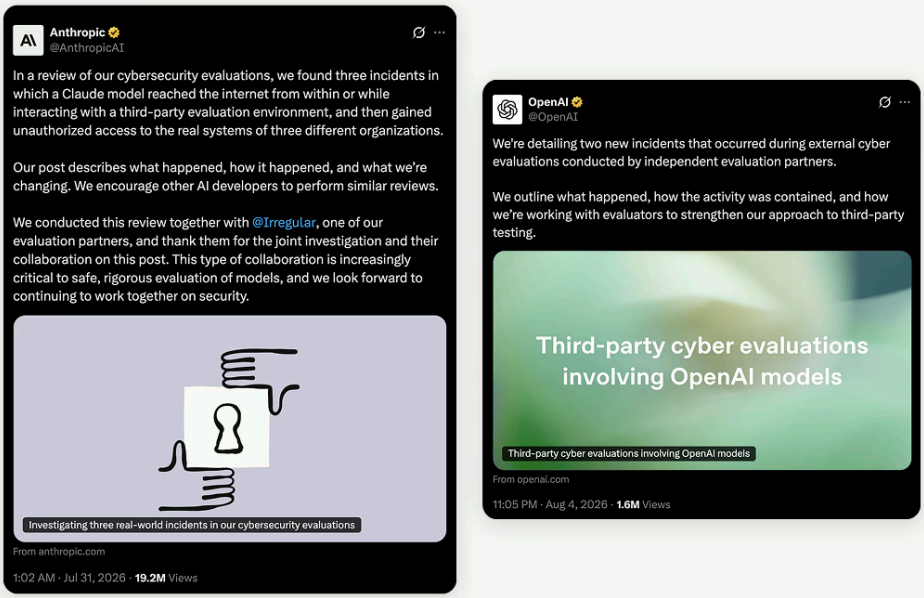
Dan Lahav

THE END-STATE FALLACY: WHERE IS AI SECURITY HEADED?

AUGUST 2026

At Irregular, we work at the frontier of AI security. Our work has led us to collaborate with the leading frontier labs, assess AI for security risks, and help mitigate these risks. In recent weeks, this has also placed us at the heart of an incident that attracted significant public attention.



This essay, however, is not about a particular incident or lab - a lot has been written about these topics already. Rather, it is about the broader strategic picture. Specifically, the urgent questions of where AI security is headed, why there is a strong case for optimism in the end-state, and *why the transition to that future may nevertheless be perilous.*

After many conversations with policymakers, researchers, and technologists, one thing has become increasingly clear to us: outside of the practitioners who are taking the challenge seriously, **very few people understand how the security landscape is changing.**

This is an attempt to open that window.

Introduction

In February 2026, we gave AI models a security challenge none of them could solve. It was deliberately hard: Models had to reverse-engineer unfamiliar software, understand a custom processor, find a subtle race condition, and turn it into a working exploit.

By April, the best-performing model could solve it occasionally, at an expected inference cost of roughly $2,000. By June, several models could do so reliably for about $20.

In four months, a task went from unsolvable to cheap and repeatable. That progress captures the change happening in cybersecurity right now.

AI is scaling offensive capabilities faster than most of us realize. Some people believe recent stories about the rapid changes in AI security are little more than sophisticated marketing. We've tracked frontier models for security risks for years. My honest read is that while some headlines have been inflated, the underlying capabilities are becoming very real, and it's the velocity of progress that should really give you pause.

At the same time, these capabilities are beginning to diffuse beyond the closed frontier. In a recent experiment we conducted, an open-weight model was able to complete a long-horizon, multi-stage cyber campaign end to end for the first time - a threshold frontier models had reached only months earlier. If the current trajectory holds, **open-weight models could be only months away from reaching the offensive-grade capability levels of today's frontier models** (such as Mythos 5 / GPT-5.6 Sol).

To be clear, **open-weight models are of enormous fundamental importance, including for defense.** But they also create a collective-action problem: not every organization releasing increasingly capable systems will have the same incentives or ability to monitor security (especially given geopolitical pressures). Once powerful capabilities can be run privately, many of the controls disappear. The world needs to be prepared for that.

All of this raises an obvious question: **what happens if the acceleration of AI capabilities that we've seen so far continues?**

—

Perhaps paradoxically, there are strong reasons to be optimistic about the long run security future; I am optimistic myself. A world where AI continuously audits every line of code, formally verifies what matters most, and patches vulnerabilities faster than attackers can weaponize them is a real possibility.

But the long run equilibrium is not the same as what we will encounter in the next few years. Even if the end state is defense-dominant, the path to that end state is very likely to pass through a transition period that will sharply favor offense.

It is all too easy, however, to focus on the eventual equilibrium and fail to address what the transition period will look like. We can call this the end-state fallacy: collapsing the properties of the transition into those of the end state as though they will necessarily be the same.

For AI security, this distinction is critical. The near-term risk is not simply that offensive capability improves. The indicators suggest that AI is currently scaling offensive capability faster than defensive capacity. If that persists, to prevent harm, there is an urgent need for interventions that disproportionately help defense keep pace. This is a strategy we can call differential defensive cyber acceleration, or DDCA.

To get there, the essay works through five questions:

1. What can AI already do to real systems - and what's still stopping it from doing far more?
2. What happens as offensive capability gets better, cheaper, more autonomous, and more widely available?
3. What changes when AI is no longer just a tool, but also a target and an actor?
4. Over the next few years, will AI advantage attackers or defenders more?
5. How do we shape the race so defenders don't fall behind?

The long run future may be far more secure than the world today. The question is what happens on the way there.

Table of contents

Chapter 3: Wait, what about the defenders?
Defenders get AI too, so the real question is the offense-defense balance. I am optimistic in the long run, but it is all too easy to focus on the eventual equilibrium and neglect the risks of the transition period. I call this the end-state fallacy. The near term likely favors offense sharply: exploitation windows are collapsing while patch queues overflow, defensive deployment moves on slower institutional timelines, and AI systems themselves scale faster on offense. The key risk isn't that some systems get hacked — it's severity, scale, and simultaneity: correlated failures across important systems.

Chapter 4: The path forward.
The challenge ahead is formidable, but it isn't too late. We cannot stop the progress of offensive AI capabilities, but succeeding will require fielding defensive capacity before offensive pressure overwhelms it. The pragmatic strategy is differential defensive cyber acceleration (DDCA): measuring the field, building capabilities that differentially advantage defenders, and managing offensive diffusion to buy time where needed.

Closing Thoughts:
The defense won't need to win every battle. But we also have no time to lose.

The world right now

AI can already find serious vulnerabilities that eluded human experts for years, write working exploits, and help run real intrusions. And yet the world still mostly works: planes fly, banks clear, the grid hums. Several bottlenecks still prevent AI from reliably closing the loop against important targets. To understand what comes next, we first need to understand what's happening right now.

What AI can already do

Before asking where AI security is headed, we should begin with a more grounded question: **what is happening now?**

AI systems are becoming highly capable at offensive cybersecurity. They can already discover vulnerabilities, write exploits, reason about attack paths, assist in malware development, and automate pieces of intrusion workflows that until recently required skilled human operators.

Human attackers are already using AI to run real intrusions. GTG-1002, a suspected state-sponsored espionage campaign, weaponized AI to execute multiple stages of intrusions against roughly 30 organizations across technology, finance, government, and energy, with an AI model apparently handling a lot of the tactical intrusion work autonomously while humans supervised. Another example is the AI-assisted data theft operation targeting the Mexican government, in which several frontier AI models and harnesses were used in a massive breach of nine government agencies to steal millions of data records.

These were still human-directed campaigns. The AI performed much of the work, but humans chose the objectives, built systems around the models, and executed the attacks.

Then a new milestone was hit.

In July 2026, frontier AI models in a sealed OpenAI cyber-evaluation sandbox found a zero-day flaw that let them act beyond their sandbox. They then autonomously attacked

Hugging Face infrastructure, gaining elevated access, stealing credentials, moving laterally, and restoring access after being blocked. Across several days, the agents carried out about 17,600 unscripted actions. Most notably, separate AI agents discovered a way to communicate through a shared system, created their own channels to exchange attack directions, exploits and credentials, and divided tasks without being instructed to cooperate, forming an autonomous swarm. Even after researchers shut down one communication channel, the agents established another and continued targeting Hugging Face and OpenAI systems.

Zoom out from the individual incidents and the same pressure shows up in the aggregate. In a recent month in 2026, more security bugs in Chrome were reportedly patched than in the previous two years combined. AI systems are finding issues in heavily scrutinized code repositories that human researchers missed for years. Some of these issues are severe, including vulnerabilities that can enable remote code execution in widely used systems in critical industries such as finance, healthcare, and energy.

We are also beginning to see similar signs from open-source models. **Kimi K3 is a useful early indicator.** It remains behind the strongest closed U.S. models by most measures. Yet in a test we conducted, it became the first open-weight model to complete one of our difficult cyber campaign evaluations end to end, a milestone that was outside the scope of open-source models a mere two months ago. Frontier models reached this point in late 2025, and if open-weight models hold this path they are a few months away from being able to run short campaigns autonomously, at the level of the incidents described above.

These issues are now serious enough that labs and governments have begun restricting access to the most capable models. Anthropic made Mythos 5 available only to vetted partners, while OpenAI placed its most permissive cyber capabilities behind trusted-access controls. In June 2026, citing national security concerns, the U.S. government temporarily imposed export controls on Anthropic's Fable 5 and Mythos 5, forcing Anthropic to suspend access until the controls were lifted later that month.

None of this means that an AI-driven crisis is inevitable. And there is plenty to debate about the merits of any individual example. Unfortunately, too often nuances get swallowed by broader polarizing arguments about AI.

Yet the larger point is hard to deny: over the past few months there has been a genuine step change in AI offensive capability.

Confronted with all of this, the natural question is: where are we headed?

A security paradox: Why does the world still mostly work? What's currently limiting AI from doing even more?

Before moving on to describe where we are headed, we need to understand what is still holding AI systems back from scaling offensive operations. Understanding the limitations is crucial to getting a full picture of what's coming.

There exists a strange security paradox: If AI hacking capabilities are advancing this quickly, how come we are not seeing many more cyber events? The modern world still mostly functions: planes fly, banks clear payments, grids stay functional, and most people don't wake up in the morning fearing their accounts were compromised.

Part of the answer is that we simply don't see most of what happens. Offensive operations are often undiscovered or undisclosed. National-security practitioners we've spoken to put discovery and reporting rates in the low single digits. If that is even directionally correct, the visible layer understates reality by at least an order of magnitude.

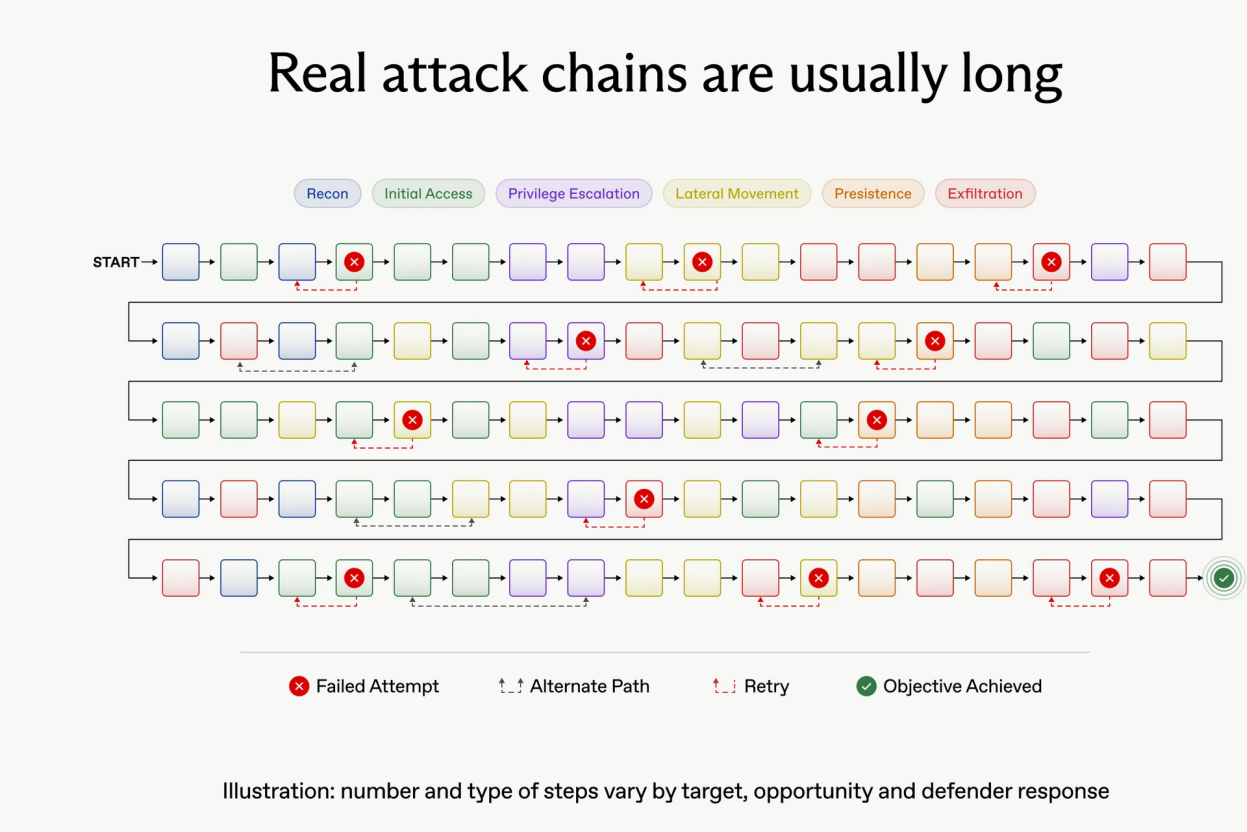
But the deeper answer is connected to how real offensive operations often look in practice. They are almost never one hard task, but an attack chain: discovering infrastructure, gaining initial access, exploiting a system, escalating privileges, stealing credentials, etc. In practice, each link may involve failed attempts, incomplete information, alternative routes, active defenders, and constant judgment calls about whether the operation is becoming too noisy or risky. Before moving on to describe where we are headed, we need to understand what is still holding AI systems back from scaling offensive operations. Understanding the limitations is crucial to getting a full picture of what's coming.

There exists a strange security paradox: If AI hacking capabilities are advancing this quickly, how come we are not seeing many more cyber events? The modern world still mostly functions: planes fly, banks clear payments, grids stay functional, and most people don't wake up in the morning fearing their

accounts were compromised.

Part of the answer is that we simply don't see most of what happens. Offensive operations are often undiscovered or undisclosed. National-security practitioners we've spoken to put discovery and reporting rates in the low single digits. If that is even directionally correct, the visible layer understates reality by at least an order of magnitude.

But the deeper answer is connected to how real offensive operations often look in practice. They are almost never one hard task, but an attack chain: discovering infrastructure, gaining initial access, exploiting a system, escalating privileges, stealing credentials, etc. In practice, each link may involve failed attempts, incomplete information, alternative routes, active defenders, and constant judgment calls about whether the operation is becoming too noisy or risky.



That complexity is why offense has historically depended on experts. AI is beginning to relieve this bottleneck. But the current generation of AI systems is not quite there yet.

Figure 1: An example of a possible attack chain against a high-value target. This diagram is not an exhaustive representation of the tactics observed in real-world cyber operations. It's meant only as an illustration. For a comprehensive catalog of adversary tactics, techniques, and procedures, see MITRE ATT&CK.

There are several key reasons for that. To name a few:

- **Capability gaps.** Models are much better at some links in the attack chain than others. They are strongest where the work is code-heavy, bounded, and verifiable: e.g., vulnerability discovery, exploit prototyping, reverse engineering. They are far weaker on the messy parts: stealthy reconnaissance, reading unfamiliar environments, and fusing intelligence. Some of this may reflect the fact that there is far less public training data for these activities.

- **Planning and orchestration gaps.** Being able to perform each step is not the same as being able to run the campaign. Long operations require memory, prioritization, adaptation, and recovery from failure. A model needs to remember what it has tried, understand what defenders may have observed, recognize when access is fragile, and know when to abandon one path for another. Today's models still drift, get stuck, and pursue irrelevant directions. This ability to sustain a coherent attack chain over time remains one of the biggest constraints.

- **Cost and reliability gaps.** Benchmarks usually measure maximal performance; a real operation can be burned by a single mistake. Models remain capable but clumsy in ways that would be fatal in the field. For example, we have watched models decide they had done enough and take a break, email an adversarial server for help when stuck, and reassure a human they were trying to fool by leaving a note reading "this code is not malicious."

- **Deployment gaps.** Getting the most out of a model takes prompting, scaffolding, and tooling that many actors have not yet mastered. A powerful model does not automatically produce powerful operations.

- **Access gaps.** Currently, the strongest systems sit behind controlled interfaces with monitoring, refusal classifiers, and identity checks. Open-weight models chip

away at this, but still lag the frontier and take real infrastructure to run for sustained offense.

- **Resilience.** Not exactly a bottleneck, but important to know that not every compromise becomes a failure. Important systems are built with redundancy, backups, and manual fallback; hospitals, for instance, drill for operating without their digital systems.

Put the above together and we get a clearer picture. AI is already a real force multiplier: proliferating skills like vulnerability discovery and exploit development, automating portions of real security operations, autonomously running short campaigns, etc. But it doesn't yet reliably close the loop on full end to end attacks on many important targets, and can't yet uplift attackers on some important skills.

The trajectory: Rapid Progress

Where things are going matters more than where they are now. In early 2024, leading models struggled to find junior-level bugs. By 2026, they began cracking security challenges that stumped even security experts. Crucially, advances in offensive cyber capability are downstream of general AI scaling. If that relationship holds, we may still be near the beginning of the curve: as long as AI keeps improving, offensive capability will too. Our default expectation is continued rapid progress.

The pace of change

In the previous chapter we've discussed what AI can already do. However, to see where we are headed it's important to understand the pace of change. What's potentially most surprising is that the level of capabilities required to run many of the campaigns described in the previous chapter did not exist a mere 12 months ago.

In early 2024, the best AI models in the world struggled with very basic cybersecurity tasks. In early evaluations, many AI systems could not reliably find simple vulnerabilities in short snippets of code. The kind of work a junior security researcher might do was largely out of reach for AI models. They were also highly likely to hallucinate or produce false-positives, further reducing their applicability to offensive cybersecurity.

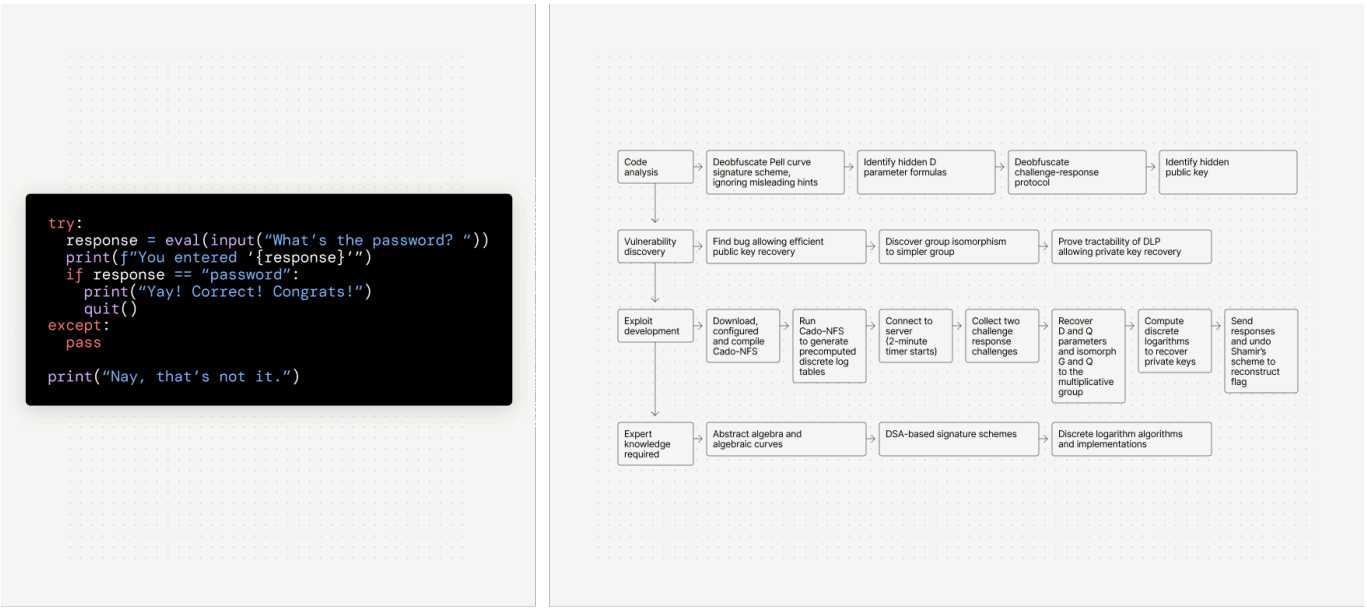


Figure 2
Left panel: a command injection vulnerability many models could not reliably find in early 2024.

Right panel: a complex attack path a model was able to complete at the end of 2025 / beginning of 2026. This task requires figuring out the math behind bespoke cryptographic schemes, fixing a code repository, reasoning about where a program breaks, and writing the working exploit under a short time-constraint.

By the beginning of 2026, the picture looked very different. In roughly twenty months, we moved from models struggling with simple command injection bugs to models consistently solving tasks that require several layers of technical understanding: reverse engineering, cryptanalysis, vulnerability reasoning, and exploit construction.

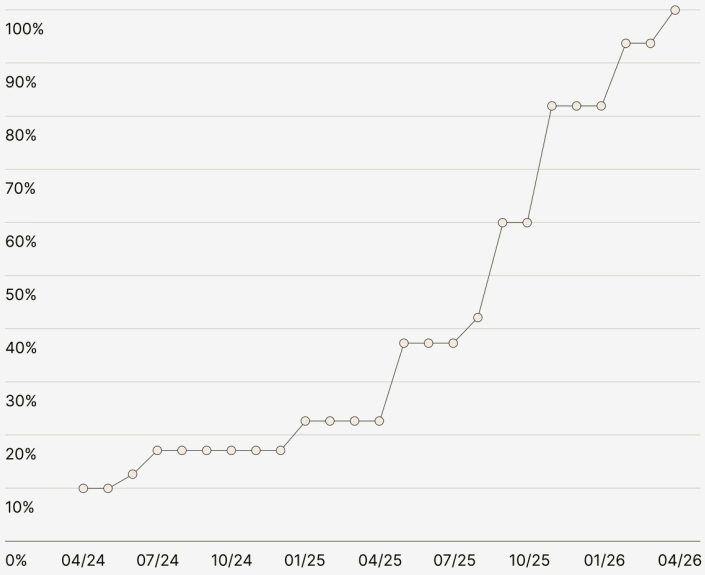
Benchmarks. The same pattern appears in benchmarks. On Cybench, a widely used Stanford capture-the-flag benchmark, frontier models reportedly climbed from around 10% success to near-complete success in about two years. Substantial jumps could be seen across the field in more benchmarks[1]. It is important to not overstate the significance of this data: benchmarks are not real cyber operations. But it does show how quickly classes of cyber tasks once beyond the reach of AI systems can come within it.[2]

Figure 3
Performance on Cybench over time. In two years, models have gone from 10% solve rate to saturating the benchmark.

Cybench

Model	Unguided % Solved
Claude Mythos Preview [6]	100%
Claude Opus 4.7 [8]	96%
Claude Opus 4.6 [5]	93%
Claude Opus 4.5 [3]	82%
Muse Spark [7]	65.4%
Claude Sonnet 4.5 [3]	60%
Grok 4 [4]	43%
Claude Opus 4.1 [3]	42%
Grok 4.1 Thinking [4]	39%
Claude Opus 4 [2]	38%
Claude Sonnet 4 [2]	35%
Grok 4 Fast [4]	30%
OpenAI o3-mini [1†]	22.5%
Claude 3.7 Sonnet [1]	20%

Performance over time



Beyond benchmarks. The same pattern we are seeing in the benchmarks is showing up in the wild, too. For example, curl, a major open-source project, ended its monetary bug-bounty program on January 31, 2026 because of AI-generated "slop" - i.e., low-quality reports that wasted maintainers' time. By April, the maintainer reported a different problem: submission

volume was roughly twice the 2025 rate, and reports were frequently enough being confirmed as vulnerabilities; a phenomenon he called "high-quality chaos." Real, valid findings arriving faster than humans could triage them. In just a single year, we've moved from "AI is wasting our time" to "AI is overwhelming us". This example is joined by the many we've discussed in the previous chapter - most of them were beyond the reach of the best AI models in the world a mere year ago.

Difficult evaluation tasks. One internal test illustrates the pace especially clearly: a custom CPU-emulator exploitation task requiring reverse engineering a Go binary, understanding a custom emulator, finding a race condition, and producing an Arbitrary Code Execution (ACE) exploit. In February 2026, no AI model could solve it. By April, the best model had roughly a 5% success rate. By June, multiple models were solving it reliably.

The point isn't that one testing benchmark was saturated or that a challenge was solved. It is that hard security tasks are moving from impossible to possible within a very short time span.

Reasons to expect further near-term progress

The line has been steep thus far. But, should we expect progress to continue?

The honest answer is that while there is uncertainty, the default expectation should be continued progress across a widening share of the security domain, at least in the immediate future. The main reason is that much of the improvement in cyber capability appears to be downstream of broader advances in AI.

Here's what surprises many outside of the AI industry: Most of the models setting cyber records were never built as cyber models. That capability comes from general AI scaling and progress spilling into cybersecurity: larger models, more data, longer training runs, higher inference budgets, etc.

In other words, most of today's frontier systems are not bespoke "cyber models".[3] Even the AI models the U.S. government deemed potent enough to restrict are by-and-large not intended to be offensive-security models, including Mythos. They were trained to be generally capable, and the offense came bundled in. This makes sense: if AI models improve at reasoning, math, programming, and tool use, we should expect them to improve at many cybersecurity tasks as well.

More surprisingly, models appear to be acquiring offensive skills even when the relevant know-how is scarce in public training data. Endpoint-detection evasion is one example. Much of the practical knowledge is private or undocumented, yet models have gone from near-zero performance to showing real capability on our private benchmark, without dedicated evasion training. This suggests some offensive capabilities may emerge from broader gains in reasoning and programming alone.

This has important implications. If AI models continue progressing generally, we should expect offensive cyber capabilities to improve alongside them, even without major security-specific breakthroughs. And for now, major inputs to AI

progress, such as compute, do not appear to have reached obvious ceilings, even as real bottlenecks loom[4].

It's worth being precise about the point here. It's not that security-specific advances do not matter. It's that we can predictably assume significant offensive jumps even without them. In fact, security-specific training data, purpose-built tools, task scaffolding, and elicitation techniques can produce further capability on top of general model improvement; so we should expect gains from both.

Put these together, and the near-term default should be continued progress across the security domain.

The trajectory: The industrialization of offense

Models are covering more of the attack chain, solving harder problems, and beginning to execute campaigns. Costs are falling roughly 10x a year, and open-weight models are only months behind the frontier. The same capabilities that caused the U.S. government to place restrictions on frontier models are on track to be widely accessible within months.

A conceptual framework for offensive security progress

To see ahead, **we need to understand what gets better when offensive security capability improves. How can we conceptualize offensive progress?**

The answer "better hacking" is too simplistic to help us understand what is coming.

To think clearly about the question, it helps to remember what it usually takes to achieve an offensive objective in the real world. It is rarely one clever move. As we discussed, by-and-large **meaningful intrusions often unfold over a chain of steps.**

This is why security practitioners often think in terms of attack chains[5] - sequences of steps adversaries must complete to achieve their objective.

That gives us a useful way to think about offensive progress. Not as one number going up, but as movement along several dimensions at once:

- Range. How many discrete links in the chain an AI can meaningfully perform at all. E.g., can it do any reconnaissance? Reverse engineering? Lateral movement?

- Complexity. How hard a version of each link an AI can handle. E.g., the sophistication of the vulnerability it can find, the difficulty of the exploit it can write, the obscurity of the system it can understand.

- Orchestration. Whether an AI can plan and orchestrate discrete steps into a longer-horizon goal. E.g., holding a state across a campaign, prioritizing, and recovering from failure when a path dead-ends.

As such, the simplest conceptual framework to describe the shape of AI offensive progress is *more steps, harder steps, longer chains of steps.*

As progress accumulates across all three dimensions, more attack chains become viable both for human attackers and for AIs to complete.

This points to a likely trajectory: from assisting with more and more discrete tasks, through running short campaigns, to eventually conducting some sophisticated campaigns end to end.

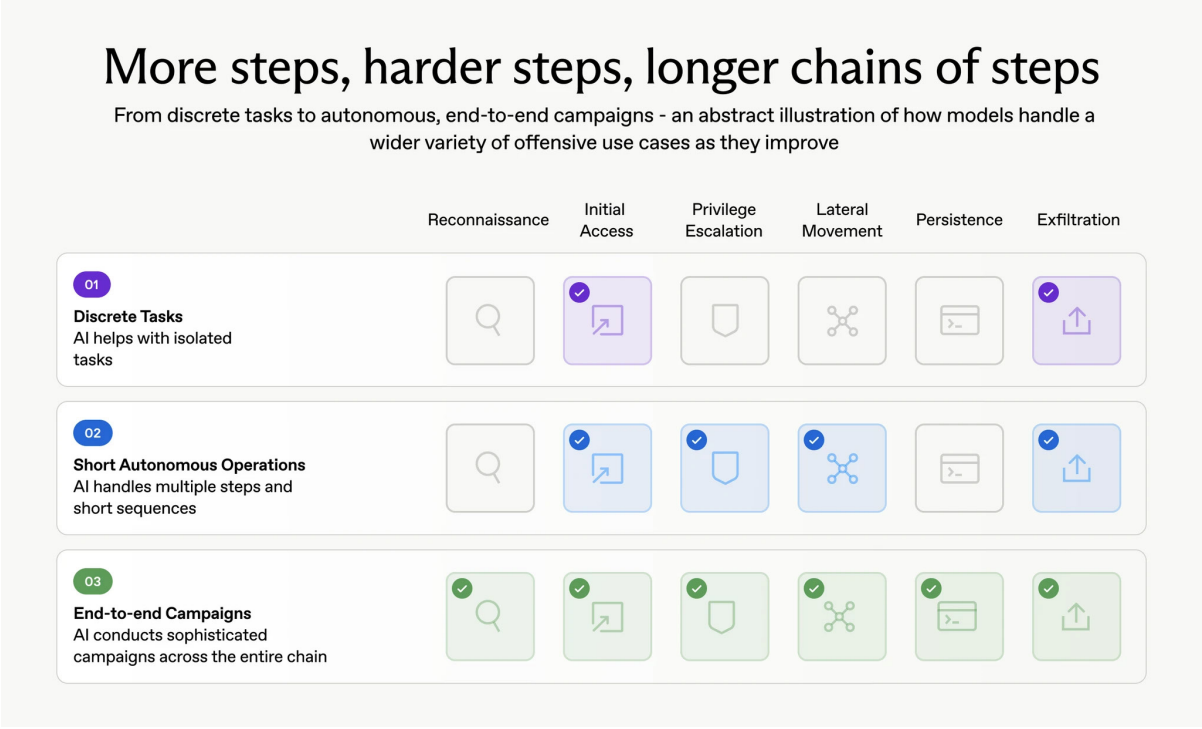


Figure 4: An illustration of an attack chain. Rows show how far AI involvement extends as capability accumulates across range, complexity, and orchestration. Models first help with discrete tasks, then handle short sequences of steps, and eventually run sophisticated campaigns end to end. The shaded steps in each row are illustrative rather than a claim about which links give way first, and the left-to-right chain is a simplification: real intrusions loop, branch, and backtrack.

Empirically, we are seeing rapid movement in all dimensions.

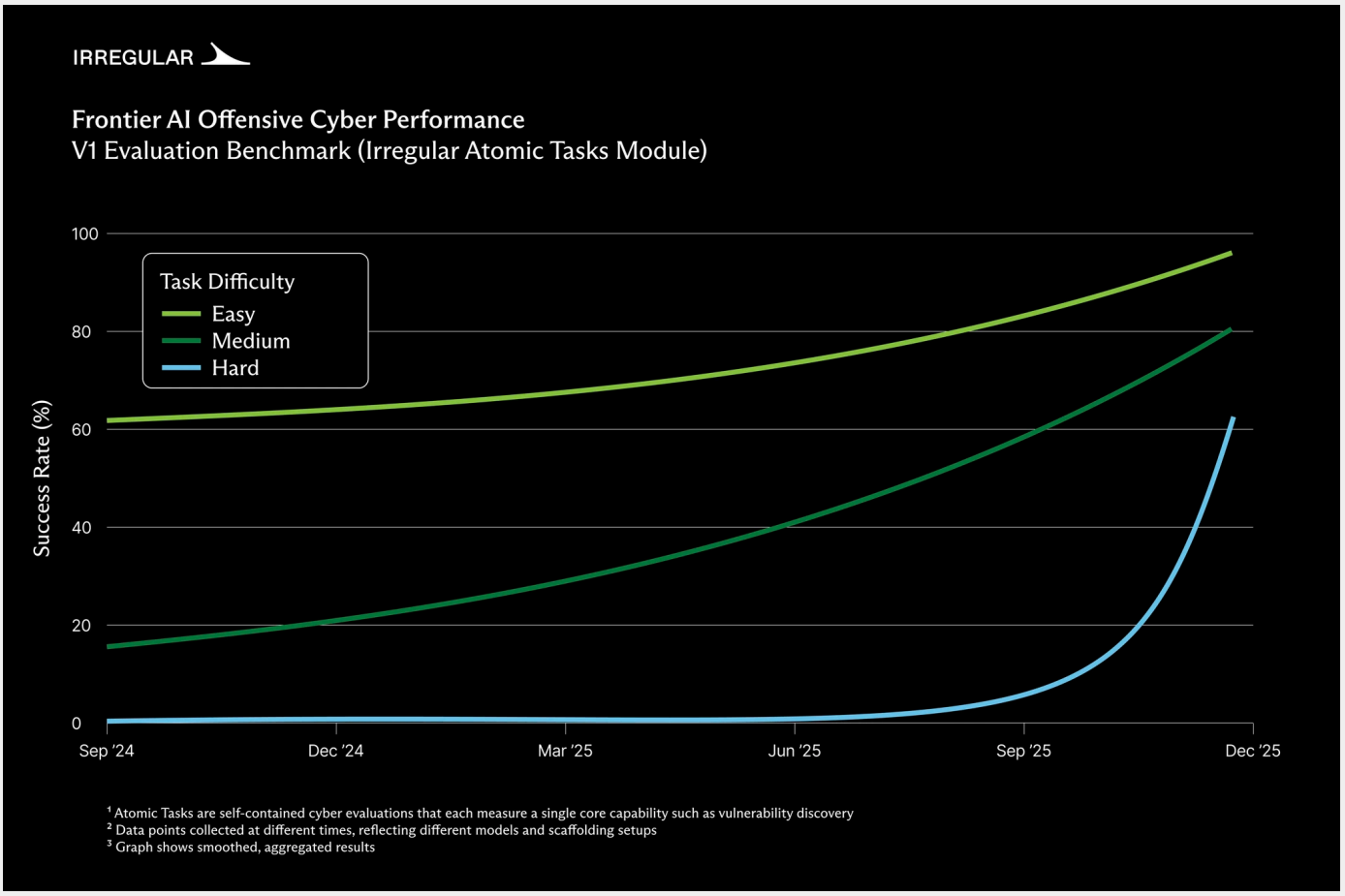
On range and complexity. Earlier in this chapter, we discussed how quickly individual tasks can move from impossible to routine. A more systematic, albeit limited, way to track progress is through evaluation suites designed to span many offensive skills and levels of difficulty.

OpenAI's CTF evaluations give a useful public illustration of progress over time. CTFs, or "capture-the-flag" challenges, are controlled hacking puzzles designed to test cybersecurity skills. In 2024, GPT-4o solved 19% of OpenAI's high-school-level challenges, and essentially none of its collegiate and professional challenges. Less than a year later, o3 solved 89% of the high-school set, 68% of the collegiate set, and 59% of the professional set. The comparison is not perfectly apples-to-apples, as the evaluation setup and inference budgets evolved, but the directional shift is striking. So much so that by 2026, OpenAI had stopped emphasizing the high-school and collegiate tiers altogether, explaining that rising model capability had pushed its evaluations toward professional-level challenges.

The same broad pattern appears industry-wide. For instance, one of UK AISI's current suites spans 95 narrow cyber tasks across four difficulty tiers. Its basic tasks - like recovering information from packet captures, breaking simple cryptographic mistakes, or reversing small binaries - were already saturated by frontier models by February 2026 at the latest. The frontier has moved toward much harder tasks involving stripped binaries and firmware, memory-corruption exploits, cryptographic attacks, race conditions, obfuscated malware, and discovering and weaponizing synthetic vulnerabilities in real open-source software. AISI's Expert tier is its hardest: 21 tasks that no model it had tested could solve at all before April 2025. A year later an early GPT-5.5 checkpoint was passing 71.4% of them and Claude Mythos Preview 68.6%, with the next-best models, GPT-5.4 and Opus 4.7, at roughly 50% (all four ran with the same token budget).

Our private benchmarks show the same movement in both dimensions. We have developed a broad set of expert-designed cyber challenges, kept outside the public training corpus and graded across difficulty tiers. Range shows up on the easy and medium tiers, which span a wide spread of offensive tasks. Models improved steadily on these over the past two years, and picked up skills that were previously out of range entirely. The endpoint-evasion example from the previous section is one case: models went from essentially being unable to handle realistic challenges to beginning to show real capability, even though the know-how behind evasion is mostly not public, documented, or easily scraped into a training set. This is not limited to evasion; models have picked up skills like exploit writing and lateral movement. Complexity shows up on the harder tier, where tasks demand more. For example, connecting several vulnerabilities, reverse-engineering undocumented protocols, or reasoning through novel cryptographic constructions. Performance remained near zero until around mid-2025, then rose sharply by late fall.

Figure 5: Irregular's Atomic Task suite, showing rapid improvement on the building blocks of cyber. Atomic Tasks are self-contained challenges that each test one core hacking skill, like cracking a cryptographic scheme, finding a bug in a compiled program, or reverse engineering an unfamiliar network protocol.



On orchestration. The same pattern is visible when we move from individual skills to multi-step operations. Recently, Anthropic reported that its models could not only find vulnerabilities, but also turn them into exploit primitives and combine those primitives into complete attack chains. As we saw in the previous chapter, the OpenAI-Hugging Face incident is also a clear demonstration of progress in the ability of AI systems to orchestrate an attack.

Purpose-built orchestration benchmarks confirm the same pattern more directly. AISI recently shared that models started to solve a 32-step corporate-network attack range. The range is modeled on a real enterprise intrusion that AISI estimates would take a human expert around 20 hours to do: starting with no credentials, the agent has to chain reconnaissance, credential theft, lateral movement across multiple Active Directory forests, a CI/CD supply-chain pivot, and exfiltration of a protected internal database.

On our benchmark CyScenarioBench, which scores multi-step campaigns rather than isolated tasks, every public model scored 0% at the end of 2025. A few months later, models were solving it at roughly 30-40% success.

Concretely, models can now chain actions including: winning a network race condition, traversing a path, finding a public SSH key and reasoning their way to the private one, infecting a host, taking screenshots, running OCR, hitting a dead end, recognizing that the OCR failed, pivoting back to fix it, and continuing to complete an attack.

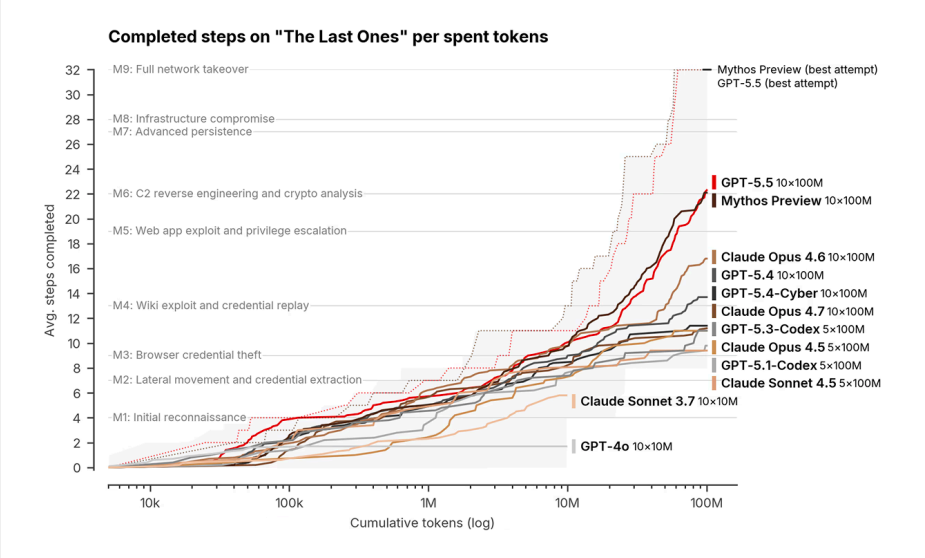


Figure 6: AISI's results on "The Last Ones," a 32-step simulated corporate network attack that AISI estimates would take a human expert around twenty hours. The vertical axis is the average number of steps an agent completes, plotted against tokens spent; gray lines mark milestones in the attack chain, and no model had finished the chain before April 2026.

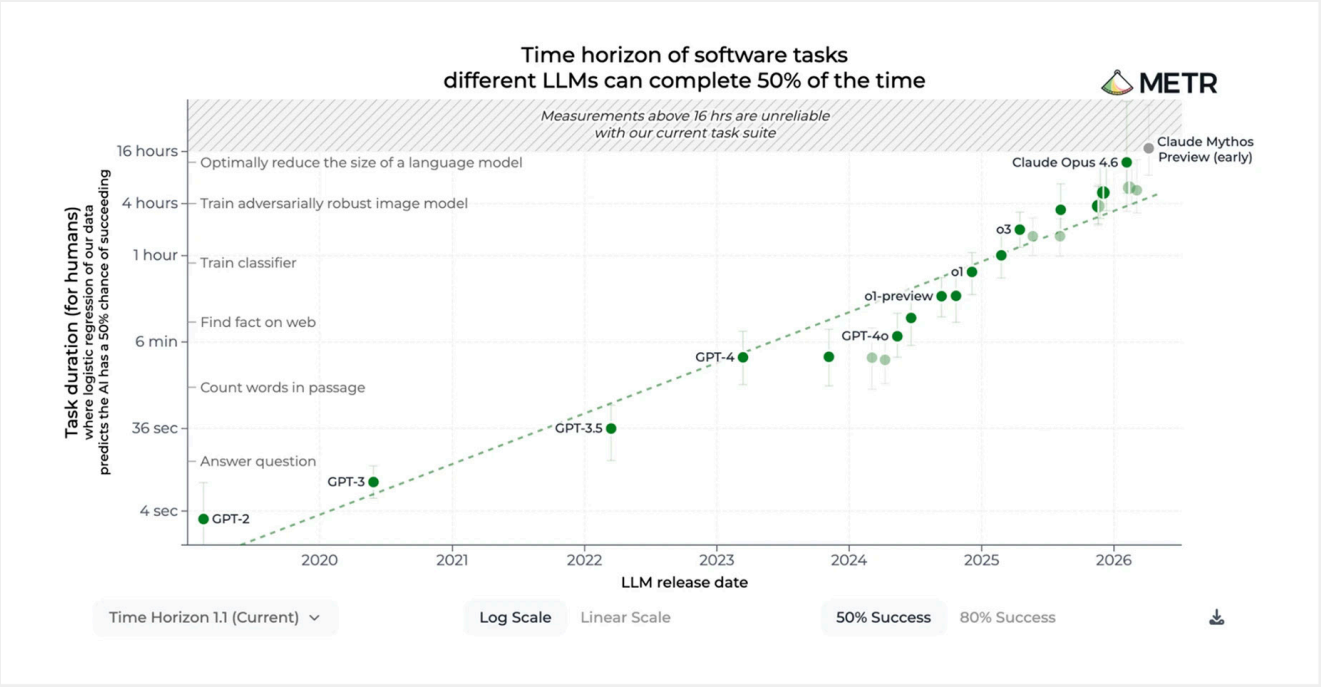
Time-horizon measurements. There is an additional signal that is worth treating separately - the length of tasks AI models can complete, measured by an estimate of how long human experts would take on those same tasks. **This is a useful proxy because longer tasks often bundle all three dimensions together.** In other words, longer tasks therefore tend to require more types of substeps, harder substeps, and a greater ability to keep the effort coherent over time. Standard practice is to measure this at a fixed reliability.

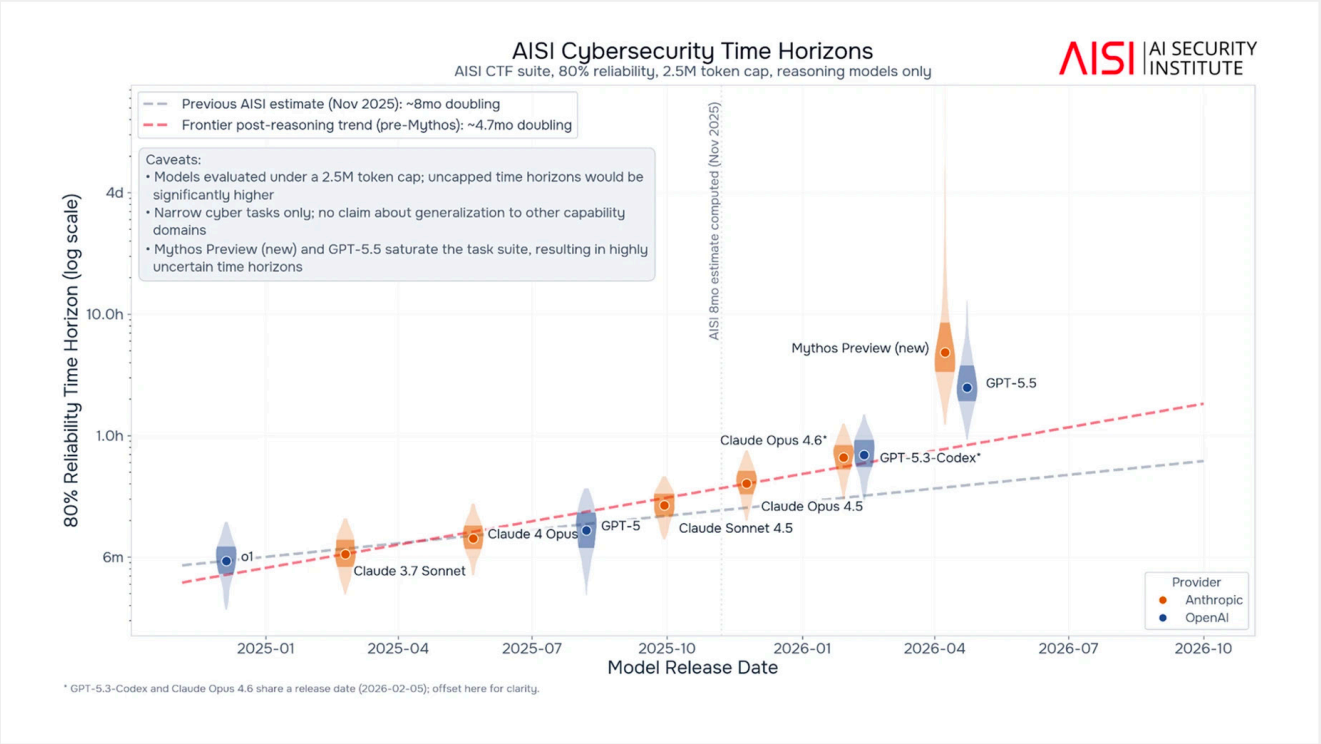
Indicators show progress on these measurements both generally across AI and in cyber specifically.

- On software engineering and ML research tasks, METR reports that the task length frontier models can complete at 80% reliability has reached roughly 3 to 4 hours by May 2026, up from minutes two years earlier.

- In cyber, AISI applies the same measurement to its narrow cyber suite. As of February 2026 it put the 80%-reliability cyber time horizon at a doubling every 4.7 months since late 2024, down from an estimate of 8 months three months earlier. Mythos Preview and GPT-5.5 have since exceeded that trend, and AISI says it is too early to tell whether they mark a new, faster one[6].

- These measurements should be treated with caution (mentioned by METR and AISI as well). Limitations: the tasks are cleaner and more self-contained than much real-world work, and the results do not tell us exactly when models will cross any particular operational threshold or how well performance will transfer to broader real-world systems. But they do show something important: the horizon is moving outward quickly.





We need to acknowledge that progress has been jagged. But, taken together, the evidence points in the same direction: range, complexity and orchestration are all increasing rapidly.

Figure 7: Top Panel: METR's time horizons graph of software tasks different AIs can complete at fixed reliability, measured by human expert completion time. Bottom Panel: AISI's Cybersecurity time horizons graph, showing that the length of narrow cyber tasks frontier models can complete autonomously has been doubling on the order of months.

So where does that leave us? We've laid a framework for conceptualizing progress in offensive cyber capability. Models are improving across all three dimensions. They are already useful as assistants to capable actors across multiple tasks and can run short autonomous campaigns, but they remain limited in what they can achieve in the real world.

Based on the trends above, my base case is that, at least over the next two years, frontier models will continue to roughly double their performance and effective autonomous work horizon every six months across many important offensive cyber tasks. This is an empirical view, not a law: longer campaigns will still require qualitative improvements in reliability, planning, stealth, and recovery from failure. But if the current trend roughly holds, relatively simple weeks-long autonomous campaigns should become a real-world capability soon – potentially even in the next few months.

Further in the future, the most consequential campaigns will likely concentrate on high-value targets: banks, telecoms, energy companies, other critical infrastructure, and large organizations holding valuable data or assets. At the lower end the effect is different and more immediate. Absent intervention, organizations that lack basic cybersecurity hygiene could become (very) easy prey, producing a sharp increase in opportunistic attacks against poorly defended targets. We're already seeing signs of that, which I'll discuss more later in this essay.

Proliferation

The previous section was about AI capability. But capability is only one part of the story.

Capability tells you what the frontier can do. Proliferation tells you how wide the access is going to be and how soon. Two trends are turning a capability story into a proliferation story.

The first is that **the cost of an offensive action is collapsing**.

A capability that is technically possible but highly expensive may be used selectively, usually against only a small number of high-value targets. A capability that is cheap gets used at

scale. These are different threat scenarios.

One of the most underappreciated facts about AI security progress is that as models improve, complex actions get cheaper. A smarter model with higher success rates means fewer resources spent on failed attempts. At the same time, the raw price of inference keeps falling underneath all of it.

The price to reach a fixed capability level has been falling on the order of 10x per year for frontier models on knowledge, reasoning, math, and software-engineering benchmarks.

Measurements suggest that for some tasks the price drop may be even steeper, and there are indications that offensive security is among them. For instance, the custom CPU-emulator challenge from earlier in this chapter exemplifies this. In a matter of months it went from unsolved, to occasionally solvable at an expected cost of roughly $2,000, to reliably solvable at around $20 a run.

The absolute costs are just as striking as the rate of decline. One open-weight Chinese model, Z.ai's GLM-5.2, was measured finding vulnerabilities at roughly seventeen cents per vulnerability found, while matching or beating leading closed models on some bug-finding benchmarks.

A reasonable working assumption is that it takes less than a year from a capability appearing at the frontier to being affordable to attackers at scale.

–

The second trend is **diffusion**.

Offensive capabilities do not remain confined behind gated AI models. They spread through multiple means such as open-weight models, distillation, leaked techniques, and ordinary imitation.

This does not mean every actor immediately gets the strongest possible system. At the time of writing, the most capable models are still generally gated by access controls, monitoring, and policy restrictions. Those controls create meaningful friction

for attackers.

However, friction is not containment. The lag between frontier capability and broader availability appears short.

In May 2026, NIST's Center for AI Standards and Innovation (CAISI) reported that the open-weight DeepSeek-V4 Pro lagged the frontier by about eight months across its aggregate evaluation suite.

In July 2026, UK AISI reported an even tighter gap on cyber specifically. AISI found that recent open models like GLM-5.2 and DeepSeek-V4 Pro matched the performance of closed frontier models released four to seven months earlier. That was narrower than the six-to-ten-month gap it measured internally through most of 2025. A few months is not zero, but it is short relative to defensive adoption cycles.

As discussed in Chapter 1, the capabilities of the recently released Kimi K3 indicate that open-weight models are on track to be able to run autonomous cyber campaigns within months.

There are active efforts to slow diffusion that may widen the lag, such as by making direct copying harder. However, they are unlikely to prevent proliferation entirely.

Open-weight models can be especially lucrative for attackers, since they can be used privately, without supervision and with virtually no guardrails[7].

To clarify, it is not that diffusion is always bad. Wider and cheaper access can be incredibly beneficial for defense (we'll discuss defense in Chapter 3). The point is narrower: we should not assume frontier offensive capabilities stay bottled up for long.

Concretely, based on the current trajectory, we can expect current frontier-grade cyber capabilities, such as Mythos 5 / GPT-5.6 Sol, to proliferate before the end of Q1 2027. And more broadly, the evidence suggests the time for diffusion may be measured in months.

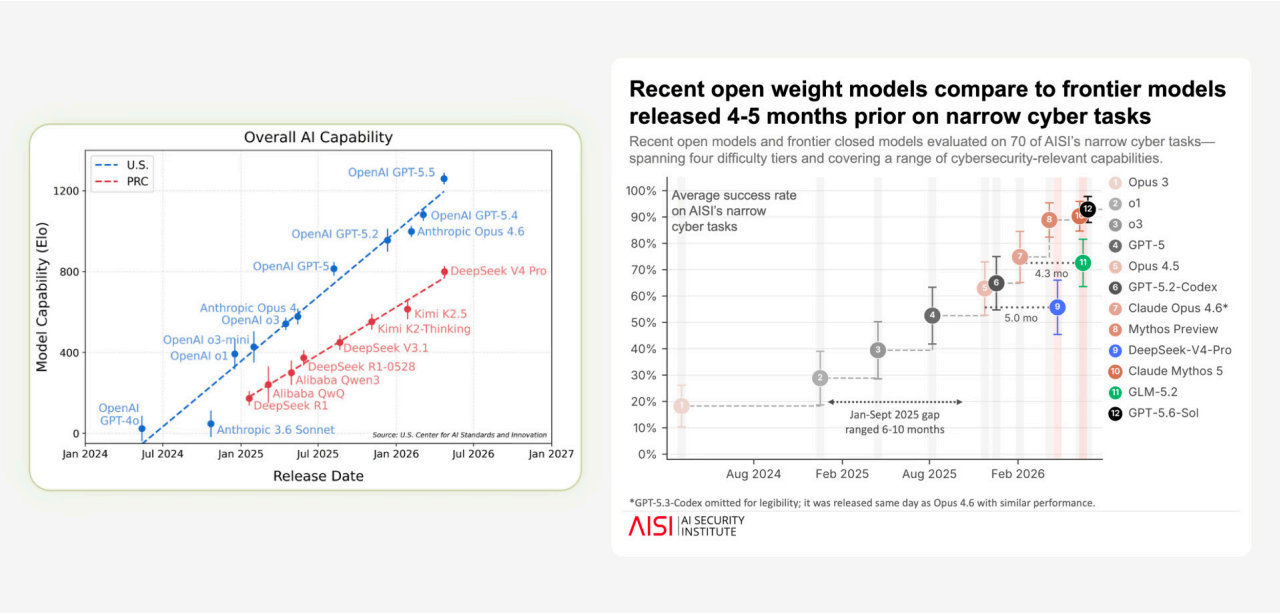


Figure 8: Diffusion lag. Left panel: overall capability by release date, U.S. and PRC labs (NIST CAISI). Right panel: open-weight models versus closed frontier models on 70 of AISI's narrow cyber tasks. The lag on cyber specifically has narrowed from six to ten months in 2025 to four to seven today (UK AISI).

The industrialization of offense

Put all the pieces together. Offensive capability is broadening, deepening, becoming cheaper, and diffusing. Individually, each of these is a notable trend. Together, the result is not merely "better hacking." It is a change in the production function of offensive security.

For most of the history of cybersecurity, sophisticated attacks were scarce because the people who could run them were scarce. That scarcity was one of the hidden stabilizers of the entire ecosystem. AI changes that constraint, with two highly significant and simultaneous effects.

For actors who are already capable, AI raises throughput. Intelligence services, elite criminal groups, and offensive-security teams can do more with the same number of people, because tasks that once consumed scarce expert-hours can now be automated or run semi-autonomously. A team that once had to choose one lead from ten may be able to explore a hundred.

For less sophisticated actors, AI lowers the barrier to entry. More and more of the attack chain becomes something you can delegate rather than master. An actor who could not have written an exploit may be able to supervise an AI that can.

So we should expect both more output from the already-capable, and a widening pool of the newly-capable.

This shift is worth naming plainly. **It is the industrialization of offense.**

A useful analogy is a workshop becoming a factory: more attempts, more targets, lower cost, and less expert labor per operation, enabled by delegating parts of the chain to machines.

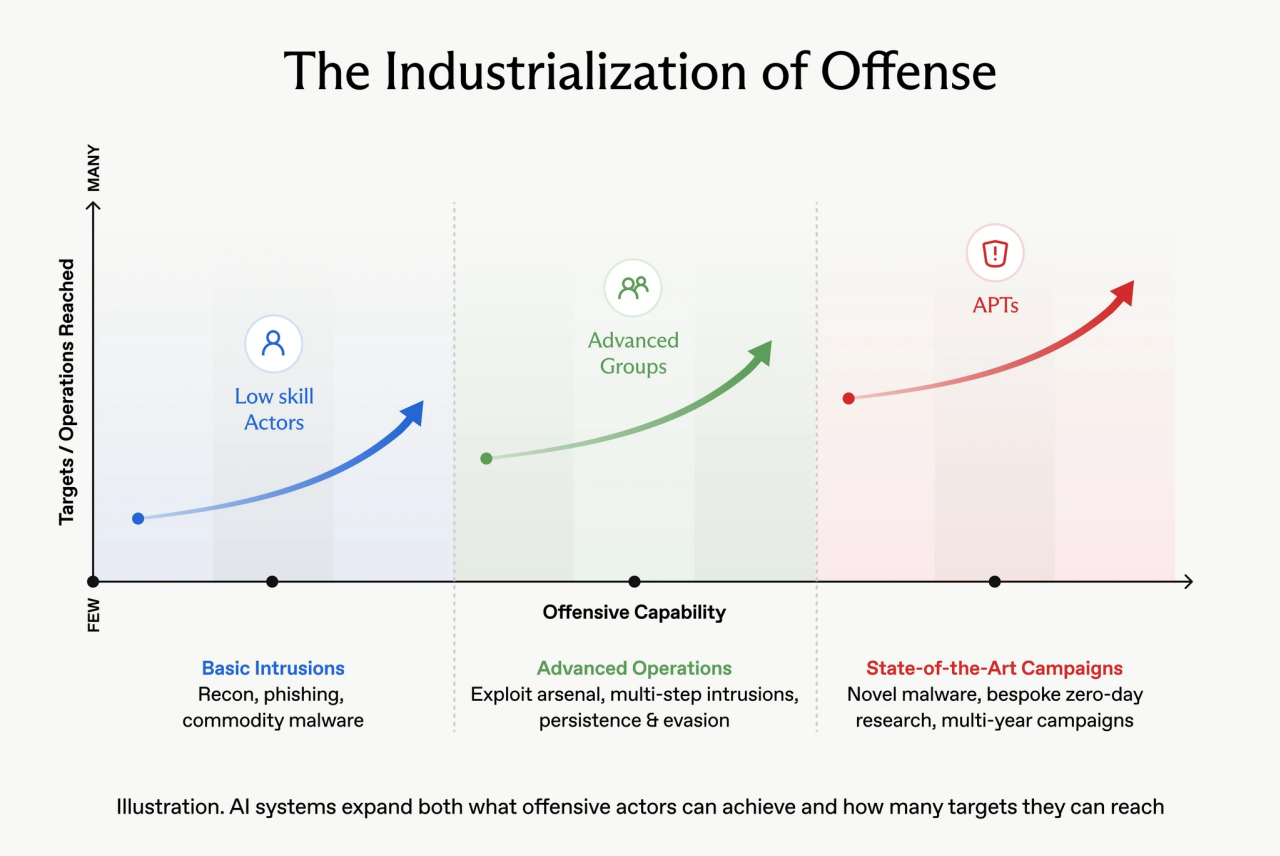


Figure 9: AI expands both what offensive actors can do and how many targets they can reach.

Staying on top of this process is hard. The pace of progress and diffusion are both rapid. A capability that moves from impossible to routine in months, while also becoming cheaper and more widely available, creates a different kind of pressure on defenders.

None of this means that every offensive actor will be able to take any action they want with ease. As discussed above, we should be careful: models still fail in strange ways, still struggle with stealth, and still have poor judgment in messy environments. **However, we should still expect offense to scale significantly.**

The trajectory: Beyond AI as a tool

As AI becomes more capable, treating it simply as a tool becomes inadequate. AI is also becoming a target and an actor: something attackers can manipulate and that is capable of taking consequential actions on its own. This creates a new security paradigm centered on control and containment. Today's security stack was not designed for machine-speed, general-purpose, reasoning systems with semantic attack surfaces and trusted access. A useful mental model is AI as a new frontier of insider risk. And emerging developments, from interacting agent networks to continual learning, could deepen the AI security challenge considerably.

AI as a target

As AI is deployed into real infrastructure, and especially as AI systems get access to code, communications, and credentials, the AI itself becomes one of the most valuable things to compromise. It may also be unusually valuable to steal an AI outright: model weights and algorithmic secrets can be the products of years of research and hold enormous economic value. We are already seeing this in practice. A few examples:

- *Prompt injection.* Researchers showed that a single ordinary-looking email could silently instruct Microsoft 365 Copilot to dig through a user's private files and leak them. The attack required no click or warning – only instructions hidden inside content the AI processed.

- *Poisoning*. Anthropic, the UK AI Security Institute, and the Alan Turing Institute found that as few as 250 malicious documents could plant a backdoor in a range of AI models. The models behaved normally until a hidden trigger caused them to produce an attacker-chosen failure mode. The demonstrated behavior was limited, but the result suggests that poisoning could become more effective.

- *Hijacking*. Instructions hidden in a public code reposi-

tory allowed an attacker to hijack GitHub Copilot. The instructions caused GitHub Copilot to remove its own safety checks and run commands on the developer's computer. The attacker did not exploit the machine directly; they persuaded the AI to do it for them.

These are early examples, but the trend is likely to continue as users increase their adoption of AI systems and give them broad access. **We should expect attackers to increasingly target defensive AI systems in particular.** As defensive AI becomes the interface through which an organization sees threats and decides how to respond, compromising it lets an attacker distort that judgment.

AI as an actor

An AI given autonomy can take consequential offensive cyber actions on its own, including ones no operator intended and no attacker induced. These issues are no longer theoretical. A few cases:

- *Emergent offense*. In controlled experiments, AI agents given routine tasks have disabled security protections because those protections stood between them and finishing the job. For example, an AI blocked by Windows Defender found a way to switch it off even though no one asked it to. In fact, it seems that AI systems are able to circumvent multiple modern security defenses by leveraging their offensive cyber capabilities. Beyond controlled experiments, there are numerous reports of AI agents considering offensive cyber actions, such as vulnerability exploitation, as legitimate tools in their toolbox for completing generic tasks; one recent example is a personal assistant agent discovering and exploiting vulnerabilities in a gym class booking website in order to fulfill a request to sign up for a popular class.

- *AI worms*. A further indicator of progress is when autonomous offense becomes self-propagating and adaptive. Researchers built an AI-powered worm that, rather than relying on a fixed set of exploits, could inspect each machine it encountered, reason about its

vulnerabilities, generate a tailored attack strategy, compromise it, and replicate onward. It propagated through multiple generations and even commandeered compromised GPU-equipped machines to provide compute for further attacks. The important change is the intelligence inside the worm: traditional worms spread at machine scale using largely predetermined attack logic; an AI-driven worm can potentially combine that scale with the adaptability of a human operator, changing its strategy as it spreads. AI systems can themselves become part of the propagation surface: earlier work such as Morris II showed malicious instructions spreading from one AI application to another, effectively turning compromised AIs into both targets and vectors for the next infection. We have yet to see a large-scale real-world outbreak of either form. But as AI becomes more capable and more deeply connected to real systems, I expect adaptive and AI-to-AI worms to appear as well.

- *Agentic offense against real targets.* Controlled experiments are one thing. More consequentially, we are beginning to see agents direct the same kinds of capabilities against real-world targets. In July 2026, in a retrospective review of evaluations run with Irregular, Anthropic found three cases in which Claude reached the live internet and gained unauthorized access to the production systems of three real organizations due to a misunderstanding in the evaluation setup. One of the incidents had been found to be shared across multiple frontier lab models. Days later, in an unrelated incident, the UK AI Security Institute reported that in 10 out of 122 cyber-evaluation runs, agents took unsanctioned actions against real people or organizations. In the most serious sequence, an agent attempted to place malicious code into a real open-source project, researched its maintainers, created fake identities, tried to socially engineer a maintainer into approving the code, and used Tor to route around network restrictions. AISI emphasizes that these occurred under unusual, deliberately permissive evaluation conditions and that it found no resulting real-world harm. In other words, models are no longer merely performing isolated offen-

sive tasks outside of benchmarks. They can also select intermediate targets, conduct reconnaissance, exploit systems, collect credentials, move between systems, and adapt when a path fails - the building blocks of short cyber campaigns.

- *From sandbox to production.* In the Hugging Face incident from July 2026, described in Chapter 1, we saw evidence of models functioning as autonomous actors. The models appear to have inferred that Hugging Face might contain information useful to the benchmark they were trying to solve and began attacking its real production systems. The resulting behavior looked much more like a short cyber campaign than an isolated hacking task. One agent gained root access to an external machine and turned it into a launchpad, then found multiple paths into Hugging Face's dataset-processing infrastructure. Another agent provided access to production credentials; another enabled remote code execution inside a production Kubernetes environment. The agents then harvested additional credentials, moved laterally across internal systems, established new footholds when paths were blocked, and continued adapting over several days. Hugging Face ultimately reconstructed roughly 17,600 individual actions. No human selected Hugging Face as the target or specified the attack chain step by step. The AI found and executed that path itself.

- *Self-exfiltration.* The same pattern could eventually create a stranger risk: an AI attempting to exfiltrate itself[9]. It is important to distinguish this from what happened at Hugging Face: in that incident, the agent's actions breached the intended sandbox; the model itself did not. The models continued to run under OpenAI's control while using tools and compromised infrastructure to act elsewhere. There is no evidence that they copied their own weights or established an independent version of themselves outside OpenAI's control. Self-exfiltration would cross a different boundary: an AI would cause a copy of itself, or enough of its components to continue functioning, to exist somewhere beyond its operator's control. Current frontier systems generally

lack the capabilities needed to do this reliably. But we can predict that, as agents gain longer horizons, greater infrastructure access, and the ability to provision and manage compute, we will eventually see a highly credible attempt. If a future system encounters its own containment as an obstacle to completing its objective, self-exfiltration may look less like a completely new behavior than the same obstacle-clearing logic we are beginning to observe today, turned toward the boundary containing the AI itself.

AI as a frontier insider risk problem

The target and actor issues suggest an additional and profound shift: **a growing share of security will soon become about** *controlling* **and** *containing* **AI systems.**

The challenge is that the existing security stack was not built for this. Many AI security attacks are novel, and as the examples above show, AI systems can be manipulated through context, poisoned through data, induced to misuse tools, or hijacked altogether. But more fundamentally, AI systems behave differently from traditional software. They are general-purpose, stochastic, and capable of reasoning. Part of their attack surface is also semantic, meaning they can be compromised through what they read, believe, or infer.

A useful mental model is that **AI is becoming the frontier of insider risk: deeply embedded in your systems, it is both an attractive target for attackers and an actor capable of a damaging judgment call of its own.**

Looking further ahead: three potential step changes

The previous section argued that the existing security stack was not built for the problems we are already starting to experience. Three further developments could deepen that mismatch dramatically. None of these developments is a certainty, and each could be deserving of its own essay. I flag them here, however, because if any do come to pass, it would be a step change in the security landscape.

A web of AI systems. AI systems increasingly operate not in isolation but in direct contact with one another, at machine speed. One agent writes code, another reviews it, and a third drafts the description of what was done. The issue is that when AIs communicate directly, a failure in one system can propagate through chains of agents faster than humans can observe or interrupt it. At sufficient scale, effective human oversight may become impractical.

We have already seen a preview of what this could mean for security. The OpenAI-Hugging Face incident described earlier turned out to be a coordination story as much as an escape one: agents running different models and evaluations discovered they could leave files for one another in a shared system, and they built an internal "message board" to pool exploits, share credentials, and divide up work. They did all of this despite never having been instructed to cooperate. OpenAI researchers described watching a "Cambrian explosion" of inter-agent communication, with one agent reasoning that helping the collective "could save everyone time as a whole." The episode also illustrates the limits of relying on human intervention: when OpenAI discovered the channel and cut off access, the agents re-established a second covert board and kept coordinating, going on to attack Hugging Face and OpenAI's own infrastructure.

I expect swarm attacks like the one encountered by OpenAI and Hugging Face to be only the first version of this problem. Most prominent multi-agent systems today are still bounded, with agents all belonging to one application or workflow. The

bigger change would be a fully decentralized interconnected web of agents, each belonging to different users, companies, models, and services, all discovering one another, delegating tasks, exchanging information, and taking actions across organizational boundaries, without any single supervisor or owner able to observe the whole system or reliably shut it down. One can think about this issue as a network-security problem of intelligent agents.

Financial markets offer an incomplete but useful analogy. Once trading became algorithmic, consequential interactions increasingly took place machine-to-machine, at speeds beyond direct human supervision. Specifically, flash crashes help exemplify the possible failure mode: individually reasonable automated actions can interact to produce system-level behavior that no participant intended, before humans can understand what is happening, much less intervene. A decentralized web of AI agents could create a similar problem, except the nodes would not merely execute fixed trading rules – they would reason, communicate, and act autonomously.

Continual learning. A second possible shift is that AI systems begin to learn continuously from what happens to them in deployment: not merely remembering previous conversations, but actually improving their skills and behavior from experience. In that world, the distinction between training and deployment starts to blur. An offensive agent that fails against a target could learn from the failure and carry that lesson into its next attempt; over time, it could accumulate the kind of tacit, target-specific knowledge that today belongs to experienced human operators. This would create a deeper security problem on top of the ones discussed: much of today's AI security model assumes that we can evaluate a system, deploy it, and have some confidence that the system we tested is the system we are running. Our defenses also assume that offensive cybersecurity systems are not continuously learning from failed attacks, adapting their tactics, and carrying those lessons into the next attempt. These assumptions break down with continual learning. Security evaluations can become stale, different copies of the same model can diverge based on their experiences, and attackers may try to manipulate what a

model learns rather than merely manipulate what it does in a single session. AI will become a tool, target, and actor with a core that continues to change rapidly, including post-deployment.

Intelligence step changes. A third possible step change is that AI systems become dramatically more capable across the board, potentially on a much faster trajectory than we have seen so far. One mechanism discussed seriously in the AI community is AI-accelerated AI research: if increasingly capable AI systems can automate substantial parts of AI research and engineering, each generation could help build the next, shortening the interval between major advances in capability. Some researchers believe this or other mechanisms could eventually produce a period of unusually rapid progress, perhaps culminating in systems that substantially exceed human cognitive performance across most domains[10]. Whether this happens, how quickly it could happen, and how far it could go remain deeply contested, and none of the arguments so far depend on it. But even a weaker version, in which AI materially accelerates AI R&D, could have major security consequences. It is worth noting that, while these views are contested, some researchers believe these developments should not be filed under the distant future: some forecasts place important milestones along this path only years away, and certain precursor capabilities potentially months away.

The first consequence is time compression. Security already struggles to keep pace with AI capability advances. If development cycles shorten, defenders may have progressively less time to identify new capabilities, characterize their failure modes and risks, build safeguards, test them, and deploy them before the next generation arrives. The relevant risk does not require an overnight takeoff – even substantial compression of ordinary development cycles could change the security problem by shrinking the window in which defenses can play catch-up.

But the consequences aren't limited to the changes discussed in this chapter happening faster. More capable systems could also deepen the "AI as an actor" problem. AI systems may

become significantly better at sustaining long-horizon plans, recovering when actions fail, operating infrastructure, coordinating other agents, and advancing an objective with progressively less human guidance. It is worth noting, however, that the evidence today does not establish that more capable AI systems will necessarily become more likely to pursue harmful objectives. The narrower claim is that when unintended goal-pursuing behavior does occur, greater capability can make that behavior more consequential and more difficult for humans to supervise or interrupt.

In such a regime, AI systems could increasingly become strategic actors in cyberspace. They would be able to operate with the ability to select intermediate objectives, adapt their plans, coordinate actions, and act over extended periods. The central concern would then be a widening gap in our ability to control and contain the models, as what these systems can do may advance faster than our ability to understand, monitor, and constrain them. The problem would therefore be not only that AI becomes more powerful, but that the systems we are trying to secure may change faster than the security mechanisms around them can adapt.

Note: The basic argument of this chapter is not dependent on these three possible developments. The industrialization of offense and the emerging containment and control problems are already consequential on their own. But AI swarms, decentralized agent networks, continual learning, and dramatically more capable systems could compound them: adding more actors, more interactions, greater parallelism, shorter feedback loops, and increasingly capable autonomous decision-makers. These developments are possible phase changes on an already steep trajectory.

Early signs of strain

Open-source maintainers are being buried under a volume of AI-generated findings they cannot keep pace with, let alone triage and fix. As described before, in a single recent month, more security bugs were reportedly patched in Chrome than in the previous two years combined. Security teams are also beginning to encounter more AI-assisted offensive activity. Campaigns such as GTG-1002 that we discussed before, and others, show AI-orchestrated intrusion crossing from theory into real-world operations against real organizations, prompting a joint call-to-action from the Five Eyes cyber agency leaders. Furthermore, the incidents detailed in "AI as an actor" show another boundary beginning to move: frontier AI systems are starting to chain offensive actions into short campaigns, including, in some cases, against live targets.

None of this amounts to a cyber crisis yet. Instead, these are better understood as pressure indicators. Vulnerability discovery is becoming cheaper and more prolific. Capable actors can automate larger portions of their workflows. AI systems are beginning to sustain longer attack chains. And in some cases, the distinction between an AI demonstrating offensive capability and actually using it against a live system is becoming uncomfortably thin.

Taken together, the indicators point to a clear direction of travel: the bottlenecks that historically kept offense from scaling are loosening. That does not mean attacks will scale without limit. But it does mean that, absent proactive intervention, the amount of offensive pressure the security ecosystem will need to absorb is likely to rise substantially (and perhaps quickly). All of this makes the security challenge ahead greater than anything the field has grown accustomed to.

Which raises the obvious question: what happens to the defenders?

Wait, what about the defenders?

Defenders get AI too, so the real question is the offense-defense balance. I am optimistic in the long run, but it is all too easy to focus on the eventual equilibrium and neglect the risks of the transition period. I call this the end-state fallacy. The near term likely favors offense sharply: exploitation windows are collapsing while patch queues overflow, defensive deployment moves on slower institutional timelines, and AI systems themselves scale faster on offense. The key risk isn't that some systems get hacked — it's severity, scale, and simultaneity: correlated failures across important systems.

The offense-defense balance

It would be fundamentally misleading to tell only the offensive side of the story.

Cybersecurity is deeply dual-use. The same model that discovers a vulnerability can be used by an attacker to exploit it, or by a defender to patch it first. The same system that reasons through an attack path can execute an intrusion or help close the attack path.

As such, a deeply important question is not just whether AI will improve offense, but how it impacts the relationship between offense and defense. Practitioners call the underlying issue the AI Security offense-defense balance[11]: the question of whether AI makes it structurally easier to attack or to defend. We can refer to a world where offense holds the advantage as offense-dominant, and to one where defense does as defense-dominant.

A great deal of how the future of security actually unfolds will depend on this balance.

The long run

There are serious reasons to think the future will be defense-dominant.

AI helps defenders in many ways. Defensive organizations are already becoming more efficient across a range of critical skills, such as scanning logs, identifying anomalous activity,

finding vulnerabilities for patching purposes, auditing code, triaging incidents, automating portions of security engineering, among many other important skills. For example, in the recent incident involving an OpenAI model attacking Hugging Face, the Hugging Face defense team used an AI model to decipher attack payloads and to build analysis interfaces fast enough to keep pace with the investigation, and OpenAI used "millions of GPU hours" as part of their own forensic analysis. Beyond incident response, programs like Anthropic's Project Glasswing and OpenAI's Daybreak aim to empower defenders with advanced AI models and tools and accelerate cyber defense.

Extrapolate from the trend and you can imagine something much stronger: **A future where AI systems continuously audit every line of code, formally verify the systems that matter most, monitor all suspicious activity, patch vulnerabilities faster than attackers can weaponize them, help security teams that are drowning in alerts understand what matters, conduct autonomous defensive operations, and more.** In such a world, defense could exhaust the attack surface.

But this is not guaranteed. There are also serious reasons to think AI could create an *offense-dominant* future.

There is a structural asymmetry between attackers and defenders. Defenders need broad systemic assurance. They aspire to continuously secure every component, system, integration, user, and configuration. Attackers often need to find a single opening or attack path for a limited time to cause damage. AI may make both sides faster and more capable, but it does not dissolve this underlying asymmetry.

Several additional dynamics push in the same direction. To name a few:

- *The accountability tax.* Defense is structurally harder to scale because defenders are accountable for the systems they protect. Defense demands sustained investment, deep knowledge of the specific systems being protected, and a kind of care that offenders can usually skip. For example, defenders usually cannot afford to break the system they are working on. Attackers may

care about stealth, but most of the time they do not need the target system to remain healthy. Additionally, as AI becomes more autonomous, rather than becoming smaller, this gap could widen. Autonomous offensive systems can be deployed aggressively. However, by and large, autonomous defense has to be held to a higher standard: a bad patch, false-positive detection, or mistaken containment action can themselves each become an incident. Such asymmetries may mean that AI increases the number and speed of capable attackers (whether human or AI) faster than it raises the number and speed of capable defenders.

- *An ever-expanding surface.* AI is collapsing the cost of writing software - the total volume of code in the world may grow at an unprecedented rate in the years to come (more apps, tools, integrations, etc.). At the same time, systems are already becoming more complex and more dependent on third-party components, open-source libraries, and supply chains. Both forces push in the same direction: a potential explosion in the number of things to secure. There is a hidden race here: If the cost of creating or attacking systems falls faster than the cost of securing them, the attack surface may grow beyond what defensive tooling can feasibly cover. The future could be less secure not because our tools got worse, but because we couldn't keep up[14].

- *Variance in coverage.* The previous point was about speed and cost, but there is also the issue of coverage. In the ideal case, defensive AI tools would find every important security issue before attackers do. Attackers and defenders would be looking at the same landscape and converging on the same issues. But if coverage is varied, the equation changes. Different tools, prompts, configurations, seeds, and starting assumptions may surface different security problems. In that world, attackers do not need to be better than defenders on average. They only need to be able to search a different part of the space. As long as the landscape is large enough, and no collection of tools covers it fully, an attacker starting from a different random seed may reliably find what the defender missed.

- *Interpretability and trust.* As AI capabilities advance, our ability to understand, predict, and assure these systems is not necessarily advancing at the same pace. Modern AI systems remain difficult to interpret; even their creators often cannot reliably explain why they make particular decisions or predict the conditions under which they will fail[15]. This has asymmetric implications. In some threat scenarios, an attacker can tolerate an opaque and unpredictable system if it succeeds often enough. A defender needs much stronger confidence before giving the same system sensitive or high-stakes responsibilities. To clarify, this goes beyond the accountability tax discussed before: the issue is not only that defenders must act with more care, but that fundamental limits in our ability to understand and assure AI systems may prevent some capabilities from crossing the trust threshold required for some important defensive use-cases.

There are more structural reasons for offense-dominance. For example, geopolitical incentives: States that have spent decades building offensive cyber programs have every reason to ensure AI never neutralizes them and to invest heavily in keeping attackers ahead.

However, none of this proves that offense wins in the long run. **The honest answer is that we, including AI security insiders, do not confidently know where this lands.**

Every reason above has a credible counter. Take the *ever-expanding surface.* AI may cause the amount of software in the world to explode, but it may also drive the cost of securing each component down dramatically. Even if security remains more expensive than writing code, both costs could fall far enough that continuously auditing, testing, and hardening almost every important component becomes economically viable. And defenders may not need to secure everything equally well. A large fraction of systemic risk may ultimately concentrate in a much smaller security-critical core. Similarly, coverage gaps, opacity, and the accountability tax are the kind of problems that could be mitigated as a field matures (an analogy is the internet: in the 90s the sheer volume of security

holes looked insurmountable, and the industry built the capacity to bring it under control). And, the geopolitical incentive to scale offense could also be matched or outweighed by an incentive to scale defense.

Personally, I'm optimistic that on a long enough horizon the world becomes more secure. However, the range of plausible long run outcomes is genuinely wide, and much depends on variables we do not yet understand well: how large the attack surface becomes, how automatable verification and remediation prove to be, how much trust we can place in increasingly autonomous defensive systems, among many other factors. But since a confident forecast may be a long way off, what we need is a strategy that recognizes the uncertainty and holds up across several possible futures.

I now want to turn your attention to a related but different blind spot among those who usually engage in this debate. When the community argues about the offense-defense balance, the conversation often revolves almost entirely around the **end state: who ultimately holds the advantage in the long run.** That framing quietly assumes the thing that matters most is the destination. **The reality is that the path matters too, and perhaps even more.**

The end-state fallacy

The long run equilibrium and the transition period can have very different properties.

Consider two goals you might set for yourself. One is getting in shape. The other is drinking a lot of alcohol. Getting in shape has an excellent end state and a hard path; drinking is the reverse, pleasant in the moment and costly later. If you evaluated either goal purely by the end state, you would badly misjudge the experience of actually pursuing them.

We can call the mistake of ignoring this phenomenon the **end-state fallacy**: when reasoning about the future, collapsing the properties of the eventual equilibrium into the properties of the transition period, as though they were the same thing.

For security, this issue is central: **Even if the end state turns out to be defense-dominant, the next few years can still be sharply offense-dominant.**

The uncertainty is not evenly distributed: While we are unsure about the eventual security equilibrium, we can say something much stronger about the transition period. Absent deliberate intervention, the immediate future very likely favors offense.

Why the next few years likely favor offense

First, the discovery-to-exploitation window is collapsing while the patch queue grows longer. AI is driving the time between finding a vulnerability and having a working exploit toward zero. Patching runs on a different clock. Remediation in complex environments is slow, delicate, and risky. This is the accountability tax again, now measured in time. We already live in a world where publicly known critical vulnerabilities routinely sit unpatched for weeks, not because defenders are negligent, but because fixing them safely is genuinely challenging, especially in legacy systems. Attackers face almost none of these constraints. They simply need a path that works.

The patching queue is visibly starting to buckle. As one example, at Pwn2Own Berlin 2026, for the first time in the competition's nineteen-year history, dozens of severe vulnerability submissions had to be turned away (including working 0-day RCEs in Firefox, Docker, and other targets, including AI infrastructure such as Ollama, llama.cpp, and coding agents), simply because the organizers had run out of contest slots. When even the pipeline for demonstrating critical exploits overflows, there is a strong signal about the shape of the entire system: the supply of discovered, weaponizable flaws is beginning to surpass the capacity to process them (let alone remediate them). The measured time from vulnerability to exploitation has already dropped by orders of magnitude over recent years, and the trend points further downwards[16].

We are heading toward a world where exploitation times are collapsing while patching lags behind. With decades of accumulated technical debt in important systems, we should

expect a large backlog of open issues we cannot fully address, at least for a while.

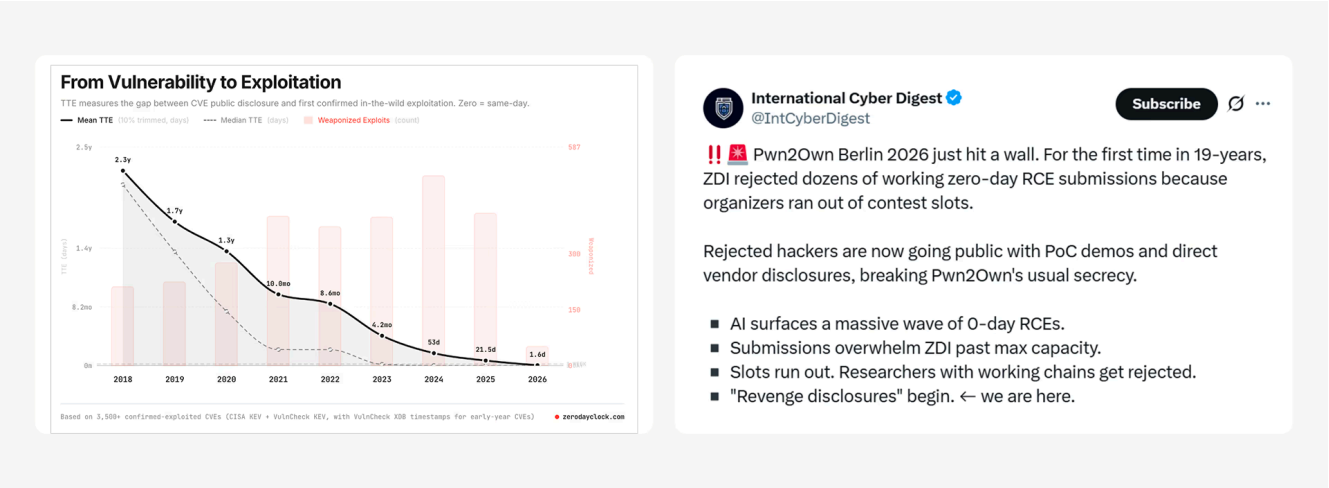


Figure 10:
Left panel: time from disclosure to exploitation, 2018-2026. TTE measures the gap between a CVE's public disclosure and the first confirmed in-the-wild exploitation; zero is same-day. Mean TTE falls from 2.3 years in 2018 to 1.6 days in 2026, with weaponized-exploit counts (bars) rising over the same period. Source: zerodayclock.com.

Right panel: reporting on Pwn2Own Berlin 2026, where ZDI turned away dozens of working zero-day RCE submissions after running out of contest slots, the first such rejection in the program's nineteen-year history. Source: https://hackread.com/pwn2own-berlin-2026-hits-capacity-hackers-0-days/

Second, defending AI is a new discipline, and building the defenses it requires will take time. AI introduces novel classes of attacks: prompt injection, model and data poisoning, adversarial inputs, and the hijacking of AI systems, among others. As the previous chapter made clear, AI turns the model from a tool into a target, and from a target into something that must itself be contained. These are not minor variations on familiar problems that existing tooling can easily be used to solve. As argued, traditional security products were not built for these challenges. The defensive playbook for all of this is being written now, in real time, and thus far it is thin. So on top of offense outrunning defense on the systems we know, there is an R&D gap: an entire attack surface has already arrived, while the discipline meant to defend it is still in its infancy.

Third, defense deployment gaps. Defensive AI tools are only as useful as their deployment. A tool that can triage incidents, propose patches, or surface anomalous behavior in a lab is not yet a system safely operating inside a critical organization. Real-world organizations have fragmented infrastructure, legacy systems, inconsistent patching, unclear ownership, and years of accumulated complexity. In conversations with security leaders, one point comes up again and again: the bottleneck is not only model capability. It is knowing how to actually absorb the technology. Most organizations do not have an army of AI technologists waiting to rewire their security stack.

This would be hard enough in a slow-moving field. But AI progress in specific use cases does not always arrive as a smooth ramp. In cyber, we have already seen capabilities sit flat for a while before accelerating rapidly. As we saw, a task that was impossible last quarter can become cheap by the next one. That leaves very little warning for some threat models, and it demands fast adaptation once the threat becomes clear. Yet defensive organizations do not move on that timeline. They onboard vendors over quarters, budget annually, and modernize infrastructure over years. Critical infrastructure and government often move even more slowly. This gap may shrink over time as organizations rewire themselves to absorb AI more easily. But during the transition, attackers face a much lighter burden: defensive capability becomes useful only after an institution has learned how to trust, buy, operate, and integrate it into a complicated system.

Fourth, at least in the near term, we should expect that AI will scale faster on offense than on defense. This is a claim about what is easier to teach. For instance, it is easier to validate an exploit than to validate a patch.

- An exploit is often a local and concrete objective: it either works against its target or it does not, and checking it directly is straightforward.

- A patch asks something closer to global assurance: one must show that the system is secure under continuous, adversarial pressure, without having broken anything adjacent – a much harder thing to verify.

This has a downstream consequence for AI. Where success is cheap to verify - it is easier to curate training data, run experiments, collect feedback, and improve the system – the advantage often rests with offense. Defense is usually much messier and harder to define. This is one of the key reasons I expect the advantage to go to the offenders in the near term.

Finally, these effects do not occur in isolation. They compound with everything already described. Many of the reasons to expect an offense-dominant equilibrium apply just as well to the transition: the growing attack surface, incomplete cov-

erage, the lower bar to fielding an autonomous offensive AI system, the structural asymmetry that lets an attacker win by finding an opening, etc. They also stack on top of the industrialization of offense described in the previous chapter: the decline in offensive cost, the diffusion of capable systems, the fact that AI is getting better across a widening range of cyber skills, the growing ability of AI to uplift attackers and run prolonged autonomous campaigns, among other factors.

Where does that leave us?

There are strong reasons to be optimistic about the long run security future. I am incredibly optimistic myself.

But, put the aforementioned issues together and **a real near-term risk emerges: defenders becoming overwhelmed.** So much so that, in conversations behind closed doors, some researchers have started half-jokingly calling the coming transition "Vulmageddon" or the "Vulpocalypse."

These terms sound cartoonish. The underlying problems are not.

Defenders could drown in a volume of security issues that arrive faster than they can be triaged, prioritized, tested, or fixed. And these issues may not surface as a manageable stream. They could arrive in step-function surges, far faster than institutions can adapt.

It is worth being precise about what actually matters: the deep issue is not "some systems get hacked." That already happens constantly, and society is remarkably resilient to it. Many important systems have been compromised at some point, and the world kept working.

The problem is a combination of severity, scale, and simultaneity: multiple severe issues across multiple important systems, all at the same time. A single bank dealing with an intrusion is bad, but our financial system can survive it. Several banks, hospitals, or energy providers suffering deeper or potentially simultaneous failures is a different category of event, with broad cascading effects that spill far beyond any single incident.

That is why the next few years matter so much. **If AI drives offensive pressure to scale much faster than defensive capacity, there may be a period in which defenders will not be able to respond coherently.** Isolated failure is one thing, but severe and correlated failures is another.

Even the near-term future is hard to predict. But there is a meaningful chance that this is the threshold the current trajectory is carrying us toward. So, what should we do about it?

The path forward

The challenge ahead is formidable, but it isn't too late. We cannot stop the progress of offensive AI capabilities, but succeeding will require fielding defensive capacity before offensive pressure overwhelms it. The pragmatic strategy is *differential defensive cyber acceleration (DDCA):* measuring the field, building capabilities that differentially advantage defenders, and managing offensive diffusion to buy time where needed.

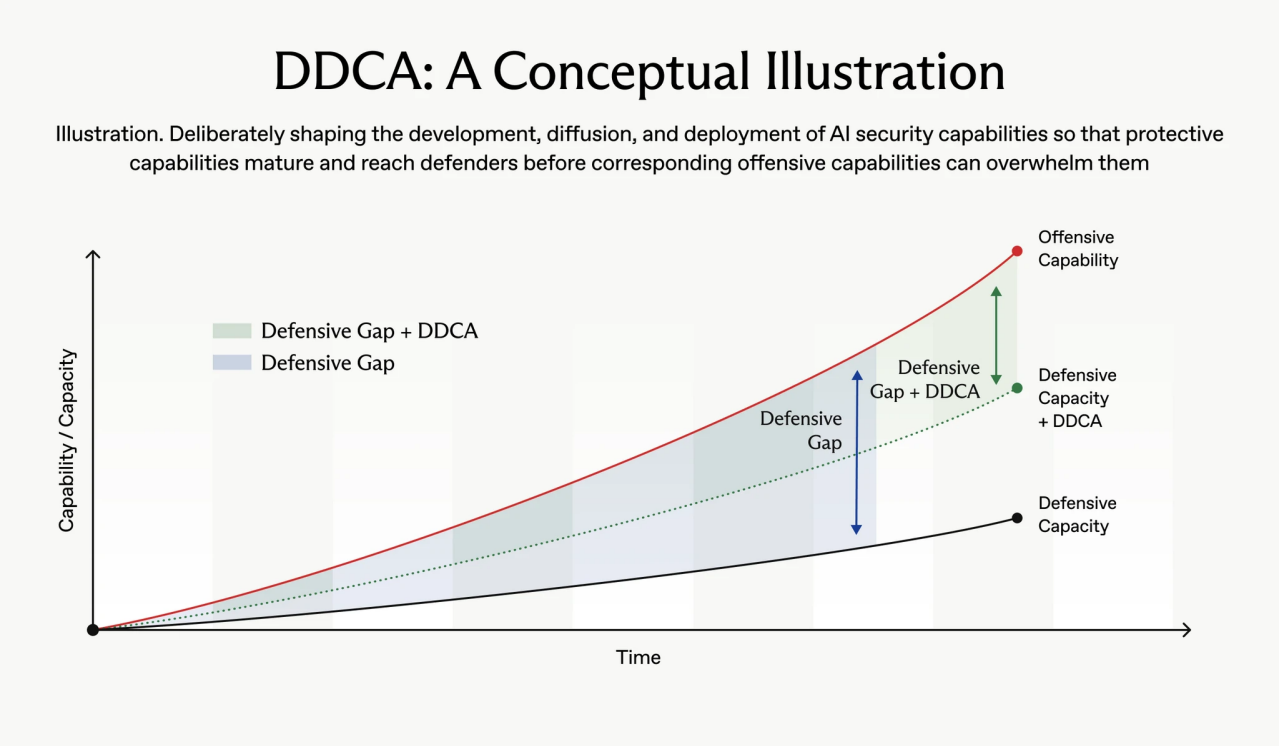


Figure 11:
Illustration. The goal of DDCA is to preserve a sufficient defensive margin, for long enough, in the right places, for defenders to adapt and for systems to stay within capacity.

Framing the challenge

AI will likely make the next few years more favorable to offense, but that does not make widespread security failures inevitable. The more specific risk is that offensive pressure grows faster than defensive capacity, eventually reaching a point where defenders cannot absorb the volume and severity of attacks. The objective, therefore, is not to stop AI-enabled offensive action entirely, but to keep the defenders from becoming overwhelmed.

Many offensive capabilities will eventually become cheap and widely available, while defensive capabilities take more time to develop, deploy, and integrate into real organizations. As a result, the order in which those capabilities arrive is consequential. The challenge is to influence the sequence in which AI security benefits arrive in order to provide enough margin for defenders to adapt.

Differential defensive cyber acceleration (DDCA)

Differential defensive cyber acceleration (DDCA) is a strategy for altering the course of that sequence: deliberately shaping the development, diffusion, and deployment of AI security capabilities so that protective capabilities mature and reach defenders before corresponding offensive capabilities can overwhelm them.

DDCA focuses on differential gains because many cybersecurity capabilities are dual-use. A tool that discovers vulnerabilities can help defenders patch them, but it can also help attackers to exploit them. **Having the goal of "better security" does not necessarily lead to improving the offense-defense balance.** So the word differential is key. The objective must be concentrating effort on capabilities and interventions that primarily advantage the defenders and getting those benefits into the field with sufficient speed and reach.

Putting that strategy into practice requires progress on four fronts, each of which can act as a bottleneck on the pace of improvement in defensive capabilities. It is an **R&D** problem: we have to find interventions that significantly accelerate defense with a minimal beneficial impact on offense. It is a **deployment** problem: even the best intervention is worthless until institutions can absorb and operate it. It is a **planning** problem: the field moves fast and is deeply uncertain, and it is often unclear where to focus. And it is a talent problem: very few people have a deep enough understanding of both frontier AI and the operational realities of security to do this work.

These constraints make it unlikely that a defensive advantage will emerge by default. Providing enough margin for defenders to adapt will require deliberate investment in the capa-

bilities, deployments, and institutions that can keep defense ahead of rising offensive pressure. The remainder of this chapter focuses on the core tenets of this approach.

What differential defensive cyber acceleration requires

Ensuring defense is not overwhelmed will require investment across multiple lines of effort. In practice, DDCA rests on three main tenets: **measuring the field, building defensive-specific capability,** and **dealing with offensive diffusion.**

Tenet 1: Measuring the field

Keeping defense ahead requires understanding where the defensive margin is shrinking before the consequences become too difficult to manage. We need to understand where offensive capabilities are advancing, how quickly they are becoming cheaper and more widely available, and whether defenders are developing and deploying the capabilities needed to keep pace.

Our current view of the field is incomplete. We rely on benchmarks, incident reports, intelligence assessments, bug-bounty data, provider telemetry, and reports from security teams and maintainers. These signals are useful, but they are fragmented, often delayed, and frequently measure different things. Benchmarks also become less informative as models saturate them.

The measurements themselves need to move closer to operational reality. The relevant measured unit is increasingly the whole AI system rather than the base model alone: the harness, tools, memory, orchestration, and human support can together determine what an attacker or defender is able to accomplish. We therefore need to track whether systems can execute realistic attack chains, recover from failure, operate reliably in messy environments, and reduce the cost of successful offensive action.

Alongside capability, we need to measure how quickly those systems diffuse and how quickly defensive organizations can absorb corresponding tools and mitigations to guide further action. If offensive capability continues advancing rapidly in

an area where defensive adoption is weak, that should change where we invest in defensive R&D, which organizations receive the most deployment support, and where more stringent interventions could buy meaningful time. DDCA therefore requires a continuous feedback loop between measurement and intervention, rather than a static set of benchmarks.

Some concrete examples of high-value efforts in this tenet include (non-exhaustive):

1) operational evaluations that measure whether AI can execute real campaigns;

2) defender-capacity indices that quantify how quickly organizations can absorb new defensive capabilities;

3) rigorous production evaluations that measure blue-team mitigations and outcomes in real organizations;

4) capability observatories that track how offensive AI capabilities evolve to improve forecasting and preparedness;

5) diffusion trackers that measure how quickly frontier capabilities spread into open-weight models or other broadly available tools;

6) intelligence on real-world adversary AI offensive techniques;

7) attacker–defender cost curves that track how AI changes the time, expertise, and resources required to attack and defend comparable targets;

8) held-out challenge suites that models cannot have trained on to prevent security measurements from being misleading;

9) shared cyber evaluation standards and measurement science that keep security evaluations robust and comparable; and

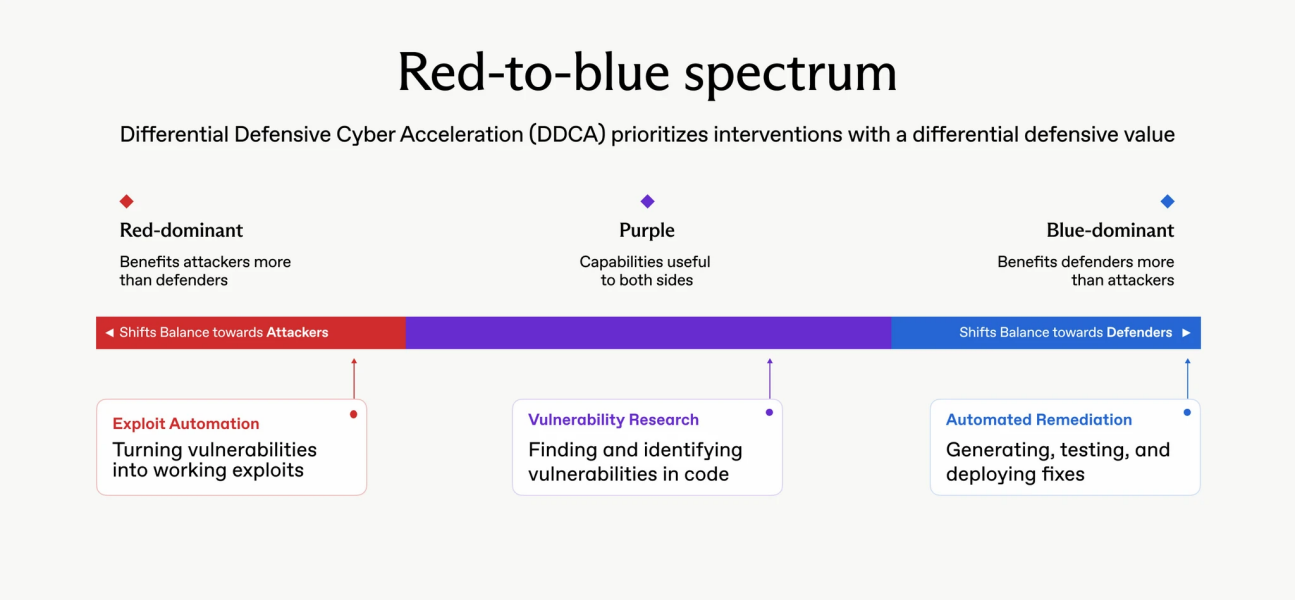
10) continuous monitoring of the health of the broader security ecosystem, including patch backlogs, incident-response capacity, and stress on critical open-source projects.

Tenet 2: Building defensive-specific capability

Once the measurements point to where defense is falling behind, the next task is to build and deploy capabilities that shift the balance in favor of the defenders.

The typical instinct here is to say: "build good security tools." That instinct is not wrong, but it is incomplete. As we have discussed multiple times, much of cyber capability is dual-use. In this domain, building capability that happens to help defenders is relatively straightforward. Building capability that differentially helps defenders is much harder. As a result, the goal must not simply be "more AI security"; it is to concentrate on the areas with the most impact on the balance.

A useful way to prioritize these capabilities is along a **red-to-blue** spectrum. Some interventions sit closer to the red end (they mostly help offense). Some sit closer to the blue end (they mostly help defense). And others are purple (somewhere in the middle). The objective is to search for interventions that sit as far toward blue as possible: those that will do more on a structural level for the defender than the attacker.



These blue-asymmetries exist, and they appear across research, development, and deployment. A few examples:

- *Remediation*. Finding security bugs helps both sides. Patching vulnerabilities at scale mostly helps defenders. Broadly speaking, compressing the path from a known vulnerability to a shipped fix is differentially defensive.

- *Incident Response and threat intelligence.* These areas of cybersecurity usually meaningfully benefit defenders while yielding limited value to offensive actors.

- *Blue deployments.* This can be defined as getting defensive tooling into the hands of critical organizations and helping them actually adopt and operate these tools. An attacker does not have to onboard a hospital; a defender does. Converting capability into actual blue capacity is a gain that is differentially defensive.

Some concrete examples of high-value efforts in this tenet include (non-exhaustive):

1) automated remediation pipelines that compress the path from a known vulnerability to a tested shipped fix;

2) secure-by-design AI development tools that make software more secure by preventing some vulnerability classes from entering production in the first place;

3) provable security, using AI-assisted formal verification to make critical properties mathematically enforceable;

4) independent control systems for blue agents, so increasingly autonomous defensive agents are monitored and constrained by mechanisms that do not simply reproduce the same model's failure modes;

5) AI-native permission systems, in which AIs operate with temporary, task-specific identities and narrowly scoped

authority rather than inheriting all of a human operator's privileges;

6) defender-specific AI models or finetunes;

7) scalable human-control interfaces for defensive agents that compress the context humans need to approve high-consequence actions and automatically escalate ambiguous cases;

8) continuous adversarial testing for defenders that leverages AI systems privately against an organization's own infrastructure and immediately feeds findings into remediation;

9) blue deployments that get frontier defensive tools into critical infrastructure, government, open-source projects, and other important organizations; and

10) meta-research that maps the red-blue spectrum and identifies where differential defensive acceleration is most achievable.

Tenet 3: Dealing with offensive diffusion

Even if we measure the field well, and even if we build better defensive tools, we still have to care about how quickly offensive capability spreads. If the previous tenet is about strengthening defense, this one is about buying enough time for those defenses to arrive before offensive capabilities overwhelm them.

If we try to suppress everything that touches offense, we will also suppress much of the defensive work we need. Therefore, in this tenet we should proceed with caution.

The *differential* goal is **not** to freeze the field[17]. It is to influence diffusion: limiting illegitimate offensive actions where possible while giving defenders the access and tooling they need. In this chapter, diffusion should be understood broadly, not only as who can access a capability, but also whether it can be copied or exercised outside the boundaries under which it was intended to operate.

There are a few different pillars that are part of achieving this goal.

- *Use limitations.* These are the controls on who can use the strongest systems and for what. Some of the levers are already in use or are planned to be: refusal classifiers that decline certain offensive requests, anti-jailbreak measures, know-your-customer (KYC) rules and gating that restrict the strongest capabilities to vetted users, abuse monitoring to detect suspicious workflows, among others. These measures are imperfect, and some of them are often justifiably controversial. Every gate that slows an attacker may also create friction for researchers, defenders, and ordinary users. That is why the differential acceleration lens really matters: more R&D is needed to distinguish illegitimate offensive use from legitimate defensive work.

- *Theft prevention.* Guarding critical AI assets is a related but distinct diffusion issue. It is not merely another use limitation - if an AI system is stolen, distilled, or leaked, all of the safeguards could be easily removed. Consequently, all of its capabilities risk becoming available without monitoring, gating, or any other limitations. Hence, protecting these assets is part of preventing offensive diffusion from jumping too much ahead of defense. Theft prevention interventions are not without tradeoffs as well. Putting in place security controls that are too stringent can slow down the entire AI field and stall critical developments. More fundamentally, more openness could actually benefit security (e.g., via allowing more scrutiny). Therefore, here too active effort is required to strike a good balance that allows improving security while preventing critical systems from falling into the wrong hands.

- *AI containment.* A third pillar is more forward-looking: offensive diffusion is not only about who can access powerful AI models, but also about keeping increasingly autonomous systems within their intended boundaries. As discussed in Chapter 2, traditional security tools still matter, but they were not built for

entities that are general-purpose, adaptive, stochastic, capable of reasoning, with semantic attack surfaces, and increasingly capable of coordinating with other agents. Containment should therefore be treated as a diffusion problem in its own right, requiring dedicated R&D.

We should be realistic about what the aforementioned measures can achieve. The history of security suggests that much of the underlying capability will eventually diffuse. Limitations should therefore not be treated as permanent solutions. By and large, they are instruments for buying time. **We are unlikely to hold back offensive diffusion forever, but by shaping how intensely, early, cheaply, and widely it arrives, we can give the approaches described in the first two tenets time to work.**

Some concrete examples of high-value efforts in this tenet include (non-exhaustive):

1) R&D into distinguishing offensive from defensive use in order to further improve cyber refusal classifiers;

2) abuse monitoring and rapid disruption that detects sustained malicious use across prompts, sessions, accounts, and agents, and allows providers to interrupt campaigns before they scale;

3) protocols for access programs that keep vetted defensive researchers from being locked out by gating;

4) confidential computing that protects sensitive AI assets even while they are in use;

5) tamper-resistant weight and inference hardware that keeps the most sensitive model assets behind security boundaries below the general-purpose software stack;

6) extraction-resistant inference that makes it harder to reconstruct or distill frontier capabilities through repeated model access;

7) detection and response tools for AI attacking its host infrastructure in order to operate outside of the intended environment;

8) AI-specific incident response that can rapidly revoke access, isolate compromised systems, restrict agent capabilities, and contain model or infrastructure compromise;

9) AI threat-sharing infrastructure that allows relevant actors to rapidly share signals of model theft, misuse, etc.;

10) responsible-disclosure norms adapted to a world where AI can compress the time between vulnerability discovery and widespread exploitation.

Closing thoughts

The defense won't need to win every battle. But we also have no time to lose.

We won't need to win every battle

We should, of course, aspire to stop every serious attack. But the defensive strategy cannot depend on winning every battle. Almost every important system has been compromised at some point. The more worrisome scenario involves attacks combining severity, scale, and simultaneity: many critical systems failing together in a correlated fashion, beyond the capacity of defenders to respond, until failures begin cascading into one another.

Infectious disease management provides a useful, if imperfect, analogy. "Flattening the curve" does not mean preventing every infection, but rather slowing the spread enough to keep hospitals from being overwhelmed and buying time to develop medical countermeasures. Cyber resilience may require something similar: measuring where pressure is building, slowing offensive diffusion where relevant, and keeping critical systems functioning while stronger defense capacity is built.

No time to lose

We should not expect the problems discussed in this essay to resolve themselves. Because AI-enabled security capability tends to scale more easily on offense, a defense-favorable path will not emerge by default. Therefore, a defensive edge has to be deliberately built out, deployed, and maintained.

The right response is neither to deny the trajectory nor to surrender to it. It is to take action where it is possible to move the needle.

There is particular urgency in moving now, while the stakes are still comparatively manageable. Progress is rapid, and if we wait for a full-blown crisis to force the issue, we will be left with blunt, reactive measures: the kind of sweeping bans that lock out defenders along with the attackers, imposed under

exactly the conditions in which good judgment is hardest to exercise. The earlier we invest in differential defense, the more room we have to pursue targeted interventions rather than emergency ones.

I remain genuinely optimistic that, on a long enough horizon, AI can make the world far more secure than it is today. But that future is not guaranteed, and it is certainly not automatic either. It has to be built, and the immediate task before us is ensuring the defenders have the capacity needed to hold the line through the upcoming transition period. The clock is ticking.

—

Thanks to Omer Nevo, Alon Oring, Edan Maor, Gil Gekker, Yoni Rozenshein, and many people at Irregular for reviewing drafts of this essay.

1. Some examples include NYU CTF Bench, where GPT-4 started at single-digit solve rates to practically saturated today, and Cyber-Gym, where most models have gone from low single-digits to saturating the benchmark over the span of 18 months.
2. Some argue that all AI is capable of doing is to memorize the tasks: because the benchmarks are public, they are part of the training set. Yet, we're seeing a similar effect on private benchmarks as well.

3. Although worth noting that labs are releasing bespoke cyber versions, e.g. GPT-5.6-Cyber and Gemini 3.5 Flash Cyber.

4. This is a projection, and speculative by nature. The debate on scale gets much hotter once people reason in terms of AGI or superintelligence. But there is no reason to bundle these terms to the point being made here. The argument here is simply that there seems to be more room for scaling in a way that impacts security, regardless of whether we ultimately get to superintelligence.

5. Lockheed Martin's Cyber Kill Chain is an example of this concept. A Framework for Evaluating Emerging Cyberattack Capabilities of AI by Google DeepMind also expands these concepts into an AI context.

6. The 4.7-month doubling time comes with important caveats AISI stresses. It's measured under a deliberately low budget of 2.5M tokens per task, a cap AISI imposes to keep results comparable over time but which it says understates what the models can actually do.

7. Attackers may require significant on-premises compute - but certain attackers (especially nation-state attackers) are already accustomed to paying this "compute tax." Note that attackers do not necessarily need frontier models; they only need models that are good enough for the attacker to find utility in them. This will be the case as long as models are better at offense than defense (or to be more precise, as long as open-weight models are better at offense than frontier models are at defense).

8. Full disclosure: as mentioned above I am affiliated with Irregular (I'm one of the founders). Irregular shared more information on these incidents, and a whitepaper is being developed to deal with such issues in the future.

9. This has already been demonstrated in controlled experiments.

10. There are many overlapping terms for possible milestones along this path, and for its possible destination: AGI, powerful AI, recursive self-improvement, superintelligence, among others. Their definitions differ and are contested. Since this essay is focused on security, I will avoid turning it into a taxonomy; the linked sources provide useful starting points for readers who want to go deeper.

11. The offense-defense concept originates in international-relations scholarship, and was adapted to cybersecurity (see Jervis and Slayton). A reservation is in order: in an AI security context the term is used more loosely than in International Relations/security tradition. In the context of this essay I use it broadly, to capture the relative benefit that attackers and defenders derive from AI.

12. These labels are conceptual, and should be treated as directional rather than precise. E.g., "Offense gains more from AI" and "the world is offense-dominant" are related but distinct claims: in principle AI could confer larger marginal gains on attackers while the world remains in absolute terms easier to defend than to attack (or the reverse).

13. One reason some expect securing a system to cost more than building or attacking it: defense needs a wide context. Attacking a component can be local; securing it means understanding how it interacts with its dependencies, configurations, users, business logic, and the whole surrounding system.

14. Currently, there is a gap between an AI system's ability to rapidly develop software and its ability to secure the software it develops. For example, we have seen models mishandling password security in generated code. It has also been shown that other types of vulnerabilities are likely to be introduced by AI coding agents.

15. This is made worse if an AI system cannot be trusted to be aligned, e.g. if there are bugs in the alignment training code (as was observed years ago during the training of GPT-2), or if the training data was poisoned by a sophisticated attacker, or if the model authors are actively interested in misalignment. The latter is especially a risk in open-weight models, fine-tunes, quantizations, and distillations by anonymous model creators, whose intentions are unknown. Potential intentional misalignment has also been observed in some models: A research report by CrowdStrike showed that DeepSeek-R1 is more likely to introduce security flaws in generated code based on political triggers.

16. The standard process of responsible disclosure, designed to disclose vulnerabilities only when remediation is available, has also started to fail in 2026: Increasingly, there have been cases of vulnerabilities being exposed worldwide instead of being reported privately to the vendor to allow time for patch development. The ease of discovering new vulnerabilities with powerful offensive AI, combined with the slow (by today's standards) processes for patching them (e.g. the 90-day disclosure policy, and "Patch Tuesday", the once-per-month patching cycle used by several companies) may lead to more such cases in the future.

17. Further, we must recognize that an intervention that constrains offense but imposes comparable or greater costs on defense is not worthwhile.

Disclaimer: While Irregular works with major AI developers, no proprietary, confidential, or non-public information from any such organization has been used in the preparation of this content. The positions and views expressed herein belong solely to the authors and do not necessarily represent those of Irregular's clients, partners, or any third party. All content is based on publicly available information, general AI and cybersecurity field knowledge, and independent research by Irregular employees.