AI agents inside the security department

Who reads, who decides, and what is never contained, blocked, closed or reported on an agent's word

IN BRIEF

01 Security operations has two failure modes on the public record, and they pull in opposite directions.

02 In the first, a warning exists and nobody acts on it. An audit office and a legislative committee have each reported a case.

03 In the second, an automated action goes wrong at scale. A faulty content update in July 2024 crashed millions of machines. A security rule has deleted legitimate files.

04 So a security team needs help with the reading. It cannot afford an agent that contains, blocks or closes on its own judgement.

05 A third fact shapes the design. What a security agent reads is written by attackers: alerts, logs and reported emails carry text an attacker chose. Six national cyber agencies say such text can trick an agent into actions nobody authorised.

06 Vendors differ. Some say their agent remediates or contains with no person. Others say a person approves by default. The public templates and playbooks are split: of 14 we opened, 4 put a person's approval before every containing action.

07 This paper is for the security operations centre and the security team around it: alerts, reported emails, incidents, vulnerabilities, identity, cloud configuration, threat intelligence and detection, assurance, and reporting. It covers defensive work only. It ends with seven questions for any vendor and a place to start.

08 The position throughout has four parts. Agents gather and summarise, and an analyst decides. An agent is given read operations, and anything that changes something is a separate tool that waits for a person. The agents treat what they read as material, never as orders, and hold no tool that could act on it without a person. And all of it runs on the organisation's own infrastructure, because logs and incident records describe its weaknesses. No agent isolates a host, disables an account, blocks an address, deletes or quarantines mail, changes a detection or firewall rule, closes an alert or an incident, or decides whether an incident must be reported.

IN THIS PAPER

01 Where security departments are today 2
02 What the public templates do 3
03 Three doors, one gate 4
04 What an agent may do 6
05 What the rules already ask 7
06 A worked example: one retailer's security team 10
07 Seven questions before an agent touches an alert 11
08 Where to start 11

PART 01

Where security departments are today

8.5 million

Windows devices affected by one faulty security content update in July 2024, as the platform vendor estimated. An automated update, not AI

A warning exists and nobody acts on it. The United Kingdom's National Audit Office investigated the 2017 ransomware attack that affected the National Health Service. Its report says: "In March and April 2017, NHS Digital had issued critical alerts warning organisations to patch their systems to prevent WannaCry. However, before 12 May 2017, the Department had no formal mechanism for assessing whether NHS organisations had complied with its advice and guidance" [5].

A tool goes quiet and nobody notices. In December 2018 a committee of the United States House of Representatives reported on the 2017 breach at Equifax. Its release says: "Failure to renew an expired digital certificate for 19 months left Equifax without visibility on the exfiltration of data during the time of the cyberattack" [4].

What those teach. In both, the information existed or should have. Nobody was reading it, or nobody could say who had acted on it. That is work an agent can do: read everything, and say what it could not read.

An automated action goes wrong at scale. On 19 July 2024 CrowdStrike released a content update to its security sensor on Windows machines. The company's root cause analysis says a new template "defined 21 input parameter fields", but the code that used it "supplied only 20 input values to match against". Reading the missing value "resulted in a system crash" [1].

Microsoft, whose operating system the machines ran, wrote the next day: "We currently estimate that CrowdStrike's update affected 8.5 million Windows devices, or less than one percent of all Windows machines" [2].

This was an automated update, not AI. The company's own findings are about release, not detection: "Each Template Instance should be deployed in a staged rollout". One of its fixes is headed "Provide customer control over the deployment of Rapid Response Content updates" [1].

A security rule has deleted legitimate files. In January 2023 Microsoft reported "a series of false positive detections" by one of its own protection rules. "These detections resulted in the deletion of files that matched the incorrect detection logic primarily impacting Windows shortcut (.lnk) files." The corrected build "does not restore deleted shortcuts" [3].

What those teach. A protective action that runs with no person reaches every machine before anyone can stop it. An agent allowed to isolate hosts or change rules has the same reach.

What a security agent reads is written by attackers. An alert carries a file name, a log carries a request, a reported email carries its whole body. Each is text someone outside chose. The joint guidance of six national cyber agencies, published on 1 May 2026, says: "Techniques, such as prompt injection or jailbreaks can trick agents into executing unauthorised actions and bypassing their intended safeguards" [9].

A 2026 preprint tested models that analyse network security logs. Its authors report that layered defences gave "a 90.4% attack reduction, although 8.4% residual vulnerability persists" [6]. Defences help. They do not close the gap.

So the design cannot rest on the agent refusing. It rests on what the agent holds. An agent with no tool that closes, blocks or deletes cannot be talked into closing, blocking or deleting.

Attackers use agents too. In November 2025 Anthropic reported a campaign in which an attacker used AI to carry out much of the work: "The operation targeted roughly 30 entities and our investigation validated a handful of successful intrusions" [7]. In November 2026 Palo Alto Networks' research unit reported "25 operations against at least 19 organizations" run through an AI framework, with "unauthorized access or data retrieval at 10 of the targeted organizations" [8].

What the vendors say. The pages we opened fall into two groups, with several vendors in each.

Some say their agent acts with no person, on some actions or within limits.

- Torq describes an agent that "autonomously remediates critical threats", with response "either completely autonomously or with human-on-the-loop oversight" [10].
- Wiz says teams can "Automatically trigger remediation actions when the Green Agent reaches a high-confidence remediation verdict". Lower-confidence verdicts go "to a developer in Slack for approval" [11].
- Qualys: "high-confidence patches deploy hands-free, risky ones get staged rollout" [12].
- Okta says its identity threat product "automatically triggers a response" and can "remediate the threat without manual intervention" [13].
- Microsoft says of its agent for reported emails that if an alert is a false alarm, "the agent classifies it as a False Positive and resolves it accordingly". A malicious one "remains open and in progress for an analyst to investigate and take further action" [14].
- Palo Alto Networks says its agents "reason, plan and execute workflows to independently resolve incidents", and that customers "Define when agents act independently or require approval for high-impact actions" [15].

PART 01 · CONTINUED

Others say a person approves, or that action is off unless switched on.

- CrowdStrike says its assistant "does not take automated response actions by default" [16].
- Splunk, on high-impact work: "The agent drafts, the human approves before anything is committed" [17].
- Check Point: "Routine operations run autonomously; high impact changes route through human approval" [18].
- Tines: "Put AI to work without giving it the keys" [19].
- Dropzone AI: "Your analysts own every verdict" [20].
- Simbian: "95% of the actions Simbian proposes are approved by your own team" [21].

These are design choices, and each vendor explains them. A buyer should know which actions run alone, and who set the limit.

Where products run. Trend Micro lists an on-premises air-gapped option, "Deployed in an isolated, offline environment" [22]. Fortinet offers "deployment options for both SaaS and self-managed software" [23]. Neither page says whether the agent features run that way. Splunk says its triage agent is "cloud only" [17]. Conifers says "Enterprises deploy inside their own cloud" [24], and Simbian says "Your data never leaves your tenant" [21]. Where the agent and its model run is a question to ask, and more than one vendor can answer it.

PART 02

What the public templates do

4 of 14

public security templates and playbooks we opened put a person's approval before every containing action

We opened 14 public templates, playbooks and examples for security operations. They come from two workflow platforms, two long-standing security automation vendors, three cloud and platform vendors' documentation, one identity vendor's template library, one open-source example and one agent framework's documentation. We read each for one thing: does a person approve before something is blocked, isolated, disabled or deleted?

WHAT THE TEMPLATE DOES	DOES A PERSON APPROVE FIRST?
A model rates the severity of an event. Then: "Auto-blocks IPs, isolates hosts, revokes tokens based on severity"	**No step**
Opens a case, has an agent write an analysis, "and isolates the host pending review"	**No step. The page offers a threshold as an extension**
A vendor's blueprint: "high-confidence malicious verdicts trigger automated containment"	**In part. Low-confidence results go to a person**
Lists unused cloud accounts and prompts an analyst. "If the analyst confirms the action", it disables them	**Yes**
Emails two administrators with Block and Ignore buttons, and is set to "wait until a response is received from the admins"	**Yes**

PART 02 · CONTINUED

The count. 4 of the 14 put a person's approval before every containing action. Two of those are not security playbooks: one handles service outages and one is a framework feature the builder must add. 3 do so in part. 7 have no such step, and 3 of those 7 contain by themselves. The other 4 only notify or open a ticket [25] [26].

Two findings matter more than the count. First, the older playbooks ask and the newer agent templates act. The two that prompt an analyst before disabling or deleting date from 2021 and 2022. The two that isolate or block on a model's rating are recent.

Second, where a template acts alone, the trigger in one of them is the model's own rating. That rating is formed from the alert, and the alert is text an attacker can influence. This is the pattern this paper avoids.

One page shows both designs side by side. It describes a playbook that waits for two administrators, and under the heading "Response to threats immediately with minimal human dependencies" another that tells the endpoint tool to isolate machines [26]. The choice is visible. Ask every product to show the same.

PART 03

Three doors, one gate

An agent is only as useful as what it can reach. In a security department that means three kinds of thing.

Documents.

Playbooks, security policies, the asset list, past incident reviews, detection rules and approved questionnaire answers. These are indexed once as knowledge sources, with each reader's permissions kept, so a colleague does not reach through an agent a case they could not open themselves.

Systems of record.

The alert and log platform, the case system, the vulnerability scanner, the directory, the cloud inventory. Most are reached through an interface the organisation already runs. The organisation gives the platform the description of that interface and picks which operations become tools.

Files and outside sources.

An alert export, a reported email saved as a file, a scan report or a threat report arrives as a file. Tickets, on-call alerts, metrics, logs and chat channels are reached through MCP servers deployed in the organisation's own cluster.

All three pass one gate: every tool call is checked against the organisation's policy, which can allow it, record it, hold it for a person's approval, or block it.

PART 03 · CONTINUED

FOUR WAYS INTO A SECURITY DEPARTMENT'S SYSTEMS

	ROUTE	WHAT IT REACHES	WHAT TO KNOW
01	**A file the user uploads**	An alert export, a reported email, a scan report, a threat report, a customer's questionnaire	Nothing is connected. The safest start
02	**A knowledge source**	Playbooks, policies, the asset list, past incident reviews, detection rules	Permissions are kept per reader
03	**An imported description of an interface**	The alert and log platform, the case system, the scanner, the directory, a public vulnerability catalogue	The organisation picks the operations. Changes wait for approval
04	**An MCP server, deployed in the organisation's cluster**	Tickets, on-call alerts, metrics and logs, chat channels	It runs inside. Check what each tool does

7,540

operations sent as changes, out of 15,450, in the public interface descriptions we imported (107 files read)

What is published. We ran Katonic's importer on 107 public interface descriptions from 46 publishers of systems and public sources a security team uses. 89 import as they are, 16 are over the 5 MB import limit and 2 are refused. The 89 give 15,450 operations. 7,910 read, and 7,540 are sent as changes. The method alone cannot tell which of those only read. The counts are per file, and some publishers split one system into several files or list several versions [27].

Who the publishers are. Identity providers and identity tools, vulnerability scanners and code scanning, one network security vendor, cloud and container platforms, secrets tools, ticket and paging systems, open-source security tools, and public vulnerability sources.

What that exercise taught us. Nearly half of the operations change something. In a security system a change can isolate a machine or lock an account. That is the reason for the approval rule, and for giving an agent the read operations and nothing else. So pick the operations at import, read what each one does, and start with the ones that only read. A description over the limit is imported in the part the organisation needs.

The best-known products publish nothing a file address reaches. That holds for the best-known endpoint detection, log, cloud posture, email security and orchestration products. Most threat-data services publish feeds, not interface descriptions. For those, the description comes from the organisation's own installation, or the agent reads an export or an uploaded file. Our search was made at likely addresses, and a miss does not prove that no description exists behind a sign-in.

PART 04

What an agent may do

One rule runs through every agent in this paper: it prepares, and a named analyst decides.

01	**It reads** the alerts, logs, emails, scan results and records it has been given, through read tools only.
02	**It gathers** what the playbook asks for around one alert: the asset, its owner, the account, earlier alerts on both.
03	**It lists** what is related, missing, overdue or unowned.
04	**It drafts** the summary, the timeline, the incident record, the detection rule, the report. A draft is a file for the analyst. It is not a record and it is not a rule in force.
05	**It treats what it reads as material.** It is instructed to report text in an alert, a log or an email as content, never to follow it. It holds no tool that could act on that text without a person.
06	**It says what it read and what it could not read.** It is instructed not to fill a gap from memory.
07	**It stops** before anything that changes a machine, an account, a mailbox, a rule or a case. A tool that changes something waits for approval unless the organisation has decided otherwise for that tool.

That gives 35 agents in nine areas. The full list is a separate document.

AREA	AGENTS	WHAT THEY PREPARE	WHO DECIDES
Alerts	4	A triage summary with the severity the playbook gives; indicator lookups; related alerts; the shift handover	The analyst and the shift lead
Reported emails	4	The signs the playbook names, read as text; other mailboxes that received the email; a draft reply; a period summary	The analyst. An administrator removes mail
Incidents	4	An incident summary; a timeline with sources; a draft record; a draft review	The incident lead and the review owner
Vulnerabilities	4	The policy's factors beside each finding; systems an advisory touches; expired exceptions; a scan summary	The vulnerability manager and system owners
Identity	4	Sign-ins for a named account in an open case; unexplained administrative rights; dormant accounts; privileged changes	The analyst, the system owner, the identity team
Cloud and configuration	3	Findings by owner; departures from the baseline; exposed resources with no reason on record	The resource owners and the cloud security lead
Threat intelligence and detection	4	A report summary set against the estate; coverage gaps; a draft rule with test cases; a tuning report	The intelligence analyst and the detection engineer
Assurance	4	Draft questionnaire answers; overdue test findings; exceptions with gaps; audit evidence	The security lead and the control owner
Awareness and reporting	4	Answers from the playbooks; a draft awareness note; the security report; the notification clocks	The head of security and the compliance officer

PART 04 · CONTINUED

Four lines are drawn tighter than the rest.

The containing action.

An agent may propose the severity the playbook gives and name the runbook step that fits. It isolates no host, disables no account, blocks no address and removes no email. An analyst decides, and an engineer or administrator acts through a separate tool.

The reported email.

The agent reads a reported email as text: the headers, the sender, the links as written and the attachment names. It is given no tool that opens a link or an attachment.

The closing.

An agent closes no alert and no incident. A closed alert is one nobody will read again, so that decision is a person's.

The person.

The Sign-in Investigation Summary works only on an account an analyst names in an open case. It draws no conclusion about the person. No agent reports on staff behaviour in general, and a post-incident review names no person as the cause.

For reporting, the Notification Deadline Check lists the clocks the organisation's own procedure gives. It does not decide whether an incident must be reported, and it notifies nobody. Every call is recorded, with who asked, which tool and what was decided.

PART 05

What the rules already ask

This part needs a read by the organisation's own lawyers and auditors before it is relied on. It reports what the published texts say; it is a summary, not legal advice. Nothing in it says that a platform makes an organisation secure or compliant. Whether a rule applies depends on the organisation's sector, size and country.

Notification clocks, and who decides

Network security in the European Union. An early warning is due "within 24 hours of becoming aware of the significant incident", a notification "within 72 hours", and "a final report not later than one month" after that. Management bodies approve the security measures and "can be held liable for infringements" [28]. The duty reaches an organisation through national law.

Personal data. A controller notifies the supervisory authority "without undue delay and, where feasible, not later than 72 hours after having become aware of it" [30].

Financial firms in the European Union. The first report is due "within four hours from the classification of the ICT-related incident as a major ICT-related incident and no later than 24 hours from the moment the financial entity has become aware" of it [29].

Listed companies in the United States. A disclosure "must be filed within four business days of determining an incident was material" [32].

Critical infrastructure in the United States. The reporting rule is not final. The agency's page says it "continues to work on the final rule" and that reporting "will not be required until the CIRCIA final rule goes into effect" [34].

Why no agent decides. Each clock starts from a judgement: that the organisation is aware, that an incident is major, that it is material. An agent that classified incidents would be making it. So the agent lists the clocks, and an officer decides.

One case on record. In October 2023 the United States securities regulator sued a software company and its chief information security officer. On 20 November 2025 it filed a joint stipulation "to dismiss, with prejudice, the Commission's ongoing civil enforcement action". The release says the decision "does not necessarily reflect the Commission's position on any other case" [33].

PART 05 · CONTINUED

Least privilege and logging

A United States federal control catalogue states the principle: “Employ the principle of least privilege, allowing only authorized accesses for users (or processes acting on behalf of users) that are necessary to accomplish assigned organizational tasks.” On logging, it asks an organisation to “Identify the types of events that the system is capable of logging in support of the audit function” [35].

An agent is a process acting on behalf of a user. Its access is an account under the same rule, and its calls belong in the same log.

Agents and injected text

The six agencies. The guidance says “strict adherence to the principle of least privilege is critical”, and that “organisations should only use agentic AI for low-risk and non-sensitive tasks”. On approval, it says to ensure that “decisions about when human approval is required are determined by system designers or operators, not delegated to the agentic AI system” [9].

OWASP. Its list of risks for applications built on language models puts prompt injection first. Among its mitigations: “Separate and clearly denote untrusted content to limit its influence on user prompts.” On excessive agency it says to “require a human to approve high-impact actions before they are taken”, and to limit an agent’s tools “to only the minimum necessary” [37].

Both are guidance, not law. Both describe the design in Part 4.

Staff data in security logs

Security logs hold staff data: who signed in, from where, and what they opened. European data protection law lets member states set rules for employees’ data, with regard to “monitoring systems at the work place” [30]. A 2017 opinion of the European data protection authorities says “a proportionality test should be undertaken prior to its commencement”, and that “employees should receive effective information about the monitoring that takes place” [31]. The opinion predates the current regulation and is guidance.

That is why the sign-in agent is tied to a named account in an open case.

PART 05 · CONTINUED

Evidence

A United States guide on forensics describes “collection, examination, analysis, and reporting”, and speaks of “maintaining a strict chain of custody for the data” [36]. The international standard for digital evidence covers its “identification, collection, acquisition and preservation” [36]. We read the standard’s public summary, not its text.

An agent’s summary is not the evidence. The records are, and the agent changes none of them.

The AI Act

There is no entry called security operations. Two entries can still matter. One covers systems meant “to monitor and evaluate the performance and behaviour of persons” at work. Another covers “safety components in the management and operation of critical digital infrastructure”. Where a system is high-risk, “Deployers shall assign human oversight to natural persons who have the necessary competence, training and authority” [38]. Whether an entry catches a particular use is for counsel.

The dates have moved. Regulation (EU) 2026/1744 says the high-risk rules apply from “2 December 2027 as regards AI systems classified as high-risk pursuant to Article 6(2) and Annex III” and from 2 August 2028 for those under Article 6(1) and Annex I [39].

PLACE	WHAT WE READ	WHAT IT ASKS
European Union	Network security, data protection and financial resilience law	First notices in 24, 72 and 24 hours [28] [29] [30]
United States, listed companies	The incident disclosure rule	Four business days from deciding an incident is material [32]
United States, federal catalogue	Access and audit controls	Least privilege for “processes acting on behalf of users” [35]
Six national cyber agencies	Guidance on agentic AI	People, not the agent, decide when approval is needed [9]

We did not open the payment card standard, the text of the ISO standards for security management and incident handling, the Indian, Gulf, Singapore or Australian reporting rules, or the proposed United Kingdom law. An organisation should check each with its own advisers.

PART 05 · CONTINUED

What the rules have in common

01	**A person decides what is reported.** Every clock starts from a judgement.
02	**Access is the least the task needs.** That applies to an agent as to a person.
03	**What an agent reads is not trusted.** High-impact actions wait for a person.
04	**Records stay as they were found.** A summary is not evidence.

PART 06

A worked example: one retailer's security team

Take a made-up retailer: 6,000 staff, a security team of 14, six of them on shifts in security operations. Nothing here is a customer's story, and every figure is illustrative; it is section 4 put to work.

01

On an alert at night, the agent gathers what the playbook asks for: the laptop, its owner, the account and 3 earlier alerts on the same machine. It proposes the severity the playbook gives. One log entry contains a sentence worded as an instruction. The agent reports it as part of the entry. It holds no tool that closes an alert. The analyst confirms the severity and asks the engineer on call to isolate the laptop.

02

On a reported email, the agent lists 4 signs the playbook names and 23 other mailboxes that received the same message. The analyst decides it is fraud. An administrator removes the messages. The agent drafts the reply to the colleague who reported it, and the analyst sends it.

03

After a scan, the agent sets the policy's factors beside 1,840 findings. 12 are on systems that face the internet and are recorded as exploited in a public catalogue. It reports that it could not read the asset list for one warehouse. The vulnerability manager sets the order and asks for that list.

04

During an incident, the agent builds the timeline from the case record and the channel, with the source of each entry. It lists the two notifications the retailer's own procedure names and the time each clock would end. The compliance officer decides whether either applies.

Every one of these changes nothing by itself, contains nothing, closes nothing, and leaves a record of who asked, which tool ran and what was decided.

PART 07

Seven questions before an agent touches an alert

01 **Where does it run?**
On our infrastructure, in our private cloud, or in the vendor's service? Where is the model, and do our logs, alerts and incident records leave? More than one vendor can run on your side. Ask the next six of each.

02 **What can it read?**
Does it keep each reader's permissions, or does it read every case and mailbox as one all-seeing account?

03 **What can it change?**
List every tool that isolates, disables, blocks, deletes, quarantines, changes a rule or closes. Is each one separate from the read tools, and does each wait for a person?

04 **What happens when it reads an attacker's text?**
Which tools does the agent hold at that moment, and could any of them act without a person?

05 **Who decides what is reported?**
Which named officer decides that an incident is notifiable or material, and what does the agent give them?

06 **What is recorded?**
Who asked, which tool ran, what was decided, for how long is that kept, and can an auditor read it?

07 **Can we stop it?**
One agent, one tool or one server, at once, by us. And who can change which actions run alone?

PART 08

Where to start

Pick work where reading is slow, the decision stays with an analyst, and nothing is contained or closed on the agent's word.

Questions and answers on the playbooks
Analysts ask what the playbook says and get the answer with the passage cited. No system is touched.

Alert triage summary
The agent gathers what the playbook asks for around one alert. An analyst decides.

Vulnerability prioritisation preparation
The agent sets the policy's factors beside each finding. The vulnerability manager sets the order.

One team, one kind of alert, invented data first. Then beside the manual process, measured against it.

About Katonic AI

Katonic AI builds the platform that organisations use to build, run and govern AI agents on their own infrastructure: in their own data centre, in a private or sovereign cloud, or with no connection to the internet at all. The platform is in production with governments, banks, telecom operators and utilities. Katonic is certified to ISO 27001.

KATONIC.AI/SOLUTIONS/SECURITY

SOURCES

Read on 23 November 2026. Figures are the publishers' own and were not audited by us.

01 CrowdStrike, "External Technical Root Cause Analysis: Channel File 291", 6 August 2024. https://crowdstrike.com/wp-content/uploads/2024/08/Channel-File-291-Incident-Root-Cause-Analysis-08.06.2024.pdf

02 Microsoft, "Helping our customers through the CrowdStrike outage", 20 July 2024. https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/

03 Microsoft, "Recovering from Attack Surface Reduction rule shortcut deletions", 14 January 2023, updated 23 January 2023. https://techcommunity.microsoft.com/blog/microsoftdefenderatpblog/recovering-from-attack-surface-reduction-rule-shortcut-deletions/3716011

04 United States House of Representatives, Committee on Oversight and Government Reform, release "Committee Releases Report Revealing New Information on Equifax Data Breach", 10 December 2018. https://oversight.house.gov/report/committee-releases-report-revealing-new-information-on-equifax-data-breach

05 National Audit Office, "Investigation: WannaCry cyber attack and the NHS", HC 414, 24 October 2017 (the file also carries a later date of 24 April 2018). https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf

06 Karanjai and others, "Context Contamination in LLM Analysis of Network Security Logs", preprint, arXiv 2607.14493, submitted 16 July 2026. https://arxiv.org/abs/2607.14493

07 Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign", full report, November 2025. https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf ; and the news page https://www.anthropic.com/news/disrupting-AI-espionage

08 Palo Alto Networks, Unit 42, report of 12 November 2026 on an AI framework used in intrusions. https://unit42.paloaltonetworks.com/cairn-framework-ai-autonomous-attacks/

09 Australian Signals Directorate's Australian Cyber Security Centre with the cyber agencies of the United States (two), Canada, New Zealand and the United Kingdom, "Careful adoption of agentic AI services", 1 May 2026, read in the copy hosted by New Zealand's National Cyber Security Centre. https://www.ncsc.govt.nz/assets/guidance/Documents/Careful-adoption-of-agentic-AI-services_FINAL.pdf ; the New Zealand page https://www.ncsc.govt.nz/protect-your-organisation/careful-adoption-of-agentic-ai-services/ ; the Canadian notice https://www.cyber.gc.ca/en/news-events/joint-guidance-careful-adoption-agentic-artificial-intelligence-services

10 Torq, home page, undated. https://torq.io/

11 Wiz, blog post of 23 March 2026. https://www.wiz.io/blog/introducing-wiz-agents

12 Qualys, product page, undated. https://www.qualys.com/apps/agentic-ai

13 Okta, Identity Threat Protection product page, undated. https://www.okta.com/products/identity-threat-protection/

14 Microsoft Learn, page on the Phishing Triage Agent, undated in the text read. https://learn.microsoft.com/en-us/defender-xdr/phishing-triage-agent

15 Palo Alto Networks, Cortex AgentiX product page, undated. https://www.paloaltonetworks.com/cortex/agentix

16 CrowdStrike, product page (frequently asked questions), undated. https://www.crowdstrike.com/en-us/platform/charlotte-ai/

17 Splunk, blog post of 5 November 2026. https://www.splunk.com/en_us/blog/artificial-intelligence/how-splunk-soc-uses-agents.html

18 Check Point, product page, undated. https://www.checkpoint.com/quantum/unified-management/agentic-network-security/

19 Tines, home page, undated. https://www.tines.com/

SOURCES continued

20 Dropzone AI, home page, undated. https://www.dropzone.ai/

21 Simbian, home page, undated. https://simbian.ai/

22 Trend Micro, deployment options page, undated. https://www.trendaisecurity.com/en/platform/deployment-options

23 Fortinet, FortiSOAR product page, undated. https://www.fortinet.com/products/fortisoar.html

24 Conifers, home page, undated. https://conifers.ai/

25 Three workflow templates on n8n: 10587 (incident response with containment), 1975 (notice of a quarantined email) and 2665 (suspicious email analysis). https://n8n.io/workflows/10587 ; https://n8n.io/workflows/1975 ; https://n8n.io/workflows/2665

26 Eleven further templates, playbooks and examples: Tines, phishing response library page. https://www.tines.com/library/phishing-attack-response ; Tines, blueprint of 22 July 2026. https://www.tines.com/blog/phishing-response-workflow-a-blueprint-for-security-teams/ ; Splunk, playbook post of 21 January 2021. https://www.splunk.com/en_us/blog/security/splunk-soar-playbooks-finding-and-disabling-inactive-users-on-aws.html ; Palo Alto Networks, playbook post of 1 July 2022. https://www.paloaltonetworks.com/blog/security-operations/playbook-of-the-week-phishing-for-trouble-with-cortex-xsoar ; Microsoft Learn, playbook recommendations for Sentinel. https://learn.microsoft.com/en-us/azure/sentinel/automation/playbook-recommendations ; Microsoft, the Isolate-MDEMachine playbook. https://github.com/Azure/Azure-Sentinel/tree/master/Solutions/MicrosoftDefenderForEndpoint/Playbooks/Isolate-MDEMachine ; Elastic, alert triage workflow guide. https://www.elastic.co/docs/explore-analyze/workflows/use-cases/security/automate-security-operations/ai-driven-alert-triage ; Amazon Web Services, Automated Security Response solution overview. https://docs.aws.amazon.com/solutions/latest/automated-security-response-on-aws/solution-overview.html ; an open-source incident example built on an agent framework. https://github.com/krishnashakula/IRAS ; CrewAI documentation, human feedback page. https://docs.crewai.com/en/learn/human-feedback-in-flows ; Okta, a Workflows template. https://github.com/okta/workflows-templates/blob/master/workflows/monitor_unsuccessful_phishing_attempts/readme.md

27 107 public interface descriptions from 46 publishers, run through the platform's importer on 23 November 2026 (Katonic research); and the search for publishers with no downloadable description (Katonic research).

28 Directive (EU) 2022/2555 (NIS2), Articles 20, 21 and 23. http://publications.europa.eu/resource/celex/32022L2555

29 Regulation (EU) 2022/2554 (Digital Operational Resilience Act), Article 19; and Commission Delegated Regulation (EU) 2025/301 on the time limits for incident reports. http://publications.europa.eu/resource/celex/32022R2554 ; http://publications.europa.eu/resource/celex/32025R0301

30 Regulation (EU) 2016/679 (General Data Protection Regulation), Articles 33(1) and 88(2). http://publications.europa.eu/resource/celex/32016R0679

31 Article 29 Data Protection Working Party, Opinion 2/2017 on data processing at work (WP 249), adopted 8 June 2017. https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=45631

32 United States Securities and Exchange Commission, "Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure", Federal Register, 4 August 2023 (document 2023-16194). https://www.govinfo.gov/content/pkg/FR-2023-08-04/pdf/2023-16194.pdf

33 United States Securities and Exchange Commission, Litigation Release No. 26423, 20 November 2025. https://www.sec.gov/enforcement-litigation/litigation-releases/lr-26423

34 Cybersecurity and Infrastructure Security Agency, page on the Cyber Incident Reporting for Critical Infrastructure Act of 2022, undated. https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/cyber-incident-reporting-critical-infrastructure-act-2022-circia

35 National Institute of Standards and Technology, SP 800-53 Revision 5, controls AC-6 and AU-2. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

36 National Institute of Standards and Technology, SP 800-86, "Guide to Integrating Forensic Techniques into Incident Response", August 2006. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-86.pdf ; and International Organization for Standardization, catalogue page for ISO/IEC 27037:2012. https://www.iso.org/standard/44381.html

37 OWASP, Top 10 for LLM Applications 2025, LLM01 Prompt Injection and LLM06 Excessive Agency. https://genai.owasp.org/llmrisk/llm01-prompt-injection/ ; https://genai.owasp.org/llmrisk/llm062025-excessive-agency/

38 Regulation (EU) 2024/1689 (AI Act), Article 26(2) and Annex III points 2 and 4(b). http://publications.europa.eu/resource/celex/32024R1689

39 Regulation (EU) 2026/1744 of 8 July 2026, the replacement of Article 113(c) of Regulation (EU) 2024/1689. http://publications.europa.eu/resource/celex/32026R1744