AI for security teams: where can we start safely?

Start with one high-volume task whose output an analyst can check in minutes, such as phishing triage, alert summaries or query writing, and run it read-only in a tool approved for that data. Add the next use case only after you have measured accuracy and time saved against your own baseline.

Reviewed	Next review	Version	Written for	Latest version
5 Nov 2026	**5 Dec 2026**	**1.1**	**SOC, security engineering and GRC teams starting to use AI**	**sheets.einblick.dev/ai-for-security/**

If you only do five things

1. **Start with tasks you can check in minutes.**
   The Australian, Canadian, New Zealand and UK cyber agencies advise verifying AI output against evidence and context before acting on it[1].
2. **Use only approved tools for security data.**
   Keep logs, incident details and customer data out of personal and consumer chatbots[1]. Data pasted there sits outside your contracts, retention rules and access controls.
3. **Keep AI read-only until it has a track record.**
   Let it summarize and suggest before it changes anything. Later, allow unattended actions only where they are narrow, pre-approved and reversible[1].
4. **Measure accuracy and time on your own cases.**
   Vendor claims and demos aren't enough to show how a tool performs in your environment[1]. Record today's time and error rate for the task, then compare.
5. **Log prompts and responses.**
   When AI output feeds an incident decision, you need to show what the model was given and what it returned. The agencies call for AI-assisted actions to be traceable and reviewable[1].

Pick your first use case

1. **Does it come up many times a week?**
   Phishing reports, alert summaries and query writing qualify. A task done once a quarter rarely repays the setup.
2. **Can an analyst check the output quickly?**
   If checking takes as long as doing the task, nothing is saved. Summaries and queries are quick to check; root-cause conclusions are not.
3. **What does a wrong answer cost?**
   Start where a mistake costs an analyst a few minutes. Leave tasks where an error means a missed intrusion or a blocked production system for later.

Good first use cases

- **Summarizing alerts and incidents.** Turn alerts, logs and forensic artifacts into an incident summary or an update for executives, with a responder checking it before it goes out[1]. Check every hostname, account and timestamp against the source.
- **Triaging user-reported phishing.** Reports arrive in volume and each ends in a verdict an analyst can check against the evidence. In a timed Microsoft trial with 167 analysts, those working with its triage agent found up to 6.5 times as many malicious emails per minute and were up to 77% more accurate[2].
- **Writing and explaining queries.** Draft KQL and SPL queries or Sigma rules from plain language and have unfamiliar command lines explained. In a 10-month study of 45 SOC analysts, explaining and writing commands was the most common task, at 31% of prompts, and about 4% asked for a verdict[3]. Test each query against known activity before you trust an empty result.
- **Reviewing scripts and infrastructure code.** Have it explain a suspicious PowerShell script or flag risky Terraform settings before merge. The agencies list scanning infrastructure as code for flaws that pattern-based scanners miss[1].
- **Drafting policies and reports.** Produce first drafts of policy text, control mappings and post-incident reports. The agencies also list an internal assistant that gives guidance based on your own policies[1]. A named owner checks every claim.

What changed in 2026

- **Five Eyes agency heads told defenders to use AI.** On 22 June 2026 they wrote that attackers already use AI to move faster and that defenders must do the same, alongside faster patching and a smaller attack surface[4].
- **ENISA recommended AI-assisted triage.** Its July 2026 paper on frontier AI calls for it in incident response, with humans reviewing the results[5].
- **Four agencies published guidance on AI in cyber defense.** In August 2026 the Australian, Canadian, New Zealand and UK agencies set out uses for each security function and questions for vendors, along with principles for oversight and testing[1].
- **Microsoft 365 E5 and E7 started including Security Copilot.** The rollout began on 18 November 2025 and continued into 2026[6].

What you can skip

- **Autonomous response at the start.** Isolating hosts, disabling accounts and purging mail can wait. The NCSC warns that AI tools can be better at finding problems than at reliably acting on them[7].

Mistakes that cost the most

- **Treating a reviewer's sign-off as a full check.** In Microsoft's phishing trial, analysts shown the agent's verdicts were 29 percentage points more likely to miss malicious emails it had been made to label benign[2].

- **Training your own model.** Start with the AI built into tools you already run. The agencies advise using frontier AI to augment existing tools and processes rather than as a standalone system[1].
- **Buying a platform before a pilot.** Run one use case on your own data first. Ask vendors for results measured in real operations, such as time to detect, time to respond and false positives[1].
- **Replacing working playbooks.** Keep fixed steps such as enrichment and ticket creation in rule-based automation, which gives the same result every time. Use AI where the input is unstructured text.

How to tell it's helping

- **Correction rate.** The share of outputs an analyst had to fix or discard, tracked per use case.
- **False negatives among AI-closed alerts.** Re-check a sample of what the AI marked benign without showing the reviewer its verdict.
- **Time per case against your baseline.** Compare the same task type before and after. Microsoft's study of live operations linked Security Copilot adoption to a 30% shorter mean time to resolve incidents but could not rule out other causes[9].
- **Cost per case.** License or compute cost divided by cases handled. For its phishing triage agent, Microsoft's usage dashboard shows cost per email processed[8].

- **Letting attacker-written content steer an AI that can act.** Phishing emails, log lines and malware strings can carry prompt injections aimed at the AI that reads them. The agencies call for safeguards when logs and artifacts go into AI systems[1].
- **Running agents on a person's admin account.** Give each agent its own identity with only the permissions it needs, as Microsoft recommends for its phishing triage agent[8].
- **Counting time saved and nothing else.** Track corrections and misses as well. The agencies warn that AI that only adds information can increase workload and risk[1].

If you run Microsoft security tools

- **Check your allowance.** E5 and E7 include 400 Security Compute Units (SCUs) a month per 1,000 paid user licenses, up to 10,000 SCUs a month[6].
- **Review the owner settings.** After auto-provisioning, access to Microsoft 365 data is on and data sharing with Microsoft is off. Roles such as Global Administrator and Security Administrator get owner access[10].
- **Turn on prompt logging.** Prompt and response content reaches Purview only through DSPM for AI, after a Security Administrator opts in[11]. Sessions inactive for 180 days are deleted[12], so copy what you need into the incident record.
- **Use Copilot in advanced hunting for KQL.** It runs read-only queries with the user's permissions. Review the query and its results before you act[13].

What this looks like at your size

Small team, no security staff

- Use the AI features already in your email, endpoint or identity products before buying anything.
- Write down which AI tools are approved and what data may go into them.
- Start with explanations of suspicious emails, scripts and sign-in alerts, with a person making the call.
- Save the prompts and responses behind any incident decision in the incident ticket.

Growing, with IT or security staff

- Pick one use case, record a baseline, then pilot it and compare.
- Turn on audit logging for prompts and responses and send the logs to your SIEM.
- Give agents their own identities and the fewest permissions that work.
- Review a sample of AI-closed alerts every week.

Enterprise or regulated

- Keep an inventory of AI use cases with owners, data, permissions and measured results.
- Run AI tools through change management and test them in your own environment before production[1].
- Ask vendors the agencies' questions on logging and data use[1].
- Add scenarios for AI failure, compromise and hallucination to your incident response plans[1].

Version history

5 Nov 2026, version 1.1	Re-checked every claim citing the joint guidance on AI in cyber defence against the NCSC-NZ co-author copy, and corrected its date to 12 August 2026 and its link.

Sources

1. Opportunities for AI in cyber defence. ASD's ACSC, Canadian Centre for Cyber Security, NCSC-NZ and NCSC-UK, 12 Aug 2026. ncsc.govt.nz
2. Randomized Controlled Trials for Phishing Triage Agent. Microsoft (arXiv preprint), 17 Nov 2025. arxiv.org
3. LLMs in the SOC: An Empirical Study of Human-AI Collaboration in Security Operations Centres. CSIRO's Data61 and eSentire (arXiv preprint), 19 Sep 2025. arxiv.org
4. The AI shift in cyber risk: why leaders must act now (Five Eyes cyber security agencies statement). ASD's ACSC, Canadian Centre for Cyber Security, NCSC-NZ, NCSC-UK, NSA and CISA, 22 Jun 2026. ncsc.govt.nz
5. ENISA's view on Cybersecurity in the Frontier AI Era. ENISA, 7 Jul 2026. enisa.europa.eu
6. Learn about Security Copilot for Microsoft 365 E5 and E7 customers. Microsoft, 19 Jun 2026. learn.microsoft.com
7. Supporting AI adoption for UK cyber defence. NCSC (UK), 23 Apr 2026. ncsc.gov.uk
8. Microsoft Security Copilot Phishing Triage Agent in Microsoft Defender. Microsoft, 2 Jul 2026. learn.microsoft.com
9. Generative AI and Security Operations Center Productivity: Evidence from Live Operations. Microsoft (arXiv preprint), 13 Nov 2024. arxiv.org
10. Understand how Security Copilot is auto provisioned for eligible Microsoft 365 E5 and E7 customers. Microsoft, 18 Jun 2026. learn.microsoft.com
11. Access the Security Copilot audit log. Microsoft, 29 May 2026. learn.microsoft.com
12. Privacy and data security in Microsoft Security Copilot. Microsoft, 8 Apr 2026. learn.microsoft.com
13. Microsoft Security Copilot in advanced hunting. Microsoft, 10 Sep 2026. learn.microsoft.com