



SPECULO · AI, CYBER & RISK

AI Controls Maturity Reference

Every AI-related control in the Speculo Control Catalogue (SCC): what to check, how to assess it, and what each maturity level looks like in practice.

12

CONTROLS

72

MATURITY LEVELS

L0–L5

MATURITY RANGE

Extracted from the Speculo Control Catalogue (Source/scc-catalog/domains/). Covers AI as an attack surface (deepfakes, synthetic media, AI-crafted phishing, AI-driven social-engineering agents), AI as an attacker's tool (AI-accelerated attack pattern recognition), and AI as an organisational asset and defensive capability (model access controls, adversarial robustness, ML-based detection, staff AI-threat awareness). One control (AIS-AGT-030) is flagged in the source catalogue as pending final review; content is otherwise as published. Most domains carry DRAFT status in the source catalogue pending full QA.

Contents



1.	AIS-AGT-010	Content authenticity verification	AI-Enhanced Attacks Defense
2.	AIS-AGT-020	Natural language processing security	AI-Enhanced Attacks Defense
3.	AIS-AGT-030	AI-powered attack pattern recognition	AI-Enhanced Attacks Defense
4.	AIS-AGT-040	Deepfake and synthetic media protection	AI-Enhanced Attacks Defense
5.	AIS-AGT-050	AI social engineering agent detection	AI-Enhanced Attacks Defense
6.	AIS-AGT-060	Biometric spoofing detection	AI-Enhanced Attacks Defense
7.	AIS-AGT-070	Deepfake detection systems	AI-Enhanced Attacks Defense
8.	AIS-AGT-080	Synthetic media verification	AI-Enhanced Attacks Defense
9.	PLS-SAT-050	AI threat awareness training	People
10.	SAS-AIM-010	Adversarial training and robustness	Service and Application Security
11.	SAS-AIM-020	AI model access controls	Service and Application Security
12.	RRC-DET-040	Machine learning-based detection	Response and Recovery

Content authenticity verification

AI-Enhanced Attacks Defense Scope: Dual Range: L0-L5



The organisation implements systems and processes to verify that emails, documents, and communications genuinely originate from the claimed party before staff or systems act on them. This includes authentication of mail paths (e.g. DMARC-aligned sending), callback and out-of-band confirmation for high-risk instructions, and where relevant document signing or provenance checks. It complements technical email security and identity controls, and requires both a defined policy and active technical enforcement to be effective.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	Nobody checks whether an email, document, or message actually came from who it claims to be from. People trust what looks right.	Ask staff how they decide if an email or document is genuine. Look for any written guidance. Optionally sample recent inbound payment or credential requests and ask how they were validated. If there is no guidance and people say they rely on appearance alone, this is L0.	A company uses its cloud email and productivity platform with no DMARC, DKIM, or SPF enforcement. Staff pay a spoofed invoice because the message looked legitimate.
1	Informal	Some staff call back or check with the sender when something feels off, but there are no agreed steps, tools, or responsibilities; it depends on the individual.	Ask for examples of when someone verified a message. Check for any written procedure or assigned owner. If verification happens only when individuals choose to, this is L1.	A finance team member calls the CEO to confirm a wire transfer after it felt odd, but only because that person is cautious; there is no policy requiring verification.
2	Planned & Tracked	Certain high-risk teams (e.g. finance, legal) have written steps and approved tools for verifying important messages before acting. Other parts of the business do not follow the same process.	Request documented verification procedures for high-risk teams. Check which tools are approved and whether training records exist. Confirm the process is not yet organisation-wide.	The finance team uses external-sender warnings in the email client and a written checklist for payment requests above a set threshold. Marketing has no equivalent process.
3	Standardised	There is an organisation-wide policy that defines when and how authenticity must be checked. Technical controls such as email authentication (DMARC enforcement where feasible), gateway verification, or document provenance checks are in place and managed centrally.	Review the organisation-wide policy. Verify technical controls are deployed. Ask for the exceptions register and governance records. Sample a week of mail-flow or authentication reports if available.	The organisation enforces DMARC reject on its domains, uses the email security platform's advanced impersonation and phishing protection features, and requires out-of-band callback for all payment instruction changes.
4	Quantitatively Controlled	The organisation tracks how well verification works: for example spoof or impersonation catch rates, time to complete verification, and false positives. These measures drive changes to tools, rules, or training.	Ask for dashboards or reports on impersonation catch rates, verification times, and false positive rates. Confirm documented changes resulted from those metrics in a recent period.	The security team uses the email security platform's investigation console monthly. After tuning impersonation rules, a report shows a large drop in false positives (illustrative outcome).
5	Optimising	Verification keeps pace with new forgery methods including AI-generated content. Threat intelligence, simulations, and incident feedback drive rapid updates to tools and procedures.	Ask how the organisation responded to the last novel forgery technique (e.g. real-time voice cloning). Look for intelligence feeds, exercise results, and policy or tooling updates within weeks, not only annual cycles.	After synthetic-voice payment fraud trends appear in sector intelligence, the organisation updates callback rules within two weeks, runs a tabletop exercise, and shares indicators with its managed security provider.

Natural language processing security

AI-Enhanced Attacks Defense Scope: Dual Range: L0–L5



This control requires email and messaging security to detect or resist AI-generated and highly fluent phishing and social-engineering content, not only obvious spam. It covers the configuration of detection features, operating procedures, and messaging security settings aimed at linguistically sophisticated lures. Authenticity and origin verification of messages, and broader SOC-level automated attack detection, are handled by separate controls in this domain.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	Email protection is limited to basic spam filtering. The organisation has not assessed whether protections address fluent, AI-assisted phishing.	Review email security configuration with the messaging admin. If only baseline anti-spam is enabled and no one has evaluated AI-capable or impersonation features, this is L0.	The organisation relies on the default spam filter in its cloud email platform. A fluent, personalised phishing message compromises an account because no impersonation or advanced detection layer is enabled.
1	Informal	Security staff sometimes notice phishing that looks unusually polished, but there are no detection rules, training, or playbooks aimed at AI-written or high-quality lures.	Ask whether the team has identified suspiciously fluent or tailored phishing. Check for documented rules, training, or playbooks. If awareness exists but nothing is formalised, this is L1.	A SOC analyst escalates an email that reads too well for typical phishing, but no rule or process captures similar messages automatically.
2	Planned & Tracked	Some email or messaging tools are configured for advanced impersonation or AI-assisted phishing detection in parts of the business. Playbooks exist for the teams that use those features.	Ask which impersonation or AI-phishing features are enabled and for which units. Request playbooks. If coverage is partial, this is L2.	Impersonation protection and mailbox intelligence run for executives and finance in the cloud email platform; other departments use basic protection only.
3	Standardised	Enterprise email and messaging are configured to detect AI-assisted or high-quality phishing organisation-wide. Policies cover reporting, quarantine, and hand-off to fraud or identity teams.	Review enterprise configuration for impersonation and advanced phishing features. Trace reporting, quarantine, and fraud or identity integration end to end.	Impersonation and advanced phishing features are on for all mailboxes. Users report via the client report button into the security team's triage queue; runbooks link to fraud.
4	Quantitatively Controlled	The organisation measures catch versus miss rates for sophisticated phishing, false positives, and reporting quality, and uses that data on a regular cycle to update rules and guidance.	Request reports on detection rates, false positives, and user reporting. Ask for evidence of rule or training updates driven by those metrics in the last six months.	Weekly reviews of investigation tooling show impersonation trends; after a spike in missed BEC-style mail, thresholds and safe-sender logic are adjusted and missed volume drops (illustrative).
5	Optimising	Defences adapt to new AI phishing tactics (e.g. multilingual or deeply personalised lures). Threat intelligence and shared indicators support rapid detection updates.	Ask how the organisation adapted to a recent novel AI phishing pattern. Look for intelligence ingestion, automated indicator use, and updates within days or weeks.	Overseas offices see multilingual AI phishing; the team updates rules within a week using ISAC or vendor intelligence feeds.

AI-powered attack pattern recognition

AI-Enhanced Attacks Defense Scope: Dual Range: L0–L5



This control requires the organisation to develop and operate detection and defence specifically against AI-assisted and highly automated attack patterns, including credential stuffing at scale, scripted reconnaissance, machine-paced password spraying, and coordinated bot activity. Coverage spans SIEM, EDR, cloud identity logs, and custom analytics designed to identify behaviour that exceeds human-paced attack rates. The scope should explicitly target automated and AI-accelerated behaviour rather than generic malware alerts, which distinguishes this control from standard SOC maturity.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	The organisation cannot distinguish automated or AI-accelerated attack behaviour from normal activity in a structured way. Only default vendor alerts exist.	Ask whether detection content targets automated or AI-assisted attack patterns. If the team relies solely on generic defaults with no tailored content, this is L0.	A SIEM runs only out-of-the-box rules. A credential-stuffing wave triggers account lockouts before anyone correlates the pattern.
1	Informal	Analysts occasionally spot automation (e.g. very fast logins across accounts) but observations are not turned into saved queries or detection rules.	Ask for examples of spotted automation. Check whether those became durable detection content. If not, this is L1.	An analyst posts about rapid API calls in chat; no rule is created and the insight is not retained.
2	Planned & Tracked	Specific analytics or rules address automated attack patterns in key systems (e.g. brute-force, impossible travel) but coverage does not span all major environments.	Request a list of rules or analytics aimed at automated attacks and which environments they cover. If only some environments or log sources are included, this is L2.	Multi-signal correlation and identity UEBA run for the primary cloud directory; other cloud workloads lack equivalent content.
3	Standardised	A managed library of detection content targets high-speed and AI-assisted attacks. Content is version-controlled, reviewed on a schedule, and spans SIEM, EDR, and primary cloud monitoring with a named owner.	Review the rule library, change history, and coverage map across SIEM, EDR, and cloud. Confirm ownership and review cadence.	Custom analytics and vendor UEBA cover identity and core cloud sign-in across multiple providers, under change control with a named detection lead.
4	Quantitatively Controlled	The team measures detection efficacy, time to detect, and coverage gaps. Purple-team or red-team exercises include automated attack scenarios; findings update the backlog.	Ask for efficacy reports and recent exercise results focused on automated or scripted attacks. Confirm backlog items closed from those results.	Quarterly exercises simulate credential stuffing and automated lateral movement; a high share of scenarios generate alerts and gaps are tracked to remediation (illustrative percentage).
5	Optimising	Detections evolve continuously from threat intelligence, hunting, and cross-domain correlation (identity, endpoint, network, cloud). Updates deploy in days or weeks, not only annually.	Ask how quickly content changed after the last relevant technique was observed. Look for cross-source correlation and intelligence-driven automation or rapid release.	After a new AI-assisted reconnaissance tool is reported, updated analytics deploy within days; identity anomalies correlate automatically with firewall or proxy alerts.

Deepfake and synthetic media protection

AI-Enhanced Attacks Defense Scope: Enterprise Range: L0–L5



This control requires the organisation to reduce harm from deepfake video, synthetic audio, and AI-generated imagery used to manipulate staff or customers. It focuses on policies, training, and procedural verification for when sensitive instructions arrive over voice, video, or rich media channels. Technical detection systems for scoring synthetic media, and the governance workflows for verifying media provenance, are covered by separate controls in this domain.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	There are no rules, training, or tools addressing fake voice, video, or AI-generated images used for fraud or manipulation.	Ask whether policy, training, or tooling covers synthetic media. If none, this is L0.	A CFO accepts a video call that appears to be the CEO requesting an urgent transfer; there is no verification policy. The media was synthetic.
1	Informal	People know deepfakes exist from news or anecdotes but there are no standing procedures or tools; response is improvised.	Ask high-risk roles what they would do if executive video or voice seemed suspicious. If they would comply without a standard, this is L1.	A company-wide email warns about deepfakes after a public fraud story; no procedures or tools follow.
2	Planned & Tracked	High-value teams (e.g. treasury, executive support) have documented steps to verify voice or video instructions (e.g. callback on a directory-listed number) before acting. Not organisation-wide.	Request procedures and training records for those teams. Confirm scope is limited.	Finance must verify payment instructions from voice or video via callback to a known number; other departments have no standard.
3	Standardised	Organisation-wide expectations exist for verifying sensitive instructions delivered by voice, video, or rich media. Procedural and technical safeguards apply wherever those channels are used.	Review policy, channel coverage list, and technical controls (e.g. meeting policies, external participant indicators).	Callback policy applies to all departments that handle sensitive transfers. Collaboration policies flag external participants; synthetic-media awareness is in annual training.
4	Quantitatively Controlled	Synthetic-media incidents and near-misses are tracked. The organisation reviews whether safeguards work (e.g. callback compliance, attempted fraud outcomes).	Ask for a register of synthetic-media-related events and management reviews of control effectiveness.	Most voice-based payment requests follow callback procedure last quarter; known fraud attempts were blocked at verification; results go to the risk committee (illustrative compliance rate).
5	Optimising	Protections refresh as synthetic-media quality improves. Intelligence on new tools, exercises, and cross-functional coordination (security, fraud, HR, communications) drive updates.	Ask how protections changed after a major public technique shift. Look for exercises, intelligence use, and coordinated rollout.	When real-time voice cloning tools gain attention, the organisation runs a tabletop within a month, adds a shared passphrase for high-value calls, and briefs staff within two weeks.

AI social engineering agent detection

AI-Enhanced Attacks Defense Scope: Dual Range: L0–L5



This control requires the organisation to detect and block AI-driven or highly automated agents used for social engineering via web chat, support portals, collaboration tools, and similar channels. The focus is on bot-like manipulation that goes beyond generic spam, including scripted or AI-driven attempts to deceive users, extract information, or influence decisions. It works alongside channel hardening and identity verification controls rather than replacing them.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	The organisation has not treated AI-powered or automated chat agents as a social-engineering threat on customer or internal channels.	Ask whether bot or automated-agent abuse is in scope for any channel risk assessment. If not, this is L0.	Support live chat has no bot mitigation; an automated script socially engineers a reset of a customer account.
1	Informal	Staff sometimes report odd bot-like chats; security handles cases one-off with no detection rules or blocks.	Ask support or collaboration owners for reports of bot-like abuse. If responses are anecdotal only, this is L1.	A support agent reports a customer chat that felt like a bot; security reviews that ticket only.
2	Planned & Tracked	Bot or automation controls exist on some channels (e.g. customer portal) but not across all relevant platforms.	List channels with bot detection, CAPTCHA, or rate limits. Identify gaps.	A CAPTCHA or bot score product protects the public portal; internal collaboration channels lack equivalent controls.
3	Standardised	Controls to detect or block abusive automated or AI-like agents cover the main external and internal channels. Owners, escalation paths, and abuse reporting are defined.	Review configuration across web, support, and collaboration stacks. Confirm ownership and abuse workflows.	Bot management on web properties, restrictions on unapproved bots in collaboration suites, and behavioural signals in the support platform are owned by a named team.
4	Quantitatively Controlled	Per-channel metrics (blocks, abuse volume, false positives) drive tuning on a schedule.	Ask for metrics and evidence of threshold or policy changes driven by data.	Monthly reports show automated interaction blocks and false-positive rate; thresholds are adjusted after review (illustrative volumes).
5	Optimising	Defences adapt quickly to new agent frameworks and tactics. Simulations and peer or vendor intelligence feed updates.	Ask how the team responded to a novel agent tactic. Look for red-team style tests and rapid signature or policy updates.	After a new agent toolkit appears in threat reporting, signatures update within two weeks and a controlled test is run against the support portal; findings go to the ISAC.

Biometric spoofing detection

AI-Enhanced Attacks Defense

Scope: Dual

Range: L0–L5



This control requires biometric authentication deployments to resist presentation attacks: photos, masks, replay video, and synthetic faces used to bypass the sensor. The organisation must specify liveness or anti-spoofing capabilities when procuring or deploying biometrics, test vendor claims, and update controls as attack techniques evolve. This applies wherever biometrics are used to gate access to systems, applications, or data.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	Biometrics are used but the organisation accepts vendor defaults without assessing spoofing resistance.	Ask whether spoofing or presentation attacks were assessed for in-scope biometric systems. If defaults only, this is L0.	Facial unlock on laptops is enabled with no review of replay or photo attack resistance.
1	Informal	The team knows spoofing is possible but has not formally assessed in-scope biometric systems.	Check risk register or meeting notes. If no assessment artefact exists, this is L1.	A news story prompts a discussion; no test or requirement change follows.
2	Planned & Tracked	New biometric deployments must include anti-spoofing or liveness; legacy systems may lag.	Review standards for new systems. Sample legacy inventory for gaps.	New mobile apps require vendor liveness APIs; older camera-only face unlock on some devices is not yet assessed.
3	Standardised	All authentication biometrics must meet an organisation-wide anti-spoofing standard appropriate to risk. Configuration is centralised and reviewed.	Review standard, compliance evidence, and central register of systems.	Policy requires liveness for all biometric login; hardware-backed or depth-camera options are approved; exceptions are rare and approved.
4	Quantitatively Controlled	Spoof attempts, bypasses, and model drift are tracked; failures trigger vendor or architectural response.	Request metrics and examples of remediation after spikes or advisories.	Quarterly review of sign-in telemetry; after a spoofing advisory, the fleet is tested and mitigations are documented.
5	Optimising	Red-team testing, vendor roadmaps, and research drive rapid compensating controls or upgrades when new spoof techniques appear.	Ask for the last published attack technique and the organisation's test and response timeline.	A new presentation-attack method is published; red team validates impact within a month; high-risk users get a second factor until patches or config updates land.

Deepfake detection systems

AI-Enhanced Attacks Defense Scope: Enterprise Range: L0–L5



This control requires the organisation to deploy and operate AI-powered or specialised systems that score, flag, or analyse video, audio, and images to determine whether media is likely synthetic. It is about operational detection capability: the tools, models, and service ownership needed to run detection consistently. Policy and procedural protections for synthetic media, and the business process governance for verifying media authenticity in decisions, are covered by other controls in this domain.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	No technical capability exists to assess whether media is synthetic; decisions rely only on human judgement of the file.	Ask fraud or security what they run on suspicious executive or customer media. If nothing, this is L0.	Fraud receives a customer video for a high-value case and has no tool; they accept it as genuine.
1	Informal	Someone used a free or trial detector once; there is no supported tool, logging, or process.	Ask if ad hoc tools were used. If one-off with no standard, this is L1.	An analyst tries a public detector on audio; result is inconclusive and the tool is not used again.
2	Planned & Tracked	Supported detection tools are used by named teams for defined scenarios. Training exists; coverage of media types or teams is partial.	List teams, tools, scenarios, and training. Confirm limits.	Fraud investigations use a specialist video analysis workstation; audio-only cases are out of scope.
3	Standardised	A production detection service (vendor or internal) has ownership, access control, logging, and privacy documentation for the media types that matter.	Review runbooks, data handling, and access logs.	A cloud content-safety API screens some channels; a specialist product supports HR and fraud cases; SOC owns access reviews and DPIA-style records.
4	Quantitatively Controlled	Accuracy, false positives, and analyst handling time are tracked; when generators improve, models or rules are refreshed.	Ask for accuracy or calibration reports and vendor or internal retraining evidence after generator shifts.	Lab testing shows strong detection on a labelled set; after new generator classes appear, accuracy drops and triggers a model or vendor update (illustrative percentages).
5	Optimising	Benchmarking against new synthetic samples, cross-team learning, and automated routing of high-risk media to experts run on a continuous cadence.	Ask how often benchmarks run and how triage automation works.	A synthetic-media intel feed and quarterly lab benchmarks feed tuning; workflow routes high-confidence synthetic flags to a specialist within a defined SLA.

Synthetic media verification

AI-Enhanced Attacks Defense

Scope: Enterprise

Range: L0–L5



This control requires the organisation to verify that media used for business decisions (identity proofing, external communications, legal or investor material) is authentic, with audit trails where required. It covers the workflows, approval steps, metadata checks, and provenance standards (such as C2PA) that govern how verification is performed and recorded. Detection technology for analysing whether media is synthetic sits with a separate control, while this one focuses on the governance layer and how verification decisions are made and recorded.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	No process checks whether photos, video, or audio used for decisions are genuine; media is trusted by default.	Ask how hiring, comms, or legal teams validate submitted media. If there is no method, this is L0.	HR files a scanned ID from email without checking for manipulation.
1	Informal	Some staff check metadata or ask for confirmation; no standard, so results vary.	Ask whether checks are required or personal habit. If inconsistent, this is L1.	One comms staffer checks image metadata before publish; others do not.
2	Planned & Tracked	Documented workflows apply to specific high-stakes scenarios (e.g. executive recordings, KYC images) with nominated tools and records of checks.	Request workflows, tools, and evidence logs. Confirm limited scope.	Corporate comms uses a checklist and metadata review before publishing executive video; scope is external comms only.
3	Standardised	Policy defines media classes that require verification. Tooling, approvals, and audit expectations apply organisation-wide.	Review policy, tooling, and sample audit logs across two departments.	Identity, external comms, and legal media require verification; C2PA-style credentials are validated where present; other formats follow manual steps; logs retained.
4	Quantitatively Controlled	Cycle time, error rate, and exceptions are measured and used to improve process and tooling.	Ask for metrics and examples of process or tool changes driven by data.	Average verification time and error rate are reported; a pilot reduces bottleneck steps for video (illustrative times).
5	Optimising	Verification evolves with new channels and forgery methods; provenance standards are adopted or evaluated; incidents and industry changes trigger fast updates.	Ask for recent updates after an incident or standard release.	Organisation adopts broader content-credential support in intake workflows; after sector news on fake earnings audio, investor-relations adds an audio verification step within a short window.

AI threat awareness training

People Scope: Enterprise Range: L0–L5



This control requires the organisation to educate personnel about AI-powered attacks, including AI-generated phishing, deepfake audio and video, automated social engineering, and the risks of interacting with manipulated AI systems. AI attack techniques are evolving rapidly, and industry practice at the higher maturity levels of this control is still developing. Technical detection of AI-enabled threats is covered in the AI-Enhanced Attacks Defense domain; training materials for this topic are produced through the content creation and management control.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	Personnel receive no training on AI-powered threats. There is no awareness that AI can generate convincing phishing, deepfakes, or automated social engineering.	Ask whether any training or communication has addressed AI-powered threats (deepfakes, AI-generated phishing, voice cloning). If none, this is L0.	Staff assume a call from the CEO is genuine and a well-written email is trustworthy. Nothing has been communicated about AI-related threats.
1	Informal	Some staff are aware of AI threats from news articles or external events, but there is no organisational training or guidance on the topic.	Ask staff whether they are aware of AI-powered threats and where they learned about them. If awareness comes from personal initiative rather than organisational training, this is L1.	A few staff share deepfake news articles in a team chat, but security has produced no guidance and most staff remain unaware of the risk.
2	Planned & Tracked	The organisation has produced at least one tracked training module on AI-generated phishing, deepfakes, or voice cloning, covering basic recognition and second-channel verification techniques.	Request the AI threat training content. Check when it was created and whether delivery is tracked. Confirm that it covers at least the basics of AI-generated threats and verification techniques.	The annual training programme gains a module explaining deepfakes and AI-generated phishing, teaching staff to verify unexpected requests by phone or in person. Completion is tracked.
3	Standardised	AI threat awareness is a standing topic in the security awareness programme, updated as techniques evolve. Role-specific guidance covers finance and executives, and verification procedures for high-risk requests are formalised and trained.	Review how AI threat content is integrated into the overall programme. Confirm it is a standing topic with regular updates. Check for role-specific content. Verify that verification procedures for high-risk requests (e.g. payment authorisation, credential changes) are documented and trained.	AI threat awareness is a quarterly training module updated with recent examples. Finance staff get extra training on deepfake payment fraud with mandatory callback verification; executives receive a voice-cloning impersonation briefing.
4	Quantitatively Controlled	Training effectiveness is measured: staff ability to identify AI-generated content, verification compliance rates for high-risk requests, and correlation with real AI-related incidents. Metrics drive content updates and reinforcement.	Request assessment results showing staff ability to identify AI-generated content. Check verification procedure compliance rates (e.g. callback completion for high-value requests). Ask for at least one example where a metric led to a training adjustment.	Quarterly assessments test staff on identifying AI-generated content; correct identification rose from 55% to 78% *(illustrative)*. Callback compliance sits at 94%. Lower scores among customer-facing staff prompted a targeted module.
5	Optimising	Training evolves in near real-time as AI attack capabilities advance. The organisation uses AI-generated content, such as crafted phishing and synthetic voice, in its own simulations, informed by threat intelligence and detection capability.	Ask how AI threat training was last updated and what triggered the change. Look for evidence that the organisation uses AI-generated content in its own simulations. Check for integration with threat intelligence on AI attack evolution. Ask whether the organisation contributes to industry knowledge sharing on this topic.	Security uses an AI tool to craft realistic phishing for simulations. When a new deepfake technique is shown at a conference, an updated module ships within three weeks, informed by a sector intelligence-sharing group.

Adversarial training and robustness

Service and Application Security

Scope: Dual

Range: L0–L5



This control covers the practices that harden AI models against adversarial manipulation: adversarial testing during model development, robustness evaluation against known attack techniques (evasion, poisoning, and prompt injection), input validation and sanitisation for model inputs, output filtering and guardrails, and ongoing production monitoring for adversarial inputs. Adversarial testing is integrated into the AI development lifecycle, with detection of AI-powered threats in the broader environment addressed separately. Without robustness testing, models can be manipulated through crafted inputs to produce incorrect outputs, bypass controls, or expose sensitive training data.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	No consideration is given to adversarial robustness. Models are developed and deployed without adversarial testing, AI-specific input validation, or output guardrails.	Ask whether AI models are tested for adversarial robustness. Check for input validation and output filtering on AI systems. If no adversarial considerations exist, this is L0.	A customer risk-scoring model is deployed without testing for crafted-input manipulation, and a knowledge-retrieval language model has no prompt injection protections. No one has considered how these models could be attacked.
1	Informal	Some awareness of adversarial AI risk exists in the data science team, but with no systematic practice. Individuals may test models informally, with no standard process required.	Ask the AI team whether they consider adversarial robustness. Check for any adversarial testing artefacts. If awareness is individual and practices are ad hoc, this is L1.	A data scientist informally tests the image classification model against perturbation techniques; it fails several tests but nothing is remediated. The customer support language model has no prompt injection protections.
2	Planned & Tracked	An AI security standard requires basic adversarial testing before deployment, input validation on AI model inputs, and output guardrails against harmful or out-of-scope content. Findings are documented and tracked.	Request the AI security standard. Check adversarial testing requirements. Verify input validation and output guardrails. Review adversarial testing findings and tracking.	The risk-scoring model is tested against input perturbation techniques before deployment. The language model has input filters blocking prompt injection and output guardrails against sensitive or out-of-scope content. Findings are tracked to resolution.
3	Standardised	Adversarial testing is integrated into the model development lifecycle for all models, covering evasion, poisoning, extraction, and prompt injection. Input validation and guardrails are standardised, red team exercises include AI scenarios, and robustness is re-evaluated on retraining.	Review the AI development lifecycle and confirm adversarial testing integration. Check coverage of attack types. Verify adversarial training techniques. Confirm standardised input/output controls. Ask about AI-specific red team exercises. Check re-evaluation on model updates.	All models undergo standardised adversarial testing: classification models against evasion attacks, language models against prompt injection and jailbreaking. The annual red team exercise includes AI-specific scenarios, and testing repeats whenever models are retrained.
4	Quantitatively Controlled	Robustness is measured against standardised benchmarks, tracking adversarial inputs detected in production, false positive rates, and improvement across model iterations. Metrics drive hardening priorities and risk assessments.	Request robustness metrics and benchmark scores. Ask for production adversarial detection rates. Check false positive rates for input/output controls. Ask for at least one example where a metric led to a model improvement.	A quarterly report shows the risk-scoring model's robustness score rising from 72% to 89% after adversarial training *(illustrative)*. Production blocks 97% of daily prompt injection attempts. A benchmark gap in the fraud model triggers targeted retraining.
5	Optimising	Robustness evolves with the threat landscape: intelligence on new techniques triggers rapid hardening, and automated adversarial testing runs continuously	Ask how adversarial robustness has adapted to new techniques. Look for evidence of continuous automated testing. Check for threat intelligence integration. Ask about research contribution.	When a new adversarial attack technique is published, the team reproduces it against the fraud detection model within a week and mitigates it. Continuous automated

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
		against production. The organisation shares benchmarks with peers and tests new architectures before deployment.		testing alerts on robustness drops, and anonymised benchmarks are shared with an industry working group.

AI model access controls

Service and Application Security

Scope: Dual

Range: L0-L5



This control covers the protection of AI assets as sensitive organisational resources: controlling who can access training data, model weights, configuration parameters, and inference endpoints; monitoring access and usage; protecting model artefacts from tampering; and managing the AI model supply chain (base models, pre-trained weights, and fine-tuning data). Training data is classified and protected according to the organisation's data classification framework, and access controls align with the broader identity and access management approach. Without AI-specific access controls, training data, model weights, and inference endpoints are freely accessible, exposing sensitive data and enabling model manipulation.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	No access controls specific to AI assets exist: training data, model weights, and inference endpoints are freely accessible, models are downloaded without provenance checks, and no model inventory exists.	Ask who can access AI training data, model weights, and inference endpoints. Check whether an AI model inventory exists. If no AI-specific access controls exist, this is L0.	Pre-trained models are downloaded without provenance checks. Customer training data sits in a shared file system open to all engineering staff, model weights are in an unprotected bucket, and endpoints have no authentication.
1	Informal	Some access controls exist for AI assets but are inconsistent and depend on individual team practice. Training data may be restricted while model weights and endpoints remain broadly accessible.	Ask how access to AI assets is managed. Check for consistency across training data, model weights, and inference endpoints. If access controls are partial and informal, this is L1.	Training data sits in restricted cloud storage, but model weights are in a general-purpose account open to all developers. Endpoints share one API key across teams, and models are downloaded without verification.
2	Planned & Tracked	An AI asset management process classifies training data, models, and endpoints, with access controls set by sensitivity. A model inventory tracks deployed models, and safe serialisation formats prevent deserialisation attacks.	Request the AI asset inventory. Check access controls for training data, model weights, and inference endpoints. Verify authentication on inference endpoints. Confirm use of safe serialisation formats.	A model inventory lists 12 deployed models with sources, storage locations, and access configurations. Training data access requires logged owner approval, endpoints use OAuth 2.0, and serialisation has moved to a safe format.
3	Standardised	Access controls are enforced consistently across AI workloads, with training data classified per the data framework and model provenance documented end to end. Training requires privileged authorisation, endpoints enforce rate limiting, and audit logging covers access, training, and inference.	Review AI access controls across all workloads. Verify training data classification. Check model provenance documentation. Confirm privileged access for training. Verify inference endpoint controls. Review AI audit logging. Check model signing and integrity verification.	Training data is classified confidential or restricted, and each model's provenance (base model, training data, initiator, hyperparameters) is documented. Training requires AI governance lead approval, endpoints enforce rate limiting, and audit logs capture every run and deployment.
4	Quantitatively Controlled	Access control effectiveness is measured: request volumes and approval rates, unauthorised access attempts, provenance completeness, and abuse detection rates. Metrics drive policy refinement and monitoring improvements.	Request metrics on access requests, unauthorised attempts, provenance completeness, and abuse detection. Ask for correlation with incidents. Check for at least one example where a metric led to an improvement.	A quarterly report shows all 12 models have complete provenance. A 50% rise in access requests prompts a streamlined approval workflow *(illustrative)*, and endpoint monitoring catches a training-data extraction attempt. A deployment gate now blocks models lacking provenance records.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
5	Optimising	Access controls adapt to new model types, deployment patterns, and threats. Automated supply chain verification checks provenance before use, continuous monitoring detects drift and poisoning indicators, and the organisation contributes to AI security standards.	Ask how AI access controls have adapted to new model types or threats. Look for evidence of automated supply chain verification. Check for continuous monitoring of model behaviour. Ask about contribution to AI security standards.	New foundation models get access controls and testing extended before production. Automated verification checks every base model's signature before registry entry, and monitoring catches a training-data drift issue. The organisation contributes to an OWASP LLM Top 10 working group.

Machine learning-based detection

Response and Recovery

Scope: Dual

Range: L0–L5



This control requires the application of machine learning models to security telemetry for pattern recognition that exceeds the capability of static rules: clustering similar alerts, identifying novel attack patterns, and predicting threat likelihood. It covers ML-based detection across services, applications, platform infrastructure, virtual infrastructure, and API integrations, improving detection accuracy and reducing false positives over time through measured model performance and retraining. The anomaly detection and behavioural analysis control provides the baseline modelling that complements ML detection; the security monitoring and detection systems control provides the data sources that feed ML models.

L	LEVEL	DESCRIPTION	ASSESSMENT METHOD	REAL-WORLD EXAMPLE
0	Non-existent	No machine learning is applied to security detection. All detection relies on manually authored rules and signatures.	Ask whether any ML-based detection capabilities are deployed. If all detection is rule-based or signature-based, this is L0.	The SOC relies entirely on SIEM correlation rules and antivirus signatures. Novel threats that don't match existing rules go undetected until damage is visible.
1	Informal	Some ML capabilities exist as built-in vendor features, such as endpoint malware detection, but the organisation has not configured, tuned, or evaluated them. ML is a default, not a managed capability.	Ask whether any security tools use ML-based detection. Check whether the organisation has configured or tuned these capabilities. If ML is present as a vendor default but not actively managed, this is L1.	The endpoint platform's ML detection engine is enabled by default, but the security team has not reviewed its configuration or evaluated its effectiveness.
2	Planned & Tracked	ML-based detection is deliberately deployed for specific use cases such as malware or phishing classification, configured for the environment, but coverage is limited to specific threat types or sources.	Request documentation of ML-based detection use cases. Check that models are configured for the organisation's environment. Verify that the team understands what the models detect and their limitations. Confirm coverage scope.	ML-based detection is configured in the endpoint platform for malware and the email gateway for phishing, tuned with custom exclusions and reviewed monthly. Network and cloud ML detection isn't deployed yet.
3	Standardised	ML-based detection spans all major domains (endpoint, network, cloud, identity, application) and integrates with the SIEM. Each model's detection scope, training data, and limitations are documented, and outputs correlate with other signals.	Verify ML detection coverage across security domains. Check integration with SIEM and incident management. Review model documentation including training data and limitations. Confirm that ML outputs are correlated with other signals.	ML-based detection spans endpoint, network, cloud, identity, and email, feeding the SIEM as enriched alerts with confidence scores. Documentation describes each model's scope, training data, and known blind spots.
4	Quantitatively Controlled	Model performance is measured across precision, recall, false-positive rates, latency, and drift, driving retraining and threshold adjustments and comparison against rule-based detection.	Request metrics on model precision, recall, false-positive rates, and drift indicators. Ask for at least one example where a metric led to model retraining or threshold adjustment. Check for comparison with rule-based detection effectiveness.	Quarterly reviews show the network model's precision dropping from 85% to 72% *(illustrative)* after infrastructure changes; retraining recovers it. ML-based phishing detection catches 30% more novel attempts than rule-based detection alone.
5	Optimising	ML detection continuously adapts to the threat landscape, with models retrained automatically or on schedule. The organisation develops custom models for its own threat patterns and contributes to sector-wide improvements.	Ask how ML models have adapted to recent threats. Look for automated retraining pipelines. Check for custom model development. Verify adversary simulation testing of ML detection.	The pipeline automatically retrains monthly using incident data and threat intelligence, and a custom model detects anomalous API usage. Red team exercises test evasion techniques, feeding results back into training and a sector data-sharing consortium.