One does not simply defend agentically

Defenders can't use AI in the same way attackers can, but there's much they can do to unlock the potential of agentic cyber defence.

One of my favourite cyber security maxims is Halvar Flake's observation that "All offensive problems are technical problems, and all defensive problems are political problems".

For an attacker conducting offensive cyber actions, the problems that need to be solved are largely technical. They might be:

'*Do I have an exploit for this vulnerability?*
or
'*How do I avoid detection?*'

Whilst defenders also have technical challenges, they mostly wrestle with what Halvar describes as 'political' problems (which for the purposes of this discussion can be framed as organisational problems). Problems such as:

'Can we get budget to replace this end-of-life system or invest in a more secure option?'

'How do I get IT operations to make time for patching?'

'How do we get a change request to put in firewall rules approved?'

In other words, defenders are most restricted by their organisational policies, whilst most attackers are restricted by technical hurdles. The reason for this difference becomes clear when you compare the core purposes of attackers and defenders.

For attackers, their core purpose is 'to conduct cyber offence' and success will cause their employer to profit. Whether that's stealing money from organisations, extorting victims, or exfiltrating information. Cyber offence is their *raison d'être*.

For the defenders, their mission is to help their organisation avoid loss. Their employer's mission isn't cyber defence; it will depend upon whatever activity that organisation does (such as developing products, treating patients, or providing digital services). Cyber defence is effectively 'a cost of doing business', one of many priorities that the organisation has to manage.

This means the cost of cyber defence has to be rigorously assessed to make sure it does not harm the organisation's competing priorities:

'Could that money earmarked to replace the end-of-life system be better used on marketing?'

'Will that patch take down the VPN?'

'How sure are we that the firewall rule will not break a business function that relies on that connectivity?'

Some board members may see little difference between a DoS attack taking down the organisation's IT, or a poorly implemented action by the cyber defence team that does the same thing. Except that the board can't shout at an attacker over the phone ...

Agentic tooling can be used for *cyber offence*, because AI is good at helping with technical problems with a clearly measurable success state. **Offensive problems** are mostly technical, and usually have a clear success state (the target program crashes, your malware calls home). But **defensive problems**, as we've established above, are *not* mostly technical. Nor do they always have a clear success state.

Using AI automation for defence therefore quickly becomes a matter of organisational politics, and someone needs to be responsible for the action taken. Defenders simply cannot put AI to work in the same way attackers can. This is an inconvenient truth, as it suggests that the threat from AI-enabled cyber attacks will grow, whilst autonomous / agentic cyber defence might struggle to keep up unless we approach things differently.

How can defenders make better use of AI?

Rather than trying to mimic attackers' use of agentic tooling (and risk breaking things), defenders need to solve the problem by *explicitly* considering the constraints. I am biased, but I think the work on Autonomous Cyber Defence by CETAS

– commissioned by the NCSC – is excellent. It identified early that defenders have different levels of appetite for where they would apply AI. But how can we break the problem down further to make it tractable?

To start with, let's examine the technical defensive tasks and consider how we might use AI automation. Firstly, there is the category of offensive techniques being applied defensively, such as penetration testing and vulnerability research or discovery. These are typically done in a way that minimises risks to the business; applications are tested before they go live, and vulnerabilities are handed to development teams to fix.

Vulnerability scanning can take a lot of buy-in across an organisation. The first scan requires convincing management that it will not break anything. However, it quickly becomes a business-as-usual service, identifying unpatched vulnerabilities.

Secondly, there are examples of where technical and dynamic approaches are used for cyber defence. Establishing a SOC is also typically a lengthy (and often expensive) process. Legal aspects and policies need to be agreed, and data then exported from live systems to a new platform where malicious activity can be detected and triaged before the team takes action.

For both the above, there is a lot of work done upfront to show that detection of vulnerabilities or missing patches will not harm the business. The results of the detection are then given to **humans** to action, whether that's development teams to implement fixes, system owners to apply patches, or incident response teams to respond to incidents [1].

There are 3 core principles at work here:

1. Technology is behind the detection, but humans respond with actions.
2. The technological detection must not harm the organisation.
3. The technological response is tightly controlled and scoped (for example to a single user account or computer in response to a clear malicious signal).

By developing these principles and splitting the problem into different dimensions, we can propose a framework for estimating the 'riskiness' of a defensive action:

Dimension	Level	Description
Potency Whether the automated process can only observe existing state or can alter systems, permissions, configuration or runtime behaviour.	**0**	AI provided data and then outputs explainable advice to a human
	1	AI provides non-explainable advice to a human
	2	Autonomously collects data using read-only privileges across APIs and web search introducing a risk of exfiltration, data being added to files
	3	Changes data or system state via APIs (Easier to guardrail than code)
	4	Executes code or makes direct system/runtime changes
Scope How bounded the affected technical estate is.	**0**	Single system
	1	Known set of systems
	2	Unknown scope
	3	Multi-system or cross-domain
	4	Enterprise-wide, fleet or shared platform
Criticality How important the affected scope is to business or mission delivery. Includes the criticality of systems within the scope.	**0**	Known not critical; not business relevant
	1	Business relevant but non-critical
	2	Known support of critical service
	3	Unknown whether it supports critical service
	4	Known critical system or service
Rollout confidence How well the proposed scope and impact can be proved before execution.	**0**	Audit mode, proper digital twin environment, or strong evidence from historical/recent logs
	1	Imperfect development environment or system twin, or partial logs
	2	Can ring-deploy, with strong test suite
	3	Single system with no failover, but robust tests
	4	No failover and no or limited tests; weak or anecdotal validation
Recoverability How cleanly and quickly harm can be reversed if the automation is wrong.	**0**	None needed because there is no state change
	1	Simple undo, for example API revert or permission rollback
	2	Code or configuration rollback with no wider systemic effects, using CI/CD

	3	Rollback possible but may leave inconsistent state, currency or dependency issues; possibly CI/CD of a complex system
	4	Manual reversal, irreversible, or recovery depends on many teams or systems

By considering the above (or something like it) we can identify the lowest-risk actions that we can start to automate. An easy win is focusing on tasks that are low potency ('advise a human' rather than 'affect a system directly'). Florian Roth has a good blog on how defenders can use generative AI to make sense of the deluge of data they face and prioritise their actions. AI is good at summarising and so is already useful to defenders in finding details, or making sense of the huge volumes of reports or threat intelligence.

Areas that can already be largely automated are good candidates for implementing agentic processing around or extending, for example, detection efforts.

Next, we look at how we can reduce the risk of tasks we want to automate but currently are uncomfortable with doing so. Can we focus on systems we know are easy to recover, following the 'cattle, not pets' model? Government teams have also been working with departments to use agentic tooling to rewrite or recreate legacy systems in modern, and therefore more re-deployable, stacks. See the Defra AI Legacy Modernisation Playbook and reporting on the government's AI code-remediation pilots.

Unlocking the potential of agentic AI actions

The NCSC and DCMS are jointly working on Cyber Shield, a national-scale agentic cyber defence ecosystem, and delivering that vision will require making autonomous defensive actions possible in a way they currently are not. As part of this, the NCSC sees a number of areas where we think specific research is needed to unlock the potential of agentic cyber defence, and will be publishing our 'AI for Cyber Defence' problem book soon.

An area where I believe we particularly need research, products, and evidence of efficiency is in how we deterministically prove that 'low risk' actions *really are* low risk. Can AI help us analyse traffic logs and show conclusively that we do know all the routes clients connect by? Can AI reverse engineer or otherwise assess a system and its binaries to show exactly which network calls it could ever make, or which processes it might need to spawn?

Answering these questions will give us, and the organisations we protect, the confidence to take automated defensive actions. Furthermore, it opens up the chance to automate the hardening of systems and reduce attack surfaces and exposure, which will be crucial in combating AI-enabled cyber attacks.

All these efforts will take time, effort, and research. Organisations cannot risk just waiting for agentic defence to roll in and protect them; they also need to be focussing on improving their security the traditional way.

Dave Chismon

CTO for Architecture

[1]There are a very small number of cases where remedial action is automated, such as security orchestration, automation and response (SOAR). In SOCs that use SOAR, automations might be enabled that, for example, make a user re-authenticate if suspicious activity is seen. Each rule is crafted, tested and, crucially, highly deterministic.

WRITTEN BY

Dave Chismon

CTO for Architecture

PUBLISHED

21 September 2026

WRITTEN FOR

Cyber security professionals

Large organisations

PART OF BLOG

NCSC publications