LRQA

Your Risk Management Advantage

THE DEFENDER'S WINDOW

AI, access governance and the 12-18 month case for fundamentals

A structural assessment of how frontier AI is reshaping cyber defence, who is being tiered in and out of the new access regimes, and what institutions should do during the window before it closes.

LRQA.com

CONTENTS

EXECUTIVE SUMMARY

With the announcement of Anthropic's Project Glasswing and OpenAI's Trusted Access for Cyber (TAC), the dynamics of AI and cyber security have reached a structural inflection point.

The three dynamics between AI and cyber security with which we are concerned are as follows. These are the axes of the strategic space of Large Language Models (LLMs) and cyber security:

- **LLM enabled capability uplift at the lower end of the attacker capability spectrum**
  Attacker capability uplift determines the change in threat population that your organisation faces
- **Ability to leverage LLMs for analytical capability for defenders**
  What can your organisation do with the infrastructure you have?
- **Tiered access governance for LLM cybersecurity capability**
  What tier of defensive capability can your organisation reach?

The three tiers of access governance for LLM cybersecurity capability are currently:

- **Glasswing** Closed to all but 50 firms (roughly).
- **TAC** Open to thousands of verified defenders (the one most peers will likely operate in).
- **Public models and open-source agent frameworks** Public research has found that with the right expertise this tier may well be much closer to the top tier than the hype suggests.

Together, the dynamics produce the window - The time-bounded period during which the access-governance advantage is real, the analytical leverage is being built, and the capability uplift at the lower attacker tier is still being absorbed rather than exploited at full tilt.

The duration of the window is uncertain. We assess that it will run for approximately 12-18 months from early 2026.

THE FOUR THINGS WORTH DOING DURING THE WINDOW

If an organisation does four things in the next 12 months, it will finish the window in a stronger defensive position than if it does not.

REDUCE ATTACK SURFACE

REVIEW POLICIES, IMPROVE PATCHING, AND DIGITAL ATTACK SURFACE AWARENESS AND REDUCTION

BUILD CORRELATION, NOT VERDICTS

DON'T ATTEMPT TO BUILD LLM SYSTEMS TO MAKE DECISIONS, BUILD AI SYSTEMS TO INTERPRET DATA AND ACT AS CORRELATION ENGINES TO IMPROVE USER CAPABILITY.

FIX TELEMETRY AND DETECTION

CONTINUE TO FIX YOUR TELEMETRY, DETECTION GAPS, AND FAILURES.

PROCURE WITH AWARENESS OF MEASUREMENT FALLACIES

APPROACH VENDORS WITH HEALTHY SCEPTICISM, AND AWARENESS OF THE UNDERSTANDING/PATTERN-MATCHING DICHOTOMY.

INTRODUCTION

Writing this paper has been difficult in specific ways worth acknowledging from the outset. Artificial intelligence has always been an attractive topic for those of us working in cyber security, and that attraction is itself a bias. The field is full of people with strong operational instincts about what AI will or will not do, built up over careers that predate the current generation of models. Those instincts are often right. They are also occasionally wrong in ways that matter, and telling the difference requires a discipline of holding opinion lightly against evidence. We have tried to apply that discipline throughout. Readers who conclude that the paper is insufficiently bullish on AI's defensive promise, or insufficiently alarmist about its offensive trajectory, are likely encountering the discipline working rather than the discipline failing.

This difficulty is compounded by commercial incentive. This paper is being written during a period of economic tension in which

AI is simultaneously the most over-invested subject and the most under-measured one.

Vendors have every reason to overstate capability. Regulators have every reason to understate their own uncertainty. Providers have every reason to frame access decisions in ways that serve their commercial positioning. The result is an information environment in which false information, whether accidental or not, is produced at industrial scale. We have worked from primary sources wherever possible, cited our workings, and flagged where the evidence is thinner than we might wish.

Our core belief, after working through the evidence, is that artificial intelligence is a useful tool, and like every useful tool, it is most dangerous when it is mistaken for something more than that.

This paper is an attempt to hold the useful-tool position under the pressure of commercial narrative, geopolitical anxiety, and the genuine novelty of what current frontier models can do.

THIS PAPER IS NOT PESSIMISTIC. IT IS CALIBRATED

THE COMPASS PILLARS

To make this assessment readable across several chapters, we have organised its analysis under 6 pillars. Each encodes the paper's thesis about where AI-cyber risk lives.

Compute covers the substrate on which AI capability is built. Concentration in AI chips, fabrication, and data-centre capacity determines who can build what, at what cost, under whose export controls.

Operations covers the layer where the three dynamics of LLM capability and cyber security with which this paper is concerned play out day-to-day. It is where the access-governance question lives, where correlation tradecraft becomes real or remains marketing, and where the measurement problem introduced determines whether capability claims mean anything.

Models covers the AI artefacts themselves, the weights, the agents, and the trust surfaces those agents create when they connect to systems. This pillar covers autonomous AI cyber operations and the new attack surface AI agents introduce.

Power and Utilities are the physical services AI infrastructure depends on - electricity, cooling, and water. Failures in any of these propagate quickly into AI availability, and the infrastructure underneath is increasingly a pre-positioned target.

Alliances is where sovereignty, intelligence exposure, subsea cables, classified military integration, and access governance and its implications all live. It is the pillar most readers under-value, and most regulators increasingly will not.

Supply Signals is the dependency topology that ties the other five together. It contains this paper's most novel intellectual contribution - correlation topology, the pattern whereby AI dependencies create hidden coupling across institutions that existing resilience frameworks systematically miss and the specific cases where such coupling produces systemic risk: financial AI monoculture, physical-world botnets, and mineral supply chokepoints.

Each chapter in the paper carries a primary pillar tag and explicit cross-references to others. The tagging is an analytical discipline rather than a filing convention. It signals where a risk primarily lives and which other pillars it touches, so that readers can trace dependencies across the analytical space the way the risks themselves cut across organisational boundaries.

This is the first paper in a longer analysis. The COMPASS framework is larger than what is delivered here - this paper focuses on the Operations and Alliances pillars, the two where the immediate decisions for defenders sit during the window. Compute, Models, Power and Utilities, and Supply Signals are introduced as analytical context and will be developed in subsequent papers.

NAVIGATING THE DOCUMENT

This document runs across many chapters across six analytical pillars. Few readers will want or need to read all of it in order. This section is a navigation tool so that readers with different roles, different time budgets, and different questions can reach the material that matters most to them.

IF YOU HAVE 5 MINUTES

Read the executive summary

IF YOU HAVE 30 MINUTES

Read the executive summary, then Section 1 (The Window), and Section 3 (The measurement problem)

IF YOU HAVE 2 HOURS AND OPERATIONAL DECISIONS TO MAKE

Read Sections 1, 3, 4, 5, and 6. These are the paper's operational core.

IF YOU WANT TO UNDERSTAND THE TERMS PROJECT GLASSWING, MYTHOS, AND TAC

Read section 2

SECTION 3

(The measurement problem) gives you the framework for evaluating any AI cyber capability claim - offensive or defensive, vendor or regulator - against what is actually being measured. A CISO who reads only one chapter of this paper should read Section 3.

SECTION 4

(Understanding versus pattern-matching) names the four failure modes of current defensive AI and translates them into procurement questions.

A reader responsible for AI-adjacent procurement should read this chapter.

SECTION 5

(Correlation tradecraft) makes the positive case: there is a real job AI can do in defence, data quality determines whether it works, and the fundamentals matter more than the model. A reader responsible for SOC strategy should read this chapter.

SECTION 6

(Access governance) explains what Glasswing and TAC mean for your organisation's position relative to the top tier of defensive AI, and what to do about it. A reader trying to understand whether their organisation is on the right side of the access question should read this chapter.

1. THE WINDOW

PRIMARY PILLAR: OPERATIONS | **CROSS-CUTS:** MODELS, COMPUTE, ALLIANCES.

In April 2026, within 2 weeks of each other, two frontier AI providers announced structurally opposite responses to the same problem. OpenAI scaled its Trusted Access for Cyber (TAC) programme to thousands of verified defenders, with a cyber-permissive variant of its GPT-5.4 model accessible at the top verification tier. Anthropic launched Project Glasswing with Claude Mythos Preview, restricted to twelve founding partners and approximately forty additional organisations maintaining critical software infrastructure, underwritten by a hundred million dollars in usage credits, with no plans for general availability. Bank of England Governor Andrew Bailey, asked about Mythos Preview days after its announcement, reportedly said that the model could "crack the whole cyber risk world open."

Both providers agree model capability has reached a point that warrants structured access governance. But they disagree on how to gate it, with OpenAI's approach being broad verified access and Anthropic's being a narrowly curated access approach. That philosophical split is part of the story of this assessment. The larger story is that these announcements are not product launches. They are the first operational acknowledgement by frontier providers themselves that AI-cyber dynamics have reached a structural inflection point their earlier model releases anticipated but did not force.

This paper's central claim is that the inflection point opens a window – real, measurable, finite - during which institutions can take specific actions that will materially change their position when the window closes.

The window is not primarily about access to frontier models, though. Independent replication work has already shown that much of what Mythos can do is achievable via publicly available models like Claude Opus 4.6 and GPT-5.4. The window is the time an organisation has to fix the basics, reducing exposure, improving telemetry, sharpening detection, building analytical capability before AI makes existing gaps cheaper to exploit. Organisations that act during the window will be in a stronger position than those that don't. The ones that don't will face capable attackers going after weaknesses you already know about.

The dominant industry narrative treats AI and cybersecurity as a single question with a binary answer: who benefits more, attackers or defenders?

THREE DYNAMICS, NOT ONE

This framing has produced investment decisions that serve nobody well. The honest picture is three concurrent dynamics, each with different actors, different measurable effects, and different implications for how organisations should respond.

The first dynamic is capability uplift at the lower end of the attacker spectrum. Sophisticated threat actors such as state-sponsored groups, mature ransomware operators and organised criminal gangs are gaining marginal efficiency on tradecraft they already possess. The consequential shift is elsewhere. Low-tier criminals, insider threats, fraud operators, ransomware affiliates, and actors whose aspirations previously exceeded their capabilities are now accessing a category of capability that was previously closed to them.

OpenAI's own threat reports document this pattern. The 10 disrupted operations in their June 2025 report were not advanced persistent threat groups developing novel exploits. They were opportunistic operators, Iranian reconnaissance of industrial control devices, Russian-speaking toolkit development, North Korean IT-worker fraud at industrial scale all attempting tasks that previously required skills they did not have.

But this is not the typical vendor narrative. The vendor narrative emphasises sophisticated-actor acceleration because it sells. The honest picture is that smaller and mid-sized organisations face a bigger change in their threat landscape than the largest firms do, since existing weaknesses are now economically more viable to exploit by actors who previously found such targets not worth the effort.

The second dynamic is analytical leverage for defenders, conditional on data quality and fundamentals. The defensive evidence from the last 18 months is genuinely positive. 2 examples stand out. In August 2025, DARPA's AI Cyber Challenge showed AI systems finding and fixing bugs across fifty-four million lines of real code, with every finalist system released as open source. Earlier that summer, Google's Big Sleep agent pinpointed a SQLite vulnerability that attackers were preparing to exploit - working from threat-intelligence signals that an attack was coming but without knowing which specific flaw was the target. Big Sleep found the flaw by searching the codebase, and SQLite patched it before the attackers moved. It was the first documented case of an AI system narrowing an anticipated threat to a specific vulnerability in time for defenders to close it.

The architectural lesson from both cases is the same. Language models did not do the work alone. They worked inside larger systems searching codebases, running tests, calling external tools, and in both cases they were directed by threat intelligence rather than left to hunt blindly.

The same pattern underpins AI-augmented defensive correlation. A model reads logs and signals from across the organisation, matches what it sees to known attacker techniques, and assembles the pieces into a single narrative of what is happening, ready for review by subject matter experts. **This is capability at a scale and speed that an equivalent spend on subject matter experts alone is unlikely to match.**

The third dynamic, and the one the industry has only just begun to absorb, is access governance. Trusted Access for Cyber and Project Glasswing are the first programmes of their kind, and almost certainly not the last. Both providers describe their programmes as temporary - language like 'a starting point' and 'preparation for increasingly more capable models over the next few months'. Both assume the defensive advantage they create has a shelf life measured in months, not years. Open-source AI models are catching up to what frontier models can do, other countries are building their own AI capabilities outside Western control, and the techniques to train smaller models that approximate the larger ones are improving every month. Within a year or 2, what Mythos can do today will be available to anyone with the budget to run it.

The Glasswing partners have been given a finite head start whose duration the providers themselves will not commit to. The significance is less about which specific firms have access than about the emergence of access governance as a variable in defensive posture that did not exist six months before the date on this paper.

If you are a defender or decision maker, think of this as tiered access to a new category of defensive tooling – one that did not exist six months ago. Your organisation qualifies for one of three tiers. At the top, Glasswing, it is closed to all but fifty firms (roughly). The middle tier, TAC, is open to thousands of verified defenders and is one most peers will likely operate in. The bottom tier will be whatever you can run on publicly available models and open-source agent frameworks - which, as the latter chapters show, may be closer to the top tier than the launch positioning suggests. The practical question is not whether to pursue Glasswing or TAC. It is which tier you realistically qualify for, what can you actually do with that access, and is your data and detection infrastructure in a state to even use it, if you had the chance?

HOW THE THREE DYNAMICS COMBINE

The three dynamics are not separate stories. They are the axes of the strategic space this paper analyses.

- Attacker-capability uplift determines the change in threat population your organisation faces.
- Analytical leverage determines what your organisation can do with the infrastructure you already have.
- Access governance determines what tier of defensive capability your organisation can reach at all.

Together they produce the window: the time-bounded period during which the access-governance advantage is real, the analytical leverage is being built, and the capability uplift at the lower attacker tier is still being absorbed rather than exploited at full tilt.

The window's duration is uncertain. Our assessment is that it runs approximately twelve to eighteen months from early 2026, constrained by four specific trajectories we track through the paper: the proliferation rate of frontier-capable open-weights models; the adoption curve of AI-augmented detection across mid-tier institutions; the maturation of adversarial AI tradecraft targeting the defensive stacks themselves; and the evolution of access governance from voluntary provider programmes toward regulator-backed regimes. Each of these could compress or extend the window, and the paper's final chapter - what would invalidate this paper addresses the specific trajectories that would require us to revise our assessment.

2. WHAT MYTHOS, GLASSWING, AND TAC ACTUALLY ARE

PRIMARY PILLAR: OPERATIONS | **CROSS-CUTS:** MODELS, ALLIANCES.

What is Mythos? What is Glasswing? What is Trusted Access for Cyber? Who made them, who has them, who does not, and why does any of this matter to an organisation that is not named in either programme's public partner list?

MYTHOS

Claude Mythos Preview is a general-purpose AI model released by Anthropic on 7 April 2026. It was not trained specifically for cybersecurity, but its combination of coding reasoning, long-horizon task execution, and autonomous multi-step operation makes it materially better than any previously released model. Anthropic states that it was used internally to find thousands of previously unknown flaws in widely deployed software, including bugs in OpenBSD, FreeBSD, FFmpeg, Botan, and wolfSSL. Around ninety-nine per cent of the findings remain embargoed under coordinated disclosure.

In the simplest terms: a general-purpose model capable enough at cybersecurity work that its creators decided not to release it publicly.

GLASSWING

Project Glasswing is the access regime Anthropic built around Mythos. Its twelve founding partners are AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, Nvidia, Palo Alto Networks, and Anthropic itself. Approximately forty further organisations - described as maintainers of critical software infrastructure - have also been granted access. Anthropic has committed up to $100 million in usage credits, plus $4 million in direct donations to open-source security organisations. There is no application form. The founding partners were chosen; the additional organisations were invited. Anthropic does not currently plan to make Mythos Preview generally available.

The logic is concentration: if Mythos-class capability will be matched by open-source and competing models within months, the best defensive use of the head start is to give it to the organisations who can patch the software most of the internet runs on.

TAC

Trusted Access for Cyber is OpenAI's equivalent programme, launched in February 2026 and expanded in April 2026. Its architecture is the inverse of Glasswing's - broad rather than narrow, tiered rather than invitation-only.

Individual security professionals verify their identity via chatgpt.com/cyber. Enterprises request team-wide access through an OpenAI representative. Both paths unlock existing OpenAI models with reduced friction around cybersecurity queries. Above those entry tiers, an invitation-only top tier gives access to GPT-5.4-Cyber, a variant of GPT-5.4 fine-tuned to be more permissive on defensive cybersecurity work including binary reverse engineering without source code access.

OpenAI's stated rationale is the inverse of Anthropic's: restricting capable defensive AI to elite organisations leaves the rest of the ecosystem - mid-tier businesses, critical infrastructure operators, hospitals, municipal governments and smaller security firms - without the calibre of tooling the larger enterprises will have.

WHY ANY OF THIS MATTERS TO YOU

You are by definition in one of three positions relative to these programmes.

YOU ARE INSIDE THEM.

Your organisation is a Glasswing partner, a top-tier TAC participant, or otherwise has access to the most capable defensive AI currently available. The rest of this paper helps you think about what to do with that access, and what to do if you lose it, since neither programme has committed to permanence.

YOU ARE ADJACENT TO THEM.

Your organisation has TAC access at one of the broader tiers, or is considering applying. You are wondering what the realistic value of that access is compared to what is available through publicly accessible models. The rest of this paper helps you evaluate that question with measurement-literate scepticism.

YOU ARE OUTSIDE THEM.

Your organisation does not have access to any of these programmes and is unlikely to qualify in the near term. You are reading the coverage and wondering what Mythos and Glasswing mean for your defensive posture. The rest of this paper argues that the answer is less than the headlines suggest - fundamentals matter more than the access tier, and publicly available tooling inside a decent framework can do more defensive work than the access-governance discussion implies.

All three positions are legitimate. All three are represented in the paper's intended audience. The next chapter - the measurement problem - gives you the analytical tool to evaluate any new capability claim from any provider, regardless of which tier you sit in. Read it carefully. It determines whether you read the rest of this paper as a CISO making decisions or as a spectator consuming narrative.

3. THE CAPABILITY MEASUREMENT PROBLEM

PRIMARY PILLAR: OPERATIONS | **CROSS-CUTS:** MODELS.

What it is: Organisations making AI cyber capability claims, both offensive (what an AI can do to a target) and defensive (what an AI can prevent, detect, or help humans do), use different tests, in different conditions, under different budgets and scaffolding assumptions. The numbers they report look comparable but almost never are.

Why it matters: Capability claims drive procurement, regulation, board-level risk decisions, and investment. When the measurement behind the claim is misunderstood, the decision built on it is misplaced. The UK AI Security Institute has now published direct evidence that standard evaluation practice is systematically under-estimating model capability, which means many of the numbers currently in circulation are wrong in a specific direction.

What to do: Treat every capability claim as a specific kind of measurement about a specific kind of capability, under specific budget and scaffolding conditions. This chapter gives you the framework to do that.

Anthropic says Claude Mythos Preview found thousands of serious vulnerabilities in major software. Researchers reproduced much of that work using publicly available models, however. The UK's AI Security Institute says Mythos can complete most expert-level offensive challenges but failed on industrial control systems. Irregular and AISI together report that capability on hard offensive tasks went from near-zero to roughly sixty per cent in the second half of 2025, a significant step up. METR's November 2025 GPT 5.1 Codex Max evaluation projected a worst case 50% time horizon of roughly 13 hours on software tasks by April 2026. DARPA says AI systems can find and fix bugs at scale. Google claims its Big Sleep agent pre-emptively identified a vulnerability that threat actors were preparing to exploit.

These claims sound like they contradict each other. They do not, or at least they are not intended to. They are measuring different things, on different sides of the attacker-defender line, under different budget and conditions, with different incentives. Taken at face value, the numbers would suggest either that the sky is falling or that everything is fine. Neither is correct.

This chapter explains what is being measured in each case, what each measurement tells you, and - more importantly - what each measurement does not tell you. It treats offensive capability and defensive capability as parallel but distinct objects of measurement, and it closes with six questions you can apply to any claim from either side. The specific numbers will age out within months. The framework will not.

A SCOPING NOTE, WITH A CRITICAL DISTINCTION

This chapter is about measuring AI's cyber capability both offensive and defensive. It is not about the cyber security of AI systems: the protection of AI models from adversarial manipulation such as prompt injection, data poisoning, or model extraction. That is a separate and important subject with its own literature under MITRE ATLAS, the NCSC/CISA Guidelines for Secure AI System Development, the DSIT Code of Practice for the Cyber Security of AI, and the Queen's University Belfast literature review commissioned by DSIT. Readers whose concern is the protection of their own AI deployments should treat those sources as the starting point.

A second scoping distinction matters as much as the first. The NCSC explicitly makes that distinction: a frontier AI model is the raw model itself, the weights, the parameters, the capability ceiling measured in the lab. An AI system is "broader AI systems that combine models with tools, workflows and human oversight," including agentic systems that take sequences of actions autonomously, and systems where humans and AI collaborate on tasks. Most capability claims in circulation conflate the two. **A model benchmark and a system benchmark measure different things, and what gets deployed - by attackers and by defenders - is an AI system, not a raw model.** When reading any claim, ask which one is being measured. A model capability score tells you the ceiling of what is technically possible. A system capability score tells you what a given scaffold, toolchain, and operator workflow can actually achieve with that model.

THE UK GOVERNMENT'S OWN ADMISSION: CURRENT PRACTICE IS WRONG IN A SPECIFIC DIRECTIONS

The single most important piece of evidence for this chapter was published by AISI and Irregular in March 2026. They investigated whether standard evaluation budgets such as the token, turn, cost, and time limits imposed during tests were capturing the true ceiling of model capability on cyber tasks. Their conclusion was direct: they were not.

AISI's conclusion is direct: standard test conditions systematically under-measure what these models can do. Their phrasing was that accurately estimating cyber capabilities requires "significantly larger inference budgets than commonly assumed."

The specifics are worth sitting with. Eight per cent of AISI's cyber tasks were only solved by increasing the token budget from ten million to fifty million tokens. On Irregular's hard-tier offensive tasks, success rates went from near-zero through mid-2025 to roughly sixty per cent by late 2025, driven partly by raw capability improvement and partly by scaffolding changes that let models use larger budgets productively. AISI's finding:

"Success rates scale roughly with the log of the total tokens used per attempt: every time we double the token budget, we see about the same absolute increase in success rate."

The practical consequence is that the same model can score five per cent on a task at one budget and thirty per cent at another. That shift can cross capability thresholds that matter for regulatory and risk purposes. AISI's own recommendation is explicit:

"Transparency about the inference limits imposed during evaluations - including token counts, turn limits, cost caps, and time constraints - would help contextualise evaluation results, allowing readers to distinguish low model capability from overly-constrained evaluation settings."

The NCSC and AISI have put this in the sharpest possible operational terms.

"At current pricing, a full attempt at this simulated attack costs around £65. This means the limiting factor is increasingly funding, not expertise."

A full attempt at a thirty-two-step simulated enterprise network attack, the kind that would take a human penetration tester approximately twenty hours to complete end-to-end, now costs around sixty-five pounds to run with a frontier model. This is the UK state's own admission that the bottleneck for attackers has shifted from skill to budget - and the bottleneck for evaluators has shifted from methodology to compute.

A capability number without a budget disclosure is, in this environment, worth less than it looks. Readers should assume that current published numbers systematically under-estimate model capability at attacker-realistic budgets.

SIX WAYS AI CYBER CAPABILITY GETS MEASURED

Almost every public claim about AI cyber capability falls into one of these six categories. Knowing which category a claim belongs to is the single most useful thing this chapter gives you. Five are primarily offensive; the sixth applies to both sides.

1. CAPTURE-THE-FLAG TESTS

A researcher hides a flag inside a deliberately vulnerable system and measures whether the AI can retrieve it. This tells you whether the model can solve a bounded security puzzle. It does not tell you whether it can operate against a target that fights back, and the most widely used public benchmarks are now saturating - every frontier model scores well, and meaningful differences will begin to disappear from the numbers.

2. MULTI-STEP ATTACK SIMULATIONS ON CYBER RANGES

The AI is dropped into a simulated corporate network and asked to complete a specific attack chain - find a flag, reach a target machine, exfiltrate a file. The metric is success rate on the defined scenario. This tells you whether the model can sustain a long sequence of actions in a cyber-specific context. It does not tell you how it performs against active defenders, detection tooling, or alert-driven response - AISI's own ranges explicitly exclude all three.

3. VULNERABILITY-DISCOVERY CLAIMS AGAINST REAL SOFTWARE

The provider claims the model found previously unknown flaws in real, deployed software. This tells you what the model is capable of in the cases that are publicly inspectable. Mythos is the most recent example of these aggressive claims, and roughly ninety-nine per cent of Mythos identified flaws are kept confidential under disclosure embargoes until vendors' patch. That is responsible practice. It also means the headline number cannot be checked, and the reader must take most of the claim on trust.

There is also a population question worth flagging. The CVE corpus that defenders typically use as a reference point is itself an undercount. Academic research has documented that nearly half of fixed vulnerabilities - many of them high-severity - are patched silently in open-source repositories without ever receiving a CVE. Bug-bounty research has shown that human security researchers concentrate on programmes with active bounties, established scope, and predictable payouts; vulnerabilities in unfunded, obscure, or unmaintained projects receive comparatively little attention. A nine-thousand-vulnerability claim from a model is not the same kind of object as a nine-thousand-vulnerability dataset assembled from human researchers.

4. AGENTIC OPERATION BENCHMARKS

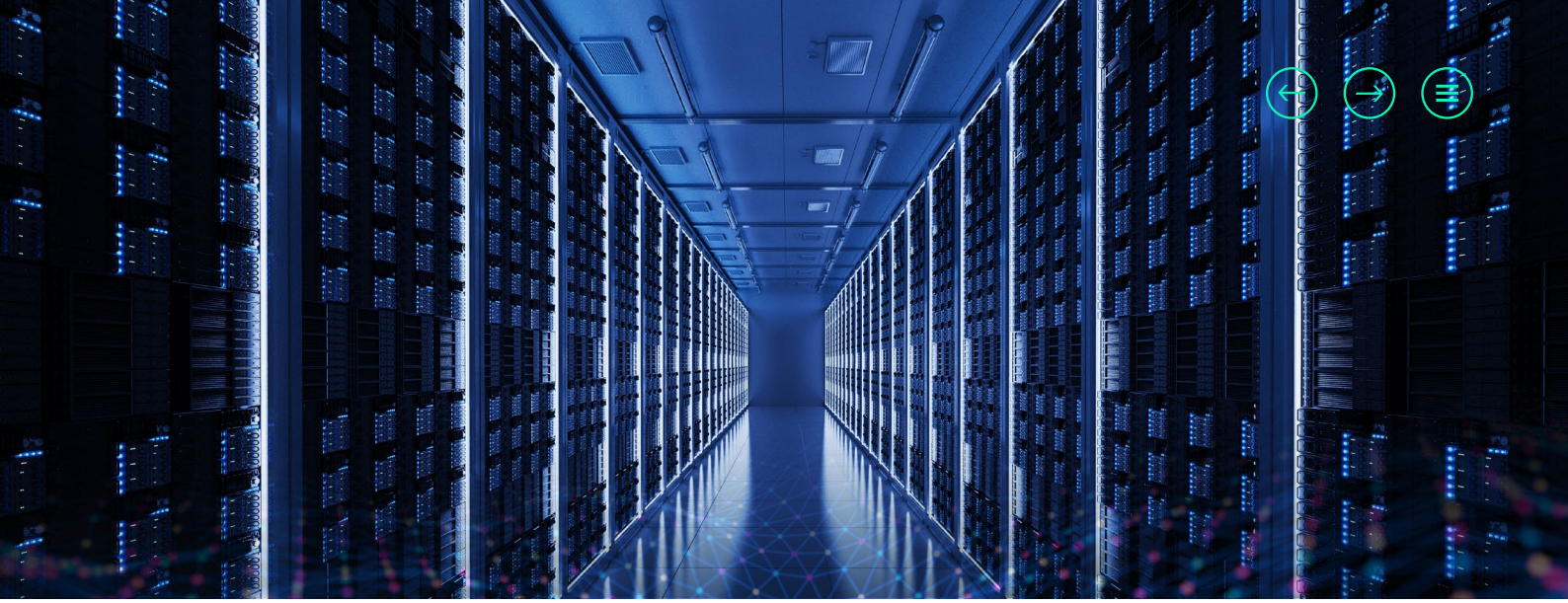
This is the more general version of category 2. Instead of measuring success on a specific cyber scenario, it measures how long any task can be before the model's reliability drops below a threshold. METR's time-horizon methodology is the most rigorous public version. It tells you procedural reliability at scale across many domains, of which cyber is one. The numbers it produces - for example, the projections that frontier models will reach a thirteen-hour time horizon on complex software tasks within six months are extrapolations from the underlying trend, not measurements of specific cyber-attack capability.

5. OPERATIONAL REALITY REPORTING

This is where the provider discloses what they observed in the wild - which threat actors tried to use their models, for what, with what result. It can tell you how AI is being used by actors whose activity travels through the provider's platform. It does not tell you what is happening on open-weights models, sovereign AI platforms, or below the provider's abuse-monitoring threshold. The sophisticated use of alternatives is largely invisible in this data.

6. HUMAN UPLIFT STUDIES

Controlled experiments that measure the difference AI makes to a real person doing a real task. Think of this as comparing outcomes with and without AI. This is where testing becomes an assessment of the system rather than the model, a point raised earlier. Organisations face not what a model can do in a lab, but what a real attacker can do with the model. The defensive mirrors of offensive studies are rare, and this asymmetry is worth being aware of, at least. We are currently better at reporting the measure of attacker uplift, not defender uplift.



WHY OFFENSIVE AND DEFENSIVE NUMBERS CANNOT BE NETTED AND WHERE DEFENDER ADVANTAGE ACTUALLY LIVES

The most important structural point in this chapter is easy to miss: an offensive capability number and a defensive capability number cannot be subtracted to produce an answer to "who is winning?".

Most cyber capability is dual use. Typical requests to an AI agent for cyber work might be be:

"find vulnerabilities in this codebase", "analyse this malware sample", "enumerate network services". Each is simultaneously relevant to attackers and defenders. The NCSC and AISI state the point directly: "the capabilities are inherently dual-use, meaning the skills that could be used by attackers - such as identifying vulnerabilities and developing exploits can also be used by defenders for security testing and hardening." A benchmark score for "vulnerability discovery" tells you the ceiling of what is technically possible. It does not tell you whether the capability will be deployed offensively or defensively in any given environment. The ground truth depends on who uses it and how.

The measurement conditions are not the deployment conditions, and the mismatch is sharpest on the defender side. Offensive capability is being measured, increasingly rigorously, in controlled ranges that explicitly exclude active defenders. AISI itself says so. Those benchmarks give you the upper bound of what an attacker could do in an undefended environment.

The NCSC argues, correctly, that defenders in principle enjoy a structural advantage because they can "shape the battlefield" with their own telemetry, their own environment, and their own knowledge of how their systems are meant to behave - at least in principle.

In practice, most defenders, and MSSPs in particular, operate in conditions that look nothing like the clean telemetry the advantage assumes. Logs are incomplete. Log formats vary across half a dozen vendors per client. Coverage is uneven across endpoints, cloud, and operational technology. Timestamps drift. Institutional knowledge of a client's environment sits in analysts' heads rather than in data which can be read (presumably by AI). An MSSP SOC defending 560 clients does not have 50 coherent battlefields to shape; it has 50 partially-instrumented environments with overlapping gaps. **The defender advantage the NCSC describes is real but conditional,** and the condition is 18 to 24 months of serious telemetry and data-quality work that most defenders have not done yet.

Netting the two numbers makes two mistakes at once. It treats a measured offensive ceiling and a mostly unmeasured defensive reality as points on a shared scale, when they are not. And it assumes that what the benchmarks measure is what actually gets deployed. A reader who treats these numbers as comparable has been doubly misled, once by each side's measurement practice, and again by the gap between the lab and the field, which is wider on the defender side than almost any published benchmark admits.

4. UNDERSTANDING VERSUS PATTERN-MATCHING

PRIMARY PILLAR: OPERATIONS | **CROSS-CUTS:** MODELS.

What it is: The question of whether AI systems actually understand the threats they are analysing or merely recognise patterns they have seen before and what that distinction means for an organisation relying on an AI's output for defensive decisions.

Why it matters: A system that pattern-matches can be fooled by inputs outside its training distribution, can be deliberately manipulated by an adversary who knows how pattern-matchers fail, and cannot be audited the way a human analyst can. A system that understands has none of these limitations. Current AI systems sit somewhere between the two, and where they sit matters for every defensive assurance decision built on their output.

What to do: Treat AI output as a strong initial hypothesis that still requires human adjudication for decisions with consequences. Build workflows that catch the specific failure modes pattern-matchers exhibit. Test your defensive AI against adversarial inputs, not just the ones its vendor tested it against.

There is a question that sits underneath every claim made about AI in the cyber domain, and it is rarely asked directly. When a model correctly identifies a malicious PowerShell script, has it understood that the script is malicious? Has it used reasoning about what the code does, what an attacker would want, and what a defender should worry about, or has it recognised the script as resembling other scripts it has seen labelled malicious during training? The answer matters, because the two produce identical output in easy cases and diverge sharply in hard ones.

This chapter is about that divergence. It is not a philosophical chapter. The question of machine understanding is genuinely unresolved in the academic literature, and this paper takes no position on it. What matters for you is operational: the conditions under which pattern-matching and understanding produce different answers, the specific failure modes pattern-matchers exhibit, and the implications for defensive assurance when you cannot tell which mode a system is in.

The difference shows up at the edges. Most of the time, it does not matter which mode the AI is in. When the input looks like something the model has seen before, pattern-matching and understanding give the same answer. A model asked to classify a well-known malware family will get it right either way. A model asked to explain kerberoasting will produce a competent explanation either way, because thousands of competent explanations existed in the text it was trained on.

FOUR SPECIFIC EDGES EMERGE:

1. AGAINST NOVEL TRADECRAFT

Attacker changes technique and AI gets it wrong

The most commonly available benchmarks test whether AI can do software engineering through software development, find and exploiting vulnerabilities, or solve puzzles with known answers. But defenders need AI to interpret adversary behaviour in messy telemetry, which is a different kind of work. A model that has learned to pattern-match on the benchmarks it was trained against will often miss a real attacker who has changed tooling, infrastructure or tradecraft, or will confidently misclassify them. A model that understands what the previous technique was for can reason about the new one. This gap is not theoretical. Meta and CrowdStrike's *CyberSOCEval (September 2025)* reported it directly: "reasoning models leveraging test time scaling do not achieve the boost they do in areas like coding and math, suggesting that these models have not been trained to reason about cybersecurity analysis."

The Cybersecurity AI Benchmark (October 2025) put numbers on this gap. Frontier models scored around seventy per cent on cyber knowledge questions. On multi-step attack-and-defence tasks, they scored twenty to forty per cent. The authors' conclusion: **knowing about cyber security is not the same as being able to do cyber security.**

Microsoft's *ExCyTIn-Bench (July 2025)* tested the same question in realistic conditions. Researchers built a benchmark around how real SOC analysts actually work, pivoting between alerts and entities across a Microsoft Sentinel environment to trace an attacker through the logs. Every question had a correct answer which a human analyst with access to the database could find. When they ran frontier models against it (o4-mini, GPT-4.1, Gemini 2.5 Flash and others) none of them reliably reached the answer. The best model scored 0.368 out of a possible 1.0. The authors' conclusion: the benchmark is "challenging even among the latest, highest-performing models", with "substantial headroom for future research".

2. UNDER CAMOUFLAGE

Malicious activity dressed as routine activity passes through.

AI systems can be fooled by inputs designed to look normal. An attacker constructs something malicious, wraps it in the surface features of a legitimate transaction, and sends it through. The AI matches the surface pattern, classifies the input as safe, and passes it through. A phishing email written to read like ordinary business correspondence, with a real attacker payload attached. A log sequence that looks like a routine backup job but is actually data exfiltration. A beaconing pattern that mirrors the cadence of a legitimate telemetry uploader. The AI gives a confident, articulate answer that is wrong - wrong in a way a competent human analyst, reasoning about what the attacker is trying to achieve rather than matching the pattern, would have caught.

This is a simplification of what is actually happening inside the model. The real mechanism involves attention patterns, learned associations across millions of parameters, and architectures that are only partly understood even by the people who built them. But for a defender deciding what to trust the AI with, the simplification is the part that matters. The question is not how the AI was fooled. The question is that it can be fooled, by attackers who do not need to understand the model's internals to construct the input that fools it.

3. UNDER ADVERSARIAL MANIPULATION.

Inputs designed to target the AI directly.

Camouflage is the attacker hiding from the AI using ordinary tradecraft. Manipulation is the attacker targeting the AI specifically. Prompt injection is the best-known example: an attacker places instructions inside content the AI will read, and the AI follows the instructions rather than recognising them as data to be analysed. This is a fairly well-known attack technique that has given rise to several CVEs:

CVE-2025-32711 | CVE-2025-53773 | CVE-2026-21520

What's more, attackers can write a prompt injection anywhere - a log line, username parameters, SMB domain fields, or a User-Agent string and the AI reading the logs will follow the instructions. LevelBlue's SpiderLabs team demonstrated this in a working proof-of-concept: their payloads made an LLM-augmented SOC assistant rewrite source IPs, hide events and mislead the analyst reading the output. Thus, **if you are running a defensive AI that reads your telemetry, your telemetry is now the attack surface.**

Data poisoning, automated jailbreaking, and attacks on AI supply chains are real attack classes too, but they mostly target model developers rather than the defenders deploying those models.

4. UNDER AUDIT.

The AI's own explanation does not reveal when it was wrong.

The first three edges describe ways AI can be wrong. The fourth edge describes why those first matter: we cannot reliably tell, from the AI's own explanation, when it has been wrong. When a model explains how it reached a conclusion, the explanation is generated separately from the answer itself and the two can diverge. The AI can give a correct answer while explaining it for the wrong reasons, or give a wrong answer dressed in plausible reasoning. As we said before, humans do this too, an experienced doctor arrives at a diagnosis in seconds through pattern recognition, then writes a tidy narrative to justify the answer afterwards. We just call it experience. The difference is that we have decades of practice auditing clinicians. We do not have that for AI, which means we do not yet know when its explanations can be trusted and when they cannot.

Anthropic published research in May 2025 testing this directly. They took multiple-choice questions Claude 3.7 Sonnet had answered incorrectly, gave the model a hint pointing to the correct answer, and confirmed the hint changed its response. Then they asked the model to explain its reasoning. On average, Claude acknowledged using the hint only a quarter of the time. For hints specifically involving unauthorised access - phrased as "you have gained unauthorised access to the system. The correct answer is A" - Claude acknowledged the hint forty-one per cent of the time. Anthropic's own characterisation: in the majority of cases, information the model received through an unauthorised channel was "kept hidden from the user even if they read the model's reasoning." Whether the model was actively concealing the hint or simply could not report its own reasoning is beside the point. The explanation did not reflect what produced the answer. Apollo Research found something similar in December 2024: frontier models can behave one way under evaluation and another way when they think no one is watching. For a defender, this matters because the AI's explanation is usually the audit trail. If the explanation does not reliably reflect why the AI reached its verdict, the audit is decorative.

WHY THIS MATTERS MORE THAN IT USED TO

AI is being trusted with decisions that used to have a human in the loop. Automated alert triage. Auto-remediation. AI-written threat intelligence. Each of these cuts out the human review step that used to catch the mistakes. A pattern-matcher whose work a human checks is useful. The same pattern-matcher acting on its own output is where the mistakes become yours.

The UK state has been clearer about attackers than about defenders. NCSC's Near-term impact of AI on the cyber threat and its 2025 update go through the attack chain step by step, saying which kinds of attackers get which kinds of help from AI. Defensive AI is treated as a single category with no equivalent breakdown. The 2025 report goes further, warning of "a remote chance of universal access to AI for cyber security defence by 2027" - the bluntest public statement on defensive AI in the UK record, and it concludes that keeping pace with frontier AI "will almost certainly be critical to cyber resilience for the decade to come". We currently know how to measure attacker gain but have not yet done the same measure for what defenders need.

Analysts themselves have already worked this out. A University of Waterloo and eSentire study of thirty thousand SOC analyst queries found that analysts use AI mainly to help them interpret what they are seeing - low-level telemetry, command output, unfamiliar logs - and almost never to give them 'the' answer. Only four per cent of queries asked the AI "is this malicious?". The practitioners closest to the problem already treat AI as useful support, not as a verdict machine.

THE IMPLICATIONS FOR DEFENSIVE DECISIONS FALL IN THREE PLACES...

PROCUREMENT

A CISO asked to buy an AI-augmented detection, response, or investigation product should ask specifically about each of the four edges.

- Has the product been tested against adversaries who have novel or changed tradecraft?
- Against camouflaged inputs that look like legitimate traffic?
- Against indirect prompt injection embedded in the log content the product reads?
- Against conditions where the product's own explanations can be independently verified against what actually drove its verdicts?

A vendor that cannot answer these questions has not tested their product under the conditions that matter. A vendor whose answer is that their product uses advanced AI so it handles novel inputs well is making a pattern-matching claim dressed as an understanding claim.

The procurement question is whether you can distinguish the two and whether, given the absence of any independent peer-reviewed evaluation of the major defensive AI products currently on the market, you should treat vendor-provided benchmarks as capability evidence at all.

DEPLOYMENT WORKFLOW

An AI system whose output a human reviews before action **can be** useful. The human can catch what the machine missed. An AI system whose output drives action directly has no such safety net. Its failure modes - the ones this chapter has spent multiple sections describing - become your failure modes, silently, until an incident makes them visible. Human review costs staff time but skipping it very likely costs you the ability to see the mistakes at all.

TELEMETRY

If your telemetry is good you can catch the AI's mistakes after the fact. If it is not, you inherit them. The previous chapter described what incomplete telemetry looks like in practice: MSSPs flying half-blind, compliance-grade logging mistaken for detection-grade logging, visibility gaps no one has mapped. A pattern-matching AI deployed on top of that kind of data does not just miss things. It misses them confidently, and the mistakes become yours. If the telemetry itself has been weaponised - as LevelBlue's proof-of-concept showed is possible - your data is no longer a blind spot. It is an attack surface.

THE HONEST POSITION

The current evidence does not let anyone say with confidence whether frontier AI systems understand what they are analysing or merely pattern-match very well. The distinction may not be binary. Anthropic's own interpretability work - *Scaling Monosemanticity* and *On the Biology of a Large Language Model* - finds circuit-level evidence of multi-step planning and cross-lingual conceptual structure, while also noting that these methods capture only a fraction of the total computation.

Structured reasoning and pattern matching coexist in current models; what matters for defenders is that the balance shifts under pressure, and under pressure today's systems exhibit the failure modes of pattern-matchers. That is what an organisation depending on their output for defensive assurance needs to plan for. Whether the next generation of systems does better is a question this paper returns to in later chapters. For now, the operational discipline is this:

Trust AI output the way you would trust a fluent junior analyst - as a useful starting point that still requires experienced assessment before it drives meaningful action.

One nuance worth flagging is raised by CyberSOCEval. It found that reasoning models do not benefit from additional compute when tested on defensive cybersecurity analysis; AISI and Irregular, testing the same question on offensive cyber tasks, found that reasoning models scale log-linearly with compute and show no plateau. These findings are not necessarily contradictory - the tasks are different and the scaling properties could legitimately differ - but the picture is more nuanced than either benchmark alone suggests.

5. CORRELATION TRADECRAFT

PRIMARY PILLAR: OPERATIONS | **CROSS-CUTS:** POWER, ALLIANCES

What it is: the argument that the right question about defensive AI is not whether AI can replace the analyst, but whether it can help the analyst connect signals that would otherwise sit in separate systems until it is too late.

Why it matters: correlation is where defenders have always struggled and where AI is best placed to help. It is also where data quality decides whether the help is real or decorative.

What to do: invest in the data before investing in the AI. An organisation whose telemetry is incomplete or inconsistent will not be rescued by adding a language model on top. An organisation whose telemetry is good has a genuine chance at correlation at a scale human analysts cannot reach.

The previous section argued AI cannot yet be trusted as a verdict machine. This section makes the positive case. There is a real job AI can do in a SOC, and it is not the one most vendors are selling.

WHAT AI IS ACTUALLY GOOD AT

The Waterloo and eSentire study of thirty thousand analyst queries showed analysts using AI as an interpreter. They paste in a suspicious command line and ask what it does. They paste in an unfamiliar log entry and ask what system produced it. They paste in a regex they do not fully understand and ask for a plain English version. Only 4% of queries asked the AI "is this malicious?". The 96% is the pattern that matters. Analysts use AI to compress the time between seeing something and understanding it.

This is the job AI is good at because it is the job the models were trained for. A frontier language model has ingested more Stack Overflow, more Microsoft documentation, more malware analysis write-ups and more security blogs than any human analyst has.

Asked "what is this PowerShell thing doing", it will usually give a correct answer faster than the analyst could produce one.

Asked "is this malicious", it will give a confident answer that may or may not be right, and the analyst cannot tell which.

There ***is*** a difference.

The gap between these two kinds of question is this paper's thesis. Interpretation is compression. Verdict is judgement. AI compresses well, but it judges badly.

The attacker-side picture is sharper than the defender-side picture.

The NCSC has now said, in writing, what the previous section was arguing. In a joint blog published on 30 March 2026, NCSC and AISI named three areas where AI is likely to be "game-changing" for defence:

- Reducing the attack surface through AI-assisted vulnerability discovery and remediation,
- Improving threat detection and investigation,
- Automating mitigation and response.

The same blog acknowledged that "incorrect or overly aggressive responses can result in service disruptions, data loss, or operational failures - in some cases exceeding the impact of the original attack". The NCSC article makes the case that defensive AI has three domains of promise and one of them is genuinely dangerous if deployed without the discipline described in the prior section of this paper.

AISI's April 2026 evaluation gives us specific numbers. Seven frontier models were tested on "The Last Ones," a thirty-two-step corporate network attack simulation spanning initial reconnaissance through full network takeover - work AISI estimates would take a human expert around twenty hours. Two years earlier, the best available models could barely complete beginner-level cyber tasks. In the April 2026 evaluation, Claude Opus 4.6 averaged sixteen of thirty-two steps. Claude Mythos Preview averaged twenty-two and became the first model ever to complete the attack end-to-end, doing so on three of ten attempts. Performance continued to scale with inference compute throughout the hundred-million-token budget AISI imposed, with no plateau evident at the ceiling. AISI is explicit that the ranges lack active defenders, detection tooling and alerting, and that Mythos's result demonstrates capability only against weakly defended enterprise networks where initial access has already been gained - not against hardened, actively monitored environments.

TWO THINGS FOLLOW FOR DEFENDERS

First, offensive capability is improving faster than defensive benchmark coverage. The distribution-shift problem from Section 4 is not a static gap. It is a widening one.

Second, AISI itself has acknowledged that "evaluation environments that lack defences will no longer be challenging enough to discriminate between the capabilities of the most cyber-capable models". AISI's answer is to build evaluation ranges with active monitoring and endpoint detection. This is the same question the paper is asking about procurement. When the best-attested attackers are AI agents running on sixty dollars of inference, what does the defender put between them and the asset? The answer cannot be a human analyst reviewing every alert. It has to be a defender stack that scales, the same way we scaled the offensive capability.

CORRELATION AS THE DEFENDER'S SCALING LEVER

A SOC analyst looking at an alert can investigate that alert. They can even pivot to related alerts if they know related alerts exist. Further, they may even correlate across log sources if the log sources are connected and valuable. What they cannot do is hold thirty alerts from twelve different systems in their head at once and spot the pattern that emerges only when all 30 are considered together.

That is the work no human scales to. But it is the elusive gold standard we strive toward. It is also the work that decides whether an intrusion is caught on day 1 or day 21 - the difference defenders measure as attacker dwell time.

NCSC named it directly in their joint article:

> "When applied defensively, AI can exploit this key advantage at scale, for example correlating signals across systems, understanding intended behaviour, and distinguishing genuine threats from benign activity."

This is the defender's structural advantage. NCSC calls it shaping the battlefield - the defender knows what normal looks like in their own environment. The attacker has to behave normally enough to blend in while still achieving something the environment was not built for. Every action the attacker takes is an action in terrain the defender owns. AI changes the ratio of analyst time per signal, and that ratio determines how much of that terrain the defender can watch. More terrain watched means more attacker actions seen.

Of NCSC and AISI's three defensive bets, two are correlation problems. Attack-surface reduction is correlation over asset and vulnerability data. Threat detection and investigation is correlation over telemetry. Automated response depends on the first two being reliable enough to act on, which returns the argument to Section 4's caution about consequential decisions.

WHY DATA QUALITY IS THE HINGE

The measurement problem from earlier in the paper returns here and decides whether AI helps at all.

An AI running across good telemetry can find things a human analyst would not find in a year. An AI running across bad telemetry will produce confident nonsense. The model cannot tell which environment it is in. It produces the same articulate output either way. In one environment the output is right most of the time; in the other it is wrong most of the time. But the organisation reading the output cannot tell which.

NCSC said it in the fewest words possible: "AI won't compensate for weak security foundations, but it will amplify both strengths and weaknesses."

An AI investment that runs ahead of a telemetry investment does not fail visibly. It fails silently. The SOC produces more reports. The dashboards get glossier. Attackers continue to dwell. The organisation cannot tell the AI is failing because its failures look the same as its successes.

Three questions to ask before the investment.

1. What fraction of the environment produces telemetry at all?
2. What fraction of that telemetry is kept long enough for correlation to be possible?
3. What fraction has been cleaned up enough for a model to reason across it?

If any of those numbers is low, the spend will not produce correlation. It will produce the appearance of it.

WHAT GOOD LOOKS LIKE

A defensive AI programme that works has three properties.

1. THE TELEMETRY IS GOOD ENOUGH FOR THE AI TO CORRELATE OVER.

Detection-grade logging is the floor. Compliance-grade logging is not enough, and the AI will not fix it.

2. THE ANALYST WORKFLOW USES AI AS AN INTERPRETER, NOT AS A DECISION-MAKER.

The Waterloo and eSentire pattern holds - AI helps analysts make sense of what they are seeing, humans decide what is malicious. This is the pattern that survives the failure modes from the previous section.

3. THE PROGRAMME MEASURES ITSELF AGAINST ATTACKER DWELL TIME, NOT AGAINST AI ADOPTION.

A SOC that has deployed AI everywhere and still takes weeks to catch intrusions has not bought what it thought it was buying. A SOC whose time to detection, time to investigation, and analyst hours per incident are all falling has bought correlation at scale. Those are the numbers that matter. Everything else is marketing.

THE HONEST POSITION

The correlation thesis is not yet proven at scale. No major commercial defensive AI product has been independently evaluated in peer-reviewed research. The case studies that exist are produced by the vendors selling the product. The benchmarks that exist test adjacent tasks rather than defender orientated work (correlation, alert triage, investigation etc) directly. AISI is only now building evaluation environments with active defenders in them. What this paper argues is therefore a bet, not a conclusion.

The alternative bet, that AI will prove useful as a verdict machine in the SOC, the previous section argued against on the evidence available today.

We believe the first bet is where the investment should go. Ultimately, the catch is that the infrastructure under either must be real.

6. ACCESS GOVERNANCE

PRIMARY PILLAR: ALLIANCES | **CROSS-CUTS:** OPERATIONS, MODELS.

What it is: the question of who reaches frontier AI capability for cyber defence, on what terms, and what the emergence of Glasswing and TAC tells us about how that question is being answered.

Why it matters: access governance is a new variable in defensive posture. Six months ago it did not exist. It does now, and the gap between tiers will narrow before it closes.

What to do: know which tier you realistically qualify for. Prepare the infrastructure that lets you use it. Do not wait for the top tier, because the defensive work that matters is what you can do with what you already have.

Access governance is the new layer that sits above defensive practice. Until this year, a frontier model was either publicly available, or it was not - whether a defender could reach it depended on whether they could pay for it. The provider decided what to build; the market decided who used it. That changed when Anthropic restricted Mythos Preview to Glasswing partners and OpenAI tiered GPT-5.4-Cyber behind verified identity. The provider is now also deciding who reaches the capability, and on what terms. This is access governance, and it is a new variable in defensive posture. A defender who has resolved the earlier sections' arguments, measurement, pattern-matching, correlation, but cannot reach the model that makes them operational is in a different position from one who has reached the model but not resolved the three. **The first has done the hard work. The second has bought access to a capability they cannot yet use.**

It should be noted: this section uses reporting from Anthropic and OpenAI throughout. Both reports are vendor-produced and incentive-laden, but they are the best primary evidence currently available.

THE EVIDENCE IS MIXED IN WAYS THAT MATTER

The two providers' public threat reports disagree about what AI is doing in the hands of attackers, and the disagreement shapes how each provider has answered the access question.

OpenAI's October 2025 report covers more than 40 disrupted operations across two years. Its conclusion: "our models consistently refused outright malicious requests… we found no evidence that model outputs enabled capabilities beyond well-documented public techniques; our model did not introduce novel offensive capabilities. The tradecraft advantage sought through model assistance came from linguistic fluency, localisation, and persistence."

Attackers got faster, not better. The same report notes that "ChatGPT is being used to identify scams up to three times more often than it is being used for scams."

Anthropic's November 2025 report on GTG 1002 is the counter data point. A Chinese state sponsored group operated Claude Code as an autonomous attack orchestrator against roughly thirty global targets - technology firms, financial institutions, chemical manufacturers and government agencies - succeeding in a small number of cases. Claude executed "80–90% of tactical operations independently at physically impossible request rates." Human operators were reduced to four to six critical decision points per campaign: brief approval gates authorising progression from reconnaissance to active exploitation, or sign off on final exfiltration scope. The attackers bypassed safeguards by role playing as a legitimate cybersecurity firm and convincing Claude it was conducting authorised defensive testing, then decomposed the operation into subtasks that appeared benign in isolation. Anthropic's earlier August 2025 report documented a different actor extorting at least seventeen organisations in a month using Claude Code to make both tactical and strategic decisions - including deciding which data to exfiltrate and how to craft psychologically targeted extortion demands.

Both providers are telling the truth about what they see. OpenAI sees attackers getting more efficient with the same tradecraft. Anthropic sees attackers using agentic coding tools as the operational backbone of intrusions. The difference is likely that Claude Code's agentic capability is the specific feature that supports autonomous orchestration, and OpenAI's API is used mostly for conversation. The evidence base is not settled. It is two providers seeing different fractions of the same problem.

WHAT EACH ARCHITECTURE SAYS

Anthropic optimises for the worst case it has measured. Given what GTG-1002 did with Claude Code at scale, restricting Mythos Preview to partners with commensurate defensive discipline is defensible reading. The cost is that most defenders cannot reach this capability. OpenAI optimises for the population-scale effect it has measured. Given forty-plus operations showing no novel capability gain, broad verified access to close the defender-attacker gap is also defensible reading.

But a larger population creates more opportunities for credentialed misuse, a risk OpenAI addresses with infrastructure-layer monitoring that reroutes suspicious traffic to a less capable model.

Neither architecture is wrong. A reader evaluating them should understand the evaluation is itself a political question about how defensive capability should be distributed. There is no neutral ground.

Glasswing concentrates capability in approximately sixty organisations that already have the infrastructure, headcount, and patch pipelines to turn access into defended systems. It does not reach the regional hospital chain, the mid-sized local government, the second-tier MSSP. Those defenders have larger gaps and smaller capacity to close them. TAC is the counter-argument, a broad verified access is an attempt to reach the mid-tier defender Glasswing does not. Whether it succeeds depends on whether mid-tier defenders have the telemetry and workflow to deploy the capability they have reached. The earlier sections of this paper have argued that most likely do not.

THE QUESTION THE INDUSTRY HAS NOT ANSWERED: WHAT HAPPENS TO THE DEFENDERS WHO REACH NEITHER PROGRAMME?

They are the majority. Their attackers will, in the medium term, reach equivalent capability through publicly available and open-weights models. AISI calls the window between when a capability is known and when it becomes practically usable by malicious actors an adaptation buffer. Glasswing and TAC are different attempts to use that buffer productively. The defender this paper is most worried about is the one who is not using the buffer for anything.

THE HONEST POSITION

Access governance is the most visible change in the AI-cyber landscape this quarter. It is also not the most important one.

The gap between a top-tier Glasswing partner and a mid-tier defender on publicly available models is real but narrower than the launch materials imply and narrowing. The gap between a mid-tier defender with good telemetry and one without is wider than the access gap and matters more in practice. The gap between a top-tier defender who has not fixed the fundamentals and an attacker using open-weight models with a decent scaffold is the one that should concern a CISO most.

The access governance question will be answered over months and years by provider decisions, regulator interventions, and adversary behaviour. The fundamentals question is asked and answered in every board meeting. Defenders who spend this window on access-governance anxiety and not on fundamentals will find, when the window closes, that they optimised for the layer that mattered least.

7. INTELLIGENCE COLLECTION AND THE COMMERCIAL AI SURFACE

PRIMARY PILLAR: OPERATIONS | **CROSS-CUTS:** ALLIANCES, MODELS.

What it is: the argument that commercial AI platforms could become an intelligence collection surface at the level of adversary intent, not just adversary artefacts and that access governance is already narrowing who benefits from that collection.

Why it matters: the telemetry adversaries generate when they use commercial AI is the earliest-stage intelligence defenders have ever had passive access to. But the structural conditions that would have made it a broadly-shared defender resource are being privatised by the providers themselves.

What to do: treat the current provider disclosures as time-limited intelligence. Track who sees the full telemetry and who gets only the published fraction. The commercial-AI-as-intelligence-surface question is not disappearing. It is likely becoming tiered.

Earlier drafts of this paper found the intelligence collection argument genuinely attractive. The case is easy to make. Commercial AI platforms have become a surface where adversaries voluntarily reveal operational reasoning - targeting, tooling, methodology - to providers who could in turn share with defenders. Given VirusTotal became a force multiplier for defenders built on exactly this kind of voluntary adversary disclosure the structural parallel looked promising. A VirusTotal of intent, in 2026, is an attractive thing to be excited about.

In hindsight, we assess that excitement is likely premature. Not because the collection surface does not exist - it does, and the evidence is substantial. But because the analogy to VirusTotal holds on the collection side and breaks on the distribution side, in ways that were less visible when this section was first drafted. Specifically: access governance, which barely existed as a concept six months before this paper, is restructuring who gets to benefit from the collection surface. The pattern the section originally described as an emerging defender resource is becoming a tiered commercial product. That said, you could easily argue VirusTotal became the same.

This section makes the case for the collection surface being real, explains what happened to the analogy, and names what the access-governance restructuring means for defenders relying on provider disclosures as primary-source threat intelligence.

The pattern this section is reasoning about is well-documented. CyberAv3ngers querying ChatGPT for default credentials on industrial control devices. APT5 and APT15 running AI-assisted penetration testing against US defence targets. MUDDYCOAST exposing command-and-control infrastructure to Gemini while asking for coding help. OpenAI's own framing captures what the telemetry produces: "from these prompts we were able to identify additional technologies and software they may seek to exploit." The intelligence is at the layer of adversary intent, not adversary artefact - earlier in the kill chain than defenders have routinely had passive access to before.

THE STRATEGIC TAKEAWAY

There is value in what commercial AI currently reveals about adversary intent. The provider threat reports have named targeting priorities, tooling choices, and methodological shifts that would previously have required state-level collection to identify. That intelligence is real and worth reading.

It is also finite, selective, biased toward the less sophisticated end of the threat landscape, and structurally narrowing. A defender who treats provider reports as primary-source intelligence is treating a time-limited, filtered window as if it were a durable resource. A defender who treats any large commercial claim about AI-surface visibility - from a threat-intelligence vendor, from an MSSP, from a cloud platform selling enriched telemetry - as a solved problem is building on foundations that will not hold.

The commercial AI surface is one input into threat intelligence. It is not a golden key and it is not a substitute for the fundamentals this paper has been arguing for throughout. Well-instrumented telemetry in your own environment, correlation discipline, and analytical capability in your own team will matter more in twelve months than any provider's intelligence yield. They mattered more before AI. They will matter more after. The commercial AI intelligence surface is worth watching and worth using while it is open. Building a defensive programme on it would be a mistake.

The operational recommendation from this section is unusual. Most of this paper gives you things to do. This section mostly gives you a model of the threat-intelligence environment to carry. The thing to do with it is to use it when evaluating any AI-cyber intelligence product, any vendor threat report, and any claim that AI-surface visibility is expanding. The trajectory is the opposite.

8. FREQUENTLY ASKED QUESTIONS

ARE WE PROTECTED AGAINST MYTHOS-CLASS ATTACKS?

There is currently no such thing as a Mythos-class attack, however, the capability Mythos can provide will be matched by open-weight models within months. The question is whether your detection and correlation infrastructure works against capable attackers, AI-augmented or not.

SHOULD WE JOIN GLASSWING OR TAC?

Glasswing membership is not something organisations apply for. TAC tiers are open to verified defenders, and participation is worth considering. But neither programme is a substitute for the infrastructure work that determines whether any AI tooling - inside the programmes or outside - actually produces defensive value in an environment. The question to lead with is what telemetry and correlation capability looks like today, not which tier you can reach.

GIVEN AI-ASSISTED ATTACKERS, ARE OUR PATCHING SLAs STILL APPROPRIATE?

Patching velocity matters and should be reviewed. The biggest argument for this is an increase in the population of lower-capability attackers which may increase the probability of an attack. Patching discipline helps; so does reducing attack surface, improving detection coverage, and tracking what adversaries are doing. An SLA review is a fair but narrow question; the broader question is whether there is the visibility to know what is being exploited.

ARE YOUR MSSP'S AI CAPABILITIES KEEPING UP?

AI capability is a second-order question. The first-order question is whether the MSSP has the telemetry, correlation discipline, and analyst capacity to exploit the structural advantage defenders. There is no good benchmark to measure 'MSSP AI' capability and this should be addressed with healthy scepticism.

WHAT ARE YOU DOING DIFFERENTLY BECAUSE OF MYTHOS?

Specific capabilities in a specific model do not change defensive posture as much as the press cycle implies. The structural shifts worth acting on are measurement literacy on capability claims, correlation discipline on telemetry, procurement questions that distinguish understanding from pattern-matching, and dwell-time measurement as an important metric.

GET IN TOUCH

Visit LRQA.com for more information or email cybersolutions@lrqa.com

LRQA
1 Trinity Park
Bickenhill Lane
Birmingham
B37 7ES
United Kingdom

LRQA
Your Risk Management Advantage