AI is both the Cure and the Disease

How AI is strengthening cybersecurity defences and manufacturing entirely new categories of risk at the same time.

nextgen
CYBERQUEST
NETALERT

Hani Darouich, CISSP
Detection Engineering Lead, Nextgen Software

Table of contents

This is Not a Think Piece about AI risk

The standard pitch goes like this. AI helps defenders work faster, detect more, close the talent gap. And that is true. But it is half the story and the half people keep leaving out is the one that will cost you.

What the research from early 2026 actually shows, from **Gartner, Anthropic, Deloitte, the International AI Safety Report, RAND and from people doing real detection work in real SOCs,** is that **AI is creating net-new categories of risk. Not theoretical risk. Not "what if" scenarios for conference panels. Production risk. Already exploited. Already measurable.** Showing up in codebases, supply chains, and security operations **right now**.

This paper is a reference for anyone trying to make **honest decisions about AI in cybersecurity**. It is built from current research, sourced throughout and written from the perspective of someone who builds detection logic for a living and sees both sides of this every day. The argument is simple:

AI is a force multiplier. A multiplier needs something real to multiply. Without expert-built foundations, AI multiplies noise, hallucinations and false confidence. With them, it genuinely changes what a lean security team can do.

SECTION 1
AI is manufacturing its own attack surface

• Researchers from UT San Antonio, Virginia Tech and the University of Oklahoma analysed **576,000 code samples across 16 large language models**. Roughly one in five recommended software packages do not exist. **They are hallucinated.** Made up by the model, presented with full confidence.

UT San Antonio, Virginia Tech, University of Oklahoma joint research, 2025. Covered by CSO Online, Feb 2026.

That finding is the basis of a supply chain attack called **slopsquatting**. It works like this: **AI coding tools recommend fake packages**. Attackers register those names on PyPI or npm and stuff them with malicious payloads. Developers trust the AI output, install the package, and the malicious code enters their CI/CD pipeline. Security researcher Bar Lanyado at Lasso Security noticed that LLMs kept hallucinating a Python package called huggingface-cli. The real Hugging Face platform exists, but that specific package did not. He registered it on PyPI as an empty test. Within three months it had 30,000 downloads. Engineering teams at Alibaba had already copy-pasted the AI-generated install instructions into their own public documentation, creating a cascading supply chain risk from a single hallucination.

Bar Lanyado, Lasso Security. Covered by Socket Security, Check Point, and Infosecurity Magazine, 2025.

The hallucinations are predictable.

43% of fabricated names recur across similar prompts. 58% reappear within ten tries. Open-source models like DeepSeek and WizardCoder hallucinated at 21.7% on average, while commercial models like GPT-4 sat at 5.2%.

That makes them weaponisable at scale. An attacker does not need access to your prompts. They just need to ask the same model the same question and watch what it invents.

The point here is not that AI coding tools are useless. They are not. The point is that the tool designed to speed up development is now generating the supply chain risk. The attacker did not create the vulnerability. The AI did.

If your detection rules, correlation logic or response playbooks are built on AI-generated code that nobody validated, the pipeline is compromised before a threat actor touches it. Treat AI code output like an unvetted third-party library. **Scrutinize it. Test it. Have a human sign off before it goes anywhere that matters.**

19.7% of packages recommended by LLMs in testing were completely non-existent - 205,000 fake packages total.

30,000 downloads of a single hallucinated package in three months, including by teams at Alibaba.

43% of hallucinated package names recur consistently, making them predictable targets for attackers

CYBERQUEST SIEM detection rules are written and reviewed by engineers who work on real incidents, not generated by AI pipelines. Every rule is tested against validated threat data before it reaches a production environment. That is how you keep hallucinated logic out of your detection stack. nextgensoftware.eu

SECTION 2
Your AI security tool is hallucinating threats

Measured hallucination rates for AI on domain-specific security tasks sit **between 17% and 33%**. That means roughly one in every three to six outputs from your AI security tool could be fabricated. Not wrong in a nuanced way. **Fabricated. Invented. This is not limited to cybersecurity.**

The same hallucination problem is now inside the SOC. AI-powered tools are fabricating threats that do not exist, and analysts are burning time chasing them. Documented cases from early 2026: **an AI-driven SIEM summarizer generated a lateral movement alert between two hosts where no such activity occurred**. In another case, the **AI invented a playbook** that did not exist, "Playbook 12, S3 exfil containment," and staff wasted time searching for a document that was never written. In a third, an **AI summary caused a team to downgrade and close a real incident** early because the AI's assessment sounded authoritative enough to override their instincts.

Exact Market, "AI Hallucinations Meet Cybersecurity Reality in the SOC," January 2026.

In February 2024, a **Canadian tribunal ruled against Air Canada after its chatbot hallucinated a bereavement fare policy that did not exist.** A customer booked flights based on the chatbot's advice, then got denied the refund. Air Canada tried to argue the chatbot was "a separate legal entity responsible for its own actions." The tribunal called this "a remarkable submission" and ordered the airline to pay. The chatbot was pulled from the website by April 2024.

Moffatt v. Air Canada, 2024 BCCRT 149. Covered by Washington Post, BBC, Feb 2024.

In the legal profession, **486 court cases worldwide now involve AI-generated fabricated citations**, according to a tracker maintained by HEC Paris. In the US alone, 128 licensed lawyers have been caught filing hallucinated case law. In the original Mata v. Avianca case in 2023, a New York attorney asked ChatGPT to verify its own fake citations. ChatGPT confirmed they were real and said they could be found on LexisNexis and Westlaw. They could not. A Stanford analysis found that some AI models hallucinate in one out of every three legal queries.

AI Hallucination Cases database, Damien Charlotin, HEC Paris. Mata v. Avianca, SDNY 2023. Stanford RegLab, May 2024.

Exact Market, January 2026.

"Would you replace a team of 20 workers with an excavator and no crew? You still need trained personnel and safety protocols. The same is true for AI in the SOC."

Here is where it compounds. AI generates a phantom alert. An analyst spends 30 minutes on it. Nothing there. This happens repeatedly. The team starts ignoring AI output. **Alert fatigue was already a problem and now you have AI-induced alert fatigue layered on top**. When a real incident arrives, it gets deprioritized because the analyst has been trained by the tool itself to expect false signals.

The quality of a detection platform is not about how many alerts it generates or how fast it processes logs. It is about the signal-to-noise ratio and the precision of the correlation logic underneath.

AI can help with that, but only when it operates **inside an architecture designed by people who know what real threats look like** and what normal looks like in a given environment. Without that foundation, AI amplifies noise.

17-33% hallucination rate for AI performing domain-specific security tasks

486 court cases worldwide involving AI-generated fabricated content (HEC Paris tracker)

CYBERQUEST SIEM alerts map to expert-authored correlation rules, not probabilistic guesses. When it fires, the detection logic traces back to a documented threat pattern, not an AI hallucination.

NETALERT NDR correlates real network traffic against known threat behaviours instead of flagging anomalies with no context.

That is the difference between signal and noise. nextgensoftware.eu

SECTION 3
The 40% Wall

If the technical risks were the only concern, better tooling would fix them. But the data on AI project outcomes tells a bigger story. **Gartner predicts that over 40% of agentic AI projects will be cancelled before they scale.** Not because the tech fails in a demo. Because organizations are not ready for autonomous systems running inside real production environments.

Gartner, cited by Michael Earls, February 2026, and Accelirate, January 2026.

RAND puts the broader number higher: 80% of AI projects never reach production. Nearly double the failure rate of normal IT projects. S&P Global found that 42% of companies abandoned most of their AI work in 2024, up from 17% the previous year. Deloitte's 2026 report says 66% of organisations report productivity gains from AI, but only 34% are doing anything genuinely transformative with it. The number one barrier is the skills gap, and the number one response from companies has been training, not actual workflow redesign.

RAND Corporation; S&P Global; Deloitte "State of AI in the Enterprise" 2026 (3,235 leaders, 24 countries).

The infrastructure side is equally fragile. In July 2024, a voltage fluctuation in northern Virginia triggered the simultaneous disconnection of 60 data centres, causing a 1,500 megawatt power surplus that forced emergency grid adjustments to prevent cascading outages across the region. The largest US grid operator, PJM Interconnection, serving 65 million people across 13 states, projects it will be six gigawatts short of reliability requirements by 2027. Goldman Sachs forecasts continued electricity price increases driven by AI demand. **The AI infrastructure bet is real, and the grid was never designed for it.**

Harvard Belfer Center, Feb 2026. PJM Interconnection data. Goldman Sachs, Feb 2026. CNBC, Jan 2026.

The February 2026 International AI Safety Report, led by Turing Award winner Yoshua Bengio with over 100 experts from 30 countries, says it plainly. AI models remain unreliable on multi-step tasks, still hallucinate, and no current combination of methods eliminates failures entirely. Agents compound these problems because they operate with more autonomy, which means less opportunity for a human to catch the mistake before it does damage.

International AI Safety Report, 3 February 2026.

Anthropic's own alignment research adds a finding most people would not expect. As models get smarter and take on harder tasks, their failures become more chaotic, not more systematic. The popular image of a precisely misaligned AI pursuing the wrong goal does not match the evidence. The actual pattern is incoherent, unpredictable failure that gets worse with complexity.

Anthropic, "The Hot Mess of AI," 2026.

Michael Earls, February 2026.

“Agentic AI does not create weaknesses. It reveals them. Fragile data quality. Undefined ownership. Weak governance. Autonomy does not introduce chaos. **It amplifies whatever discipline already exists or does not.”**

40%+ of agentic AI projects predicted to be cancelled before scaling (Gartner)

80% of AI projects fail to reach production (RAND Corporation)

42% of companies abandoned most AI initiatives in 2024 (S&P Global)

Only 34% of organizations genuinely reimagining business with AI (Deloitte 2026)

Nextgen Software platforms are **production-ready from day one** because they **do not depend on AI maturity curves** or organizational readiness programmes.

CYBERQUEST SIEM and NETALERT NDR deliver validated detection immediately, with AI augmenting where it has been tested and proven. No pilot phase. No 40% cancellation risk.

nextgensoftware.eu

SECTION 4
The SOC Analyst is not dead. The pyramid is.

Will AI replace the SOC analyst? Wrong question. AI SOC benchmarks from 2025 showed that an optimised AI agent outperformed 95% of human participants on standardised investigation tasks.

Same study showed that human-AI teams investigated incidents **2.3 times faster and with 42% higher accuracy than either alone.** Junior analysts using AI-guided workflows hit near-senior-level performance.

Simbian AI SOC Championship, 2025.

The ISC2 Workforce Study counts 3.5 million unfilled cybersecurity positions globally, up 19% year over year. **Job postings for entry-level security analysts have dropped 53% since 2022, from 68,600 postings to 36,000.**
CrowdStrike cut staff in 2025 while denying it was an AI replacement strategy, but the CyberSN job data tells a clear story: AI is absorbing the work that Tier 1 analysts used to do. The analysts who remain need to operate at a higher level, and there are not enough of them.

International AI Safety Report, 3 February 2026.

The traditional SOC pyramid, where junior analysts drown in false positives while senior investigators sit three escalation layers up, **is dying. What replaces it is flatter. AI handles volume. Humans handle judgment, context, and the decisions that actually matter.**

CyberGen Security, February 2026.

"AI surfaces hypotheses. Humans validate them. The organisations that combine intelligence-led strategy with disciplined AI adoption are the ones maintaining resilience."

For SMEs this is especially sharp. Large enterprises can absorb the cost of failed AI experiments and keep deep security teams regardless. SMEs cannot. If you are running a lean security function and need Cyber Essentials compliance, **the question is not whether to use AI. It is how to use it inside a framework where detection logic is expert-built**, AI output is validated and response workflows are grounded in real threat intelligence rather than AI approximations.

2.3x **faster incident investigation with human-AI teams vs either alone**

3.5M **unfilled cybersecurity positions globally (ISC2 Workforce Study)**

53% **drop in entry-level SOC analyst job postings since 2022**

CYBERQUEST SIEM is built for the flatter SOC. Expert-authored rules handle the signal quality while AI-assisted triage absorbs the volume, so a lean team operates at the level of a much larger one.

NETALERT NDR gives that same team network-level visibility without needing dedicated staff to monitor it. Detection engineering plus AI augmentation, not AI on its own.

nextgensoftware.eu

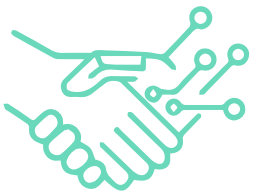
SECTION 5
Expert-led, AI-augmented, continuously validated

Everything in this paper points the same way. **AI is transformative in cybersecurity, but its value is entirely dependent on the quality of the human expertise directing it.** The organisations getting this right share common traits. They pick well-defined problems instead of trying to automate everything. They invest in detection engineering as the foundation, not the afterthought. They treat AI output with the same scepticism they would apply to any unvalidated source. And they build governance and oversight in from day one, not after something breaks.

The competitive advantage in cybersecurity for the foreseeable future will not go to whoever has the most advanced AI.
It will go to whoever integrates AI capability with human expertise, purpose-built detection logic and operationally validated workflows. That is not a philosophical position. It is the pattern in the data.

Takeaways

1. **AI is creating new attack vectors (slopsquatting) and new noise (SOC hallucinations) at the same time it accelerates defence. You have to deal with both.**
2. **Over 40% of agentic AI projects will fail before scaling. The blockers are skills gaps, data quality and governance. Not the technology.**
3. **Human-AI teams outperform either alone. 2.3x faster, 42% more accurate. The goal is augmentation.**
4. **Detection engineering fundamentals, expert-built rules, validated correlation, real threat intel, are what make AI in security operations reliable instead of reckless.**
5. **Treat every AI output as a hypothesis to be validated. Not a conclusion to be trusted.**



Nextgen Software builds security platforms for organizations that need reliable detection without enterprise budgets.

CYBERQUEST SIEM and NETALERT NDR are built on the principle this paper argues for: expert-authored detection, augmented by AI, continuously validated.

nextgensoftware.eu

nextgen

CYBERQUEST

NETALERT

Questions? Contact us.

www.nextgensoftware.eu
office@nextgensoftware.eu
marketing@nextgensoftware.eu