Why cyber defenders need to be ready for frontier AI

Understanding the threats and staying ahead of the adversary

Artificial intelligence (AI) is no longer a distant or speculative issue for cyber security. The most advanced systems, often referred to as frontier AI models, are already showing results in specific steps of cyber operations, such as in helping identify zero days in widely used software, or solving cryptographic challenges. These models are changing the cost, speed and scale of operations for both attackers and defenders. It means tasks which once required specialist skills – such as writing exploit code, understanding system architecture or using attack tools – can increasingly be automated using AI in certain circumstances.

Recent findings from the AI Security Institute (AISI) highlight an accelerated increase in the cyber capabilities of frontier models, and at a far faster pace than many expected. In specific cyber tasks, models already exceed what a skilled practitioner could achieve at lower cost. In parallel, publicly available examples demonstrate how these capabilities are already being used in practice, and how attackers could adopt them more widely.

The implication is clear: defenders should assume that at least some attackers already have access to capable AI tools. Since frontier AI capabilities potentially strengthen cyber attackers, cyber defenders must use the same capabilities to drive defensive advantage.

This blog helps cyber security professionals, decision makers and risk owners to better understand:

- how frontier AI capabilities are evolving
- how attackers are already using – or could use – these capabilities
- where frontier AI is most likely to deliver game-changing improvements in cyber defence

It also explains the critical structural advantage defenders hold and how to retain that advantage to stay ahead of the attackers.

Glossary of terms used

+ Show All

AI + Show

(unless otherwise stated) refers to generative AI (GenAI) – the class of AI systems that can produce text, code and other outputs in response to prompts. (This is distinct from earlier forms of AI which have been used in the cyber security context for many years).

Frontier AI models + Show

refer to the most capable models available at any given time. It's worth noting that capabilities developed in frontier models can be transferred into smaller, cheaper, or open-weight models through a process called distillation – meaning advances at the frontier set the direction of travel for the whole ecosystem.

AI systems + Show

refer to broader AI systems that combine models with tools, workflows and human oversight. Examples include:

- agentic systems that take sequences of actions autonomously
- systems where humans and AI collaborate on tasks

> Many of the benefits and risks discussed in this blog are delivered by AI systems rather than raw model capability alone.

How frontier AI capabilities in cyber operations are evolving

Whilst safeguards applied by responsible model developers can limit misuse of AI, these protections can often be bypassed (and in open-weight models can be removed or are absent from the start). Publicly-available AI model weights can be modified, and safeguards removed entirely. Reporting from multiple frontier labs over the last few years has shown attackers are using frontier AI models to aid their operations. For example:

- Anthropic's Disrupting the first reported AI-orchestrated cyber espionage campaign
- Google's Adversarial misuse of generative AI
- Open AI's Disrupting malicious uses of AI

In its recent research into measuring AI agents' progress on multi-step cyber attack scenarios, AISI evaluated the cyber capabilities of 7 frontier AI models, released before March 2026. Importantly, 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.

The models were given specific tasks in 2 simulated environments (an enterprise network and an industrial control system) and left to operate autonomously.

On the 32-step enterprise network attack, estimated to take a human cyber security expert approximately 14 hours to complete end-to-end, the best-performing model (Claude Opus 4.6, released February 2026):

- Averaged 15.6 steps, with extended processing time - which corresponds to roughly 6 of the 14 hours a human expert would need.
- Averaged 9.8 steps without extended processing time – up from fewer than 2 steps 18 months earlier.
- Completed its single best run in 22 of 32 steps.

As of March 2026, no public model has completed the full scenario end-to-end.

On the more complex industrial control system attack scenario, AI performance was significantly more limited. But even here there were early signs of progress: the most recent models were the first to make any consistent headway, and in some cases found attack approaches the scenario designers hadn't anticipated.

> In just 18 months, the best AI models went from barely making any progress on a realistic simulated enterprise attack to completing over half of it, and the cost of a full attempt is now around £65.

Factors driving the rapid pace of improvement in offensive AI cyber capabilities

There are 2 reinforcing trends driving this acceleration:

1. **The capability ceiling is rising fast.** Each new generation of AI model is better at working through complex attack sequences than the last. The best model in early 2026 completed nearly 6 times more attack steps than the best model 18 months earlier.
2. **Running these attacks is getting cheaper, not harder.** Giving the same model more processing time reliably improves results with no additional attacker skill required. At current pricing, a full attempt at this simulated attack costs around £65. This means the limiting factor is increasingly funding, not expertise.

Current limitations of the AI models in attack scenarios

Despite rapid improvement, AI models released before March 2026 still fall short of end-to-end completion of these complex attack scenarios. The main reasons for this limitation are:

- The amount of processing time. In several evaluations, models were still taking useful actions when they reached the end of the allotted processing time, meaning the results likely understate their full capability.
- Specialist knowledge gaps in areas such as reverse engineering, cryptography, and malware development. Performance drops sharply when attacks transition from reconnaissance and web exploitation to phases where there is less training data. Complex, multi-step coordination is unreliable. Models struggle with operations that require managing several concurrent processes in real-time.
- Models lose track over long operations which leads to lost context and missed opportunities.
- Results are inconsistent. The same model with the same amount of processing time can produce very different results across individual runs.

It's important to note that these aren't permanent barriers. They are areas where the rate of improvement has already been rapid, and where even modest extensions to processing time or human-AI teaming can result in substantial gains. It's also worth noting that these results likely underrepresent what current models are capable of, given the evaluations used a standard setup with no specialist tools or human involvement. Purpose-built approaches would almost certainly perform better.

> An important near-term advantage for defenders is that the activity of frontier AI models released before March 2026 tends to generate noticeable security alerts and is relatively easy to detect. Current models would likely be identified and disrupted before they managed to achieve the levels of progress outlined above – but only in environments with effective monitoring and the ability to respond.

For a broader assessment of how AI-enabled cyber threats may evolve in the near term, see the NCSC's assessment on AI and cyber threat.

How frontier AI capabilities may be used to strengthen cyber defence

With attackers already making use of AI to support their operations, defenders should also look to where AI can assist their work and increase their ability to protect systems at scale and pace.

AI is already being used across many cyber security workflows, from analysing threat intelligence to alert triage, ensuring policy compliance, and finding vulnerabilities in source code. In some areas, these tools are delivering tangible benefits. More importantly, there are some areas of application where AI-enhancement is likely to be a game changer in how defenders operate in practice. Here we outline 3 of our top bets:

+ Show All

1. Reducing the attack surface – or system hardening + Show

AI-enabled security testing tools are making vulnerability discovery and penetration testing faster and more comprehensive. In limited settings these tools can now:

- scan systems continuously at machine speed
- identify vulnerabilities and misconfigurations
- test exploitability
- map complex attack paths that would traditionally take human testers many hours to uncover

The majority of these tools still only identify problems rather than fix them, but systems are beginning to demonstrate autonomous remediation, including generating and applying code patches. Examples emerging from initiatives such as DARPA's AIxCC challenge, Google's CodeMender, and OpenAI's Codex Security point to a future where vulnerability

discovery and remediation happen far more quickly, reducing attackers' windows of opportunity. Use in critical systems remains challenging, but the direction of travel is clear.

2. Improving threat detection and investigation

+ Show

AI, and in particular LLM-enabled tools and systems, have potential to help defenders:

- triage alerts
- make sense of patterns from diverse logs
- write summary reports to support analyst decision-making

More significantly, in the future AI may enable a shift away from detection models that assume a human must review each alert. Future systems may retain far more contextual information over time, increase logging, initiate targeted investigations, or deploy deception techniques, such as honeypots, to gather intelligence on suspicious activity. This could improve detection of slow, subtle intrusions that often evade traditional approaches.

Although these systems are still emerging and bring new risks, such as over-reliance on automated judgement and reduced transparency, they nevertheless represent a meaningful change in how defenders may detect and understand adversary behaviour.

3. Automating mitigation and response

+ Show

Some organisations are exploring automated response capabilities that can contain or remediate threats without any human intervention, such as:

- blocking traffic flows
- quarantining suspicious processes
- revoking user accesses

When carefully designed, this can significantly reduce the time between compromise and containment. But it's important to note that automating responses does also introduce substantial risk. 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.

Shaping the battlefield: how to retain the defender advantage in AI

Despite the rapid improvements seen in successful use of AI capabilities for offensive purposes, defenders do still have some advantages over attackers.

Defenders can (and must) work collaboratively across the globe, sharing their insights transparently and openly. Furthermore, market forces should drive the community towards building strong defences from AI attacks. But the key advantage in defenders' favour is that they have the ability to 'shape the battlefield'; that is to shape their environment to make it work better for them and disadvantage the adversary.

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. Used well, this can deliver disproportionate benefits to defenders (requiring attackers to be super stealthy and to successfully hide every time).

> But this advantage is not guaranteed. Where attackers adopt AI more effectively than defenders, or where baseline cyber security is weak, the quality of defensive data degrades and the information gap narrows quickly.

AI-enhanced cyber security tools also introduce new complexity, creating additional dependencies and failure modes that may be difficult to detect or control. In line with the Government's Code of Practice for the security of AI, these tools should be designed and deployed securely in their own right, and treated as part of the attack surface.

The most effective actions are not novel or experimental. Since AI will more quickly enable rapid scaling of attacks against 'soft' rather than 'hard' targets, strong baseline security is vital. To limit what AI-enabled attackers can achieve, ensure you have the cyber security basics in place, such as:

- accurate asset inventories
- robust access controls
- secure configuration
- comprehensive logging

AI won't compensate for weak security foundations, but it will amplify both strengths and weaknesses. Organisations that invest now in strong security baselines and carefully deployed AI-enhanced defence will be best placed to retain defender advantage as AI increasingly shapes the cyber risk environment. We'll continue to share our informed thinking on the use of AI in cyber security, and will be updating the Intelligent security tools guidance in due course.

Paul J, Technical Director for Cyber AI Research, NCSC

Alan Steer, Cyber Security Researcher, Cyber and Autonomous Systems Team, AISI, DSIT

WRITTEN BY

Paul J
Technical Director for Data Science Research, NCSC

Alan Steer
Cyber Security Researcher, Cyber and Autonomous Systems Team, AISI, DSIT

PUBLISHED

30 March 2026

WRITTEN FOR

Cyber security professionals Large organisations Public sector Small & medium sized organisations

PART OF BLOG

NCSC publications