Familiar weaknesses, exploited faster: defending against agentic cyber attacks

How to protect your organisation from cyber attacks created using ‘agentic offensive frameworks’.

Since September 2026, the NCSC has seen a marked rise in AI-assisted cyber attacks. Organisations across multiple sectors are reporting incidents as a result of threat actors using ‘agentic offensive frameworks'; software originally designed for red teaming being applied beyond its defensive purpose.

The incidents involved automated identification and exploitation attempts across organisations' internet-facing services and assets in quick succession. This resulted in downtime to some services or the compromise of systems, providing attackers with a foothold for further agent-driven activity (or in some cases, human-driven activity within their network).

It’s worth noting that:

- none of the weaknesses being exploited are new or novel
- the tools the agents have access to are common and provide no new capability the typical threat actor would have access to

What **has** changed, is how quickly and persistently threat actors can conduct attacks – at low cost – using these frameworks combined with cheap, capable open models or illegal API resellers.

This blog explains how agentic offensive frameworks are being used, what this means for cyber defenders, and the steps you should take to help protect your organisation.

What is an agentic offensive framework?

An agentic offensive framework allows AI models to interact with a set of off-the-shelf cyber security tools for the purpose of probing digital systems for weaknesses. Typically, a framework will:

- use an LLM to plan a sequence of attack steps
- use the tools against a target
- retrieve the results and decide what to do next

At the time of writing (November 2026), this type of activity has become synonymous with the ARTEX framework, a multi-agent penetration testing and red-teaming tool originally developed for authorised security testing. Different tools come and go (others include Cairn, PentAGI, RedAmon, Xalgorix and Darkmoon), so rather than concentrating on a particular framework, here we’ll focus on how the combination of *a human attacker* plus *a capable model* plus *a framework* creates an uplift in capability.

Compared with a human attacker working without tools, an agent plus a framework:

- **works quickly** (it moves from discovery to exploitation in minutes)
- **is exhaustive in its search** (it will try every endpoint, parameter and credential it finds)
- **instantly learns from failures** (it creates memory, reads error messages, responds and adapts its next attempt)
- **is cheap to run, at scale** (the cost of running automated attacks is low)

This does mean that today's frameworks are noisy. They trade stealth for speed, generate a lot of traffic and will interact with almost anything they find, which makes them easier to detect. This may not always be the case if future frameworks prioritise stealth.

Notably, the issues being exploited currently are not novel memory corruption vulnerabilities (or even known but unpatched vulnerabilities), but unmitigated weaknesses that are often the result of years of technical debt that organisations have accumulated. Agentic attacks are exploiting low hanging fruit such as:

- credentials mistakenly committed to code repositories
- misconfigurations in infrastructure as code (IAC)

- default credentials or lack of authentication
- abuse of weaknesses in APIs

Once an initial foothold has been gained, the agents explore the environment, exploiting the fact that many credentials in an organisation are over-privileged for their needs, and networks are often flat or interconnected.

Defending against agentic offensive frameworks

Organisations need strong operational discipline to avoid being overwhelmed by the volume, intensity and duration of attacks from agentic offensive frameworks. Answering the following questions will help you develop a holistic approach that can protect your organisation.

1 **Do you know what's exposed?**

Agents start by mapping every device and service that's reachable. You should know which systems and services are exposed to the internet, and who is responsible for them. The NCSC's Asset Management Guidance can help with this.

Services successfully compromised by agentic offensive frameworks are sometimes those that an organisation has forgotten about such as old websites and legacy equipment. Organisations should be conducting regular security reviews of the assets they know about and using either internal skills or a consultancy to perform reconnaissance against their perimeter.

2 **Can you take services off the public internet?**

Rather than try to harden all services individually, you should look at making services only available to authenticated users. There are a few ways this can be done:

- place services behind an identity aware proxy, ideally ensuring strong phishing resistant authentication, following modern zero trust methods
- use firewalls to restrict access to services unless traffic comes from the corporate VPN or internal ranges
- configure mutual TLS certificates to require authentication by a user prior to interacting with a service

In most organisations, the majority of services currently accessible from the perimeter can be locked down quickly so they can only be accessed by staff or partners. Specific hardening can then focus on the remainder of services that can't be locked down. If you were to assume everything you have connected to the internet is vulnerable, how would this change your approach to security?

3 **Can you detect the use of agentic offensive frameworks?**

Agent activity tends to look different from human activity: bursts of requests followed by short thinking phases, a wide variety of endpoints and parameters (including garbage or example variants which would never work), and increased error rates. Are you able to correlate across log sources to detect these signals? Collect and retain logs needed to detect, investigate, and recover from an incident; the NCSC's logging guidance can help with this.

Like all threat actors, agents leave a trail and appropriate logging helps responders to understand what assets the agent interacted with (and what it did) in order to remediate. Unlike human threat actors, an AI agent reads every error it receives and uses that information to refine its next attempt. Error messages shown to external users should reveal as little as possible to reduce the efficacy of the next action, so make sure your error messages give nothing away.

4 **Are known vulnerabilities patched and misconfigurations fixed?**

Agents can test known vulnerabilities and exploits across targets very quickly, so patching internet-facing systems needs to keep pace. The same applies to configuration, particularly identity and access (IAM) settings that quietly drift from original designs. To help identify, prioritise and fix vulnerabilities within assets that are accessible from the internet, you can use the NCSC's external attack surface management guidance in conjunction with our vulnerability management guidance.

Companies assured under the NCSC's penetration testing scheme (CHECK) can conduct penetration testing to help you discover vulnerabilities and misconfigurations. Smaller organisations should refer to our guidance on adversary simulation to see if it's right for your organisation.

5 **Have you installed bot challenges to protect abuse targets?**

Account and identity flows, inventory, checkout, and resource-intensive endpoints (such as search, report generation or LLM-backed features) are high priority assets that should be protected by deploying 'bot challenges'. These require human interaction in an attempt to prevent automated tools from abusing services. Challenges can be set to low interaction (essentially invisible to human users) unless risk signals become elevated. When deploying bot challenges, provide an accessible alternative when a challenge is triggered.

6 **Are your APIs protected adequately?**

Bot challenges are presented in the browser, on pages such as login or submission forms, but if the underlying API doesn't enforce the challenge, an attack can call the API directly and bypass it. Many APIs have limits to prevent large volumes of requests, perhaps from specific IP addresses. However, agents routinely rotate through proxies and VPNs, so blocking IP addresses is often ineffective. Instead you should:

- apply rate limits to accounts, API tokens and authenticated sessions
- look at behaviour as well as volume: lots of varied requests, rapid failures or rapid retries can be as telling as raw numbers
- read the NCSC's API security guidance

If your organisation publishes public code repositories, enable pre-commit automation to prevent secrets being pushed. Our guidance on keeping secrets out of the codebase and using a cloud platform securely can help.

7 **Who can create an account, and how quickly?**

If anyone is able to create an account in seconds, then account-based controls (including the rate limits above) are easy to get around. Agents can sign up for new accounts as fast as you can block them. Consider throttling account creation, and heavier throttling of sign-ups that use disposable email services. Agents can use default credentials and use tools to brute-force weak passwords. You should protect accounts with strong authentication and appropriate access controls in line with the NCSC's authentication guidance.

8 **Could you exploit an agent's perseverance?**

The traits that make agents effective also make them susceptible to deception. A honeytoken or tripwire placed somewhere no legitimate user should touch gives a high-confidence alert the moment it's used. An agent's perseverance and exhaustive searching mean decoy systems, tarpits and fake 'successes' can use up an agent's time or attempt budget, and keep it away from your real assets. Microsoft has reported that deception measures reduced the attack traffic reaching real systems during a recent campaign using deception strategies.

Cyber security fundamentals matter more than ever

The NCSC believes the adoption of AI-enabled cyber can close the gap between attackers and defenders. We recently blogged about unlocking the potential of agentic cyber defence across a range of areas including enhanced threat detection and automated incident response. However, it will take time for defensive AI-driven programmes to mature, so organisations can't sit back and wait for agentic defence to protect them. AI-enabled cyber defence is not the sole answer to agentic attacks; the foundations of cyber security and basic cyber hygiene remain the priority.

As we stated in the NCSC's 2026 'The State of the Cyber Nation' Annual Review, AI does **not** make established cyber security controls obsolete. In fact building strong foundations has never mattered more. AI is making weaknesses easier to find and faster to exploit; a forgotten key or an unprotected API which might once have gone unnoticed for years can be found in minutes.

Nearly everything the NCSC has seen can be prevented by implementing cyber security best practice, the application of modern cyber security approaches like zero trust, and by the diligent application of technical controls in line with your organisation's technologies and processes. For organisations seeking a structured approach:

- zero trust principles can help protect against agentic attacks; explicitly verify identity, ensure least privilege everywhere, and deploy controls as if attackers are already inside your network
- the Cyber Assessment Framework helps organisations responsible for essential services or critical functions assess and improve their cyber resilience
- the NCSC's Cyber Essentials sets out the minimum technical controls for organisations, as recommended by the government

Although they may not always be applied in the right place, the controls that mitigate the threat from agentic offensive frameworks will already be familiar to many organisations. Today's frameworks may be replaced by quieter, more capable

ones, and tools will change. But the cyber security fundamentals won't.

This is your chance to get ahead of the threat. Consider how your organisation would answer the questions above, and make sure your answers hold up against an attacker that never needs a break and never gets tired of asking.

Kris B, Incident Handler, Incident Management

Ellen T, Lead Data Scientist, Vulnerability Management Team

David Chismon, CTO for Architecture

WRITTEN BY

Kris B
Incident Handler, Incident Management

Ellen T
Lead Data Scientist, Vulnerability Management Team

David Chismon
CTO for Architecture

PUBLISHED

24 November 2026

WRITTEN FOR

Cyber security professionals

PART OF BLOG

NCSC publications