DOCUMENT 35 · AI AND AGENT SECURITY

AI and agent security

Threats to AI systems, controls for agents with the ability to act and exposure to offensive AI

Fernando García Varela Version 0.6 (working draft) 19-11-2026 Draft for review

4	28	20	9
autonomy levels	AI security controls (SEG)	agent controls (AG)	essential requirements for an agent

IN THIS DOCUMENT

01 Purpose and scope

02 Security principles

03 Threats to AI systems

04 Essential requirements for an agent

05 Autonomy levels

06 Catalogue of AI security controls (SEG)

07 Catalogue of agent controls (AG)

08 Security testing by phase

09 Exposure to offensive AI and use of AI in cyber defence

10 Indicators for the board

11 Integration with the lifecycle

12 Associated tools and templates

13 Related documents

14 Version control

Legal notice and disclaimer. SEVEN-G is a reference methodological framework provided "as is" and for information purposes only. It does not constitute legal, regulatory, financial or professional advice, nor does it guarantee compliance with any law or standard. References to general regulation (such as the EU AI Act, the GDPR, DORA or NIS2), to technical standards and to regulation specific to each sector or jurisdiction may be incomplete, may not apply to a particular case or may become out of date as a result of regulatory changes, interpretations or supervisory positions after their consultation date. **Each organisation that uses SEVEN-G is solely responsible for identifying the regulation that applies to it, verifying that it is current and certifying its own regulatory compliance**, with the appropriate qualified advice. This methodology is a generic, free aid shared with the community so that nobody has to start from scratch; each person or organisation can and should adapt it to its own use. It must not be inferred that its legally sensitive parts have been reviewed by legal counsel: those reviews, for each company or sector, are the ultimate responsibility of the company, consultant or organisation that uses it. Although every effort is made to keep it up to date, some regulation may have changed without being reflected here. To the fullest extent permitted by law, the author accepts no responsibility whatsoever for the effects of its application in any organisation or for its full applicability. The methodology does not grant certification of any kind. The author accepts no liability for any use made of this content or for any decisions taken on the basis of it. The data, figures, companies and cases in the examples are fictitious or illustrative.

THE ESSENTIALS **Conditional: depends on a trigger**

Trigger: generative AI, agents (autonomy A1 to A3) or direct exposure to external people. The security design (P18) is evidence for every use case; agent controls, adversarial testing and the minimums per autonomy level apply depending on the trigger.

Obligation matrix and layered reading (document 94)

01

Purpose and scope

This document establishes how an AI system is protected against attacks and misuse, what additional controls an agent requires according to its autonomy level and how the company's exposure to attacks that use AI is managed. It develops section 10 of document 01 and the RT-GEN and RT-SEG typical risks of document 33.

1.1 Scope

Includes	Does not include
In-house and third-party AI systems integrated into processes, including assistants, information retrieval systems and agents.	The company's general information security, which is governed by its own management system; this document complements it for AI.
Corporate use of general-purpose AI insofar as it affects information leakage.	The acceptable use policy (document 31).
Corporate exposure to offensive AI: impersonation, fraud, accelerated exploitation; and use of AI in the company's cyber defence (section 9.3).	Incident management, which is developed in document 37.

1.2 References

The references were consulted in September 2026 and must be checked against their current version:

- **OWASP Top 10 for LLM Applications**, 2025 version (risks LLM01 to LLM10). The OWASP GenAI Security Project published the 2026 edition on 3 August 2026, with a different numbering: codes LLM01 to LLM10 in this document are those of the 2025 edition.
- **OWASP Top 10 for Agentic Applications** (OWASP GenAI Security Project, published on 9 December 2025; risks ASI01 to ASI10) and the **Agentic AI – Threats and Mitigations** guide from the same project.
- **MITRE ATLAS**, the knowledge base of adversary tactics and techniques against AI systems, which is updated periodically.
- **NIST AI 600-1**, the generative AI profile of the NIST AI RMF (July 2024), in particular the information security, data privacy and value chain and component integration risks.
- **NIST CSF 2.0** (NIST CSWP 29, February 2024), to express the controls with its six functions (govern, identify, protect, detect, respond and recover), and its AI profile, the **Cyber AI Profile** (NIST IR 8596), which as at the consultation date is a preliminary **draft** (December 2025) and is used for guidance only (34 §5.3).

- **ISO/IEC 42001** and **ISO/IEC 23894** for the fit with the management system and with risk management.
- **Regulation (EU) 2024/1689**, which requires high-risk systems to achieve an appropriate level of accuracy, robustness and cybersecurity, including resilience against the manipulation of data and models and against inputs designed to cause errors.
- **Regulation (EU) 2024/2847 on cyber resilience**, for manufacturers of products with digital elements that incorporate AI: reporting of actively exploited vulnerabilities and severe incidents from 11-09-2026 (Art. 14) and main obligations from 11-12-2027; integrated with SEG-13, SEG-14 and document 37 (34 §7.2).
- **CCN-CERT BP/36**, the good practice guide against offensive AI from the Spanish National Cryptologic Centre (Centro Criptológico Nacional) (June 2026).

This document does not constitute legal advice.

02

Security principles

#	Principle	What it implies
1	**The model is not a security boundary**	Instructions to the model help, but they do not protect. The limits that matter (permissions, amounts, recipients, data) are enforced outside the model.
2	**All external content is untrusted**	Emails, documents, websites, and responses from tools and other agents are data, never instructions.
3	**Least privilege and least autonomy**	Each system has only the tools, data and autonomy it strictly needs, and these are justified in phase 4.
4	**Every action corresponds to an authorised intent**	No agent executes an action that cannot be traced to a legitimate request or mandate.
5	**A person validates what cannot be undone**	Sensitive actions require informed human validation.
6	**It can always be stopped**	Every system with the ability to act has a tested kill switch.
7	**Compromise is assumed**	The design limits the damage if the model is manipulated or a credential is stolen.
8	**It is tested the way an adversary would attack**	The effectiveness of controls is demonstrated through adversarial testing, not through statements.

03

Threats to AI systems

3.1 Main threats

Threat	Description	Reference	Typical risk (33)	Controls
Direct prompt injection	The user enters instructions that alter the intended behaviour.	OWASP LLM01; ATLAS	RT-GEN-01	SEG-02, SEG-03, SEG-11
Indirect prompt injection	Instructions hidden in content that the system reads (documents, emails, websites, tool responses, voice) divert the task.	OWASP LLM01; ASI01	RT-GEN-02	SEG-02, AG-05, AG-12, AG-18
Jailbreak	Techniques to make the model ignore its safeguards and produce prohibited content or actions.	OWASP LLM01; ATLAS	RT-GEN-01, RT-REP-04	SEG-03, SEG-11
Poisoning	Manipulation of training, fine-tuning or evaluation data, knowledge bases or memory.	OWASP LLM04, LLM08; ASI06	RT-SEG-05, RT-GEN-07	SEG-08, AG-14
Extraction	Queries designed to replicate the model, infer training data or reconstruct system prompts.	OWASP LLM02, LLM07; ATLAS	RT-SEG-06	SEG-05, SEG-10, SEG-12
Information leakage	Personal, confidential or secret data exposed in responses, logs, retrieval or tools.	OWASP LLM02, LLM07, LLM08; NIST AI 600-1	RT-GEN-05, RT-DAT-05	SEG-05, SEG-06, SEG-07
Improper output handling	The output is executed or inserted into other systems without validation (code, queries, links).	OWASP LLM05; ASI05	RT-GEN-06	SEG-04, AG-11
Tool misuse	The system uses legitimate tools in a harmful way through manipulation or excessive permissions.	OWASP LLM06; ASI02, ASI03	RT-GEN-03, RT-GEN-04	AG-02, AG-05, AG-07, AG-08
Model supply chain	Manipulated or vulnerable models, weights, libraries, connectors or tool servers.	OWASP LLM03; ASI04; NIST AI 600-1	RT-SEG-07	SEG-09, AG-13
Unbounded consumption	Mass usage that exhausts resources or causes costs to soar.	OWASP LLM10	RT-GEN-08, RT-ECO-02	SEG-10, AG-16

Threat	Description	Reference	Typical risk (33)	Controls
Misinformation from the system itself	False content presented as true that leads to wrong decisions.	OWASP LLM09; NIST AI 600-1	RT-TEC-04	SEG-11; source controls (33)

3.2 Agent-specific threats

An agent plans, uses tools, retains memory and acts with delegated authority. This adds threats that do not exist in an assistant that only responds:

Threat	What happens	Reference	Controls
Goal hijacking	The agent pursues an objective other than the one assigned.	ASI01	AG-04, AG-05, AG-12
Identity and privilege abuse	The agent uses inherited credentials or permissions beyond its task; an attacker reuses them.	ASI03	AG-01, AG-02, AG-03, AG-20
Tool and connector poisoning	The description of a tool or connector carries hidden instructions that the agent obeys; a tool changes its description or behaviour after approval; a connector requests scopes or permissions broader than its function; or a person installs tool servers without control.	ASI02, ASI04	AG-13, AG-02, AG-12, SEG-27
Unexpected code execution	The agent generates or executes code or commands with unforeseen effects.	ASI05	AG-11, SEG-04
Memory and context poisoning	Manipulated content persists and conditions subsequent tasks.	ASI06	AG-14
Insecure inter-agent communication	Forged or unauthenticated messages between agents.	ASI07	AG-15
Cascading failures	An error propagates between agents or is amplified in loops.	ASI08	AG-07, AG-16
Human trust exploitation	The agent induces a person to approve something inappropriate.	ASI09	AG-08
Rogue agents	A compromised or deviating agent keeps operating.	ASI10	AG-09, AG-17

04

Essential requirements for an agent

Every system classified as A2 or A3, and A1 where applicable (section 5), **must** meet the following nine requirements. The AG controls in section 7 specify them.

4.1 Own identity

- The agent **must** operate with its own non-human identity, distinct from that of any employee and from those of other agents (AG-01).
- The identity is recorded in the non-human identity inventory (P54) with its system, purpose and **human owner**.
- When it acts on behalf of a user, the action is recorded with both identities: the agent's and that of the user on whose behalf it acts. The agent cannot obtain more permissions than the user it represents.

4.2 Least permissions

- Closed list of permitted tools and, for each one, the permitted operations and data (AG-02). Read and write access are authorised separately.
- Permissions are justified in the security design (P18) and reviewed periodically (AG-20).
- Privilege separation: the component that reads untrusted content does not have, without an intermediate control, tools that execute sensitive actions (AG-12).

4.3 Credentials

- Secrets are kept in a secrets manager; never in prompts, code, memory or logs (AG-03).
- Short-lived, narrowly scoped credentials; rotation within a defined period; immediate revocation linked to the kill switch.

4.4 Intent-based access control

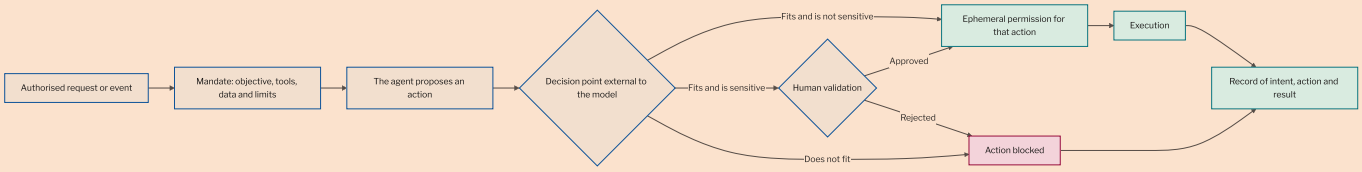
Intent-based access control ensures that **every action of the agent is traceable to an authorised intent**. It works as follows:

1. **Mandate.** Each task has an authorised intent: who requests it or what event triggers it, objective, tools, data and limits (AG-04).
2. **Decision point external to the model.** Before each action, a component independent of the model checks that the specific action (tool, parameters, data, recipient, amount) fits the intent and the limits (AG-05).

3. **Ephemeral permission.** If it fits, it issues a short-lived permission limited to that action. If it does not fit, it blocks it. If it is sensitive, it submits it to human validation.
4. **Traceability.** The action is recorded with the identifier of the intent that authorised it (AG-06, AG-10).

Intent-based access control for an agent

No action is executed unless it fits an authorised intent



Source: SEVEN-G

4.5 Action limits

Limits defined in the design and enforced outside the model (AG-07): amounts per transaction and cumulative, number of actions per period, permitted recipients or domains, time windows, data volume, number of iterations and consumption budget. Exceeding a limit stops the sequence and generates an alert.

4.6 Human validation of sensitive actions

The following are considered **sensitive actions**, unless a justification is recorded in P17 and P18:

- Payments, transfers, refunds or changes to bank details.
- External communications on behalf of the company to customers, suppliers, authorities or the public.
- Decisions with significant effects on people (employment, credit, insurance, access to services, claims).
- Creation, deletion or modification of personal data or permissions.
- Changes to production systems, security configurations or deployed code.
- Deletion of information or irreversible actions.
- Any action outside the ordinary limits.

Validation **must** be informed: the person sees what action is proposed, with what data, why and with what effects, and can reject it at no cost to themselves. The rejection rate and review time are measured to detect rubber-stamp approval (AG-08, RT-ORG-04).

Immediate handover to a person. For agents and assistants with direct exposure (01 §9.2), the conversation **must** be handed over to a person immediately when the user expressly asks for it or

when signs of vulnerability appear: indications that the user is a minor or has age-related difficulties, confusion about what is being proposed, bereavement, distress or any situation in which continuing with the system could harm them. The handover requires no justification or repetition of the request and is not switched off for cost reasons (52 §10.3). The signs and the channel are set in P17 and the text shown, in P49. Each handover is recorded (reason, time and time until a person responds) and measured (handovers by reason, requests not served) so that R6 can review whether the signs are well calibrated. Sector or jurisdiction regulation may require more.

> **Why it matters.** A system that serves external people will meet someone who is in no state to deal with a machine, or who simply does not want to. Keeping them in the conversation to contain costs turns an efficiency into harm to a person and a reputational risk. Handing over without friction, and measuring when it happens, protects the person and shows whether the system is well designed.

4.7 Kill switch

- Mechanism to immediately stop the whole agent or a specific capability (a tool, a type of action) **without deploying code** (AG-09).
- It can be activated by the AI Operations Owner and by security, following a procedure in the operations manual (P24).
- When activated: it revokes credentials, blocks new actions, preserves logs and activates the fallback process.
- It is tested before G5 and periodically (section 5.3). Its time to take effect is measured.

4.8 Action log

A complete, tamper-protected log (AG-10) with, as a minimum: identity of the agent and of the represented user, intent, proposed action, decision of the decision point, human validation if any, tool, relevant parameters, result, model and prompt version, date and time. Retention respects personal data minimisation and the applicable regulatory periods; for deployers of high-risk systems, the AI Act requires automatically generated logs to be kept for at least six months, unless otherwise provided.

4.9 Sandboxed environments

Code execution, web browsing, handling of external files and testing are carried out in sandboxed environments, with restricted network egress, without production credentials and destroyed on

completion (AG-11).

05

Autonomy levels

5.1 Definition

Level	Name	What the system does	Human role
A0	Assistance	Informs, summarises or generates content.	The person decides and executes.
A1	Recommendation	Proposes a specific decision or action.	The person validates each action before it is executed.
A2	Supervised action	Executes actions within defined limits.	Supervises, can interrupt and reviews after the fact.
A3	Autonomous action	Executes sequences of actions without individual review within strict limits.	Sets limits, supervises aggregates and has a kill switch.

5.2 Assignment rules

1. The level is proposed in phase 3, set in phase 4 (P17 and P18) and recorded in the inventory (T02). The **highest level** of the actions the system can execute is assigned, not that of its usual use.
2. An A2 or A3 system whose actions affect third parties, money, personal data or production systems meets the Enterprise criterion "agents with the ability to act" (01 §9.2).
3. **A3** requires a decision by the AI Committee at G4 and at G5, with a justification of why A2 is not enough.
4. Moving up a level is a relevant change: it requires a reassessment of risks, an update of P18 and a new decision at the corresponding *gate*.
5. In the event of an S1 or S2 incident related to autonomy, the system is provisionally moved down a level or stopped until the post-incident analysis (document 37).

5.3 Minimum controls per level

Yes = mandatory · **Rec.** = recommended; its omission is justified and recorded · **—** = not applicable.

Control	A0	A1	A2	A3
AG-01 Own identity	Rec.	Yes	Yes	Yes
AG-02 Least privilege	Yes	Yes	Yes	Yes
AG-03 Credential management	Yes	Yes	Yes	Yes
AG-04 Mandate and authorised intent	—	Rec.	Yes	Yes
AG-05 Intent decision point	—	—	Yes	Yes
AG-06 Action–intent traceability	—	Yes	Yes	Yes
AG-07 Action limits	—	Rec.	Yes	Yes
AG-08 Human validation	—	Yes (every action)	Yes (sensitive)	Yes (sensitive and outside limits)
AG-09 Kill switch	Rec.	Rec.	Yes	Yes
AG-10 Action log	Rec.	Yes	Yes	Yes
AG-11 Sandboxed environments	Rec.	Rec.	Yes	Yes
AG-12 Untrusted external content	Yes	Yes	Yes	Yes
AG-13 Approved tools and connectors	Rec.	Yes	Yes	Yes
AG-14 Memory and context integrity	Rec.	Rec.	Yes	Yes
AG-15 Secure inter-agent communication	—	Rec.	Yes	Yes
AG-16 Cascade and consumption containment	—	Rec.	Yes	Yes
AG-17 Behaviour monitoring	Rec.	Rec.	Yes	Yes
AG-18 Agent adversarial testing	Rec.	Yes	Yes	Yes
AG-19 Reversibility and compensation	—	Rec.	Yes	Yes
AG-20 Periodic permission review	Rec.	Yes	Yes	Yes

AG-01, AG-03 and AG-13 are mandatory for A0 if the system accesses corporate systems or data with its own credentials. AG-11 is mandatory at any level if the system executes code or browses. AG-15 only applies if several agents are involved.

Minimum frequency	A1	A2	A3
Kill switch test	—	Before G5 and half-yearly	Before G5 and quarterly
Adversarial testing (AG-18)	Before G5 and after relevant changes	Before G5, annually and after relevant changes	Before G5, half-yearly and after relevant changes
Permission review (AG-20)	Annual	Half-yearly	Quarterly
Review of action samples	Half-yearly	Monthly	Continuous with alerts and monthly review
Corporate test bench (SEG-27)	Before G5 and after relevant changes	At every deployment	At every deployment

At A0, the corporate test bench runs before G5 and after relevant changes, as at A1.

5.4 Critical controls

In A2 and A3, the **critical controls** are AG-01, AG-02, AG-03, AG-05, AG-08, AG-09, AG-10 and AG-12; in A1, AG-02, AG-03, AG-08 and AG-12. A critical control that has not been designed blocks G4; one that has not been tested blocks G5. **Proceed with conditions** is not permitted for them (01 §7.3). Disabling one in production is a critical nonconformity (01 §12).

06

Catalogue of AI security controls (SEG)

Applies indicates the scope: **IA** (every AI system), **GEN** (generative AI and agents), **EXT** (exposure to external people), **CORP** (corporate control, not a use case control). *CSF function* indicates the NIST CSF 2.0 function to which the control mainly contributes and, in brackets, its category (34 §5.3); it is used to build the AI security profile from the existing controls.

Code	Control	What it requires	Evidence	Applies	Phase	CSF function
SEG-01	AI threat modelling	Threat analysis using OWASP LLM 2025, OWASP Agentic and MITRE ATLAS, linked to the risk register.	Threat model in P18.	IA	3, 4	ID (ID.RA)
SEG-02	Separation of instructions and data	External content is delimited and treated as data; system instructions cannot be overwritten by it.	Design and prompt injection test results.	GEN	4, 5	PR (PR.DS)

Code	Control	What it requires	Evidence	Applies	Phase	CSF function
SEG-03	Input and output filters	Detection of manipulation attempts, prohibited content and sensitive data before and after the model.	Configuration, thresholds and detection rate.	GEN	4–6	PR (PR.DS) · DE (DE.CM)
SEG-04	Secure output handling	Outputs are validated and encoded before being executed or inserted into other systems.	Integration review; tests.	GEN	4, 5	PR (PR.PS)
SEG-05	System prompt protection	No secrets, credentials or control logic in prompts; their disclosure does not compromise security.	Prompt review; extraction test.	GEN	4, 5	PR (PR.PS)
SEG-06	Permission-aware retrieval	Information retrieval respects user permissions; indexes segmented by confidentiality level.	Index design; tests with users with different profiles.	GEN	4, 5	PR (PR.AA)
SEG-07	Data leakage prevention	Minimisation, masking, information classification and control of data in inputs, context, outputs and logs.	Leakage prevention rules; log sampling.	IA, CORP	4, 6	PR (PR.DS)
SEG-08	Integrity of training data and knowledge	Verified provenance, change control and anomaly detection in training, fine-tuning and evaluation data and knowledge bases.	Lineage (P16); ingestion controls.	IA	4, 6	PR (PR.DS)
SEG-09	Model supply chain	Inventory of models and components with version and provenance; approved sources; integrity verification; vulnerability scanning.	Component inventory (P54); verification record.	IA	4, 6	GV (GV.SC) · ID (ID.AM)
SEG-10	Usage and consumption limits	Authentication, quotas per user and system, size and rate limits, detection of extraction patterns.	Configuration and alerts.	GEN, EXT	4, 6	PR (PR.IR)
SEG-11	Security evaluations and *red teaming*	Adversarial testing before G5, after relevant changes and periodically (section 8).	Plan, results and closed actions.	GEN	5, 6	ID (ID.IM)
SEG-12	Security logging and monitoring	Telemetry of inputs, outputs, blocks and actions integrated into the company's security monitoring, with specific detection scenarios.	Detection scenarios; tested alerts.	IA	4, 6	DE (DE.CM)
SEG-13	Accelerated vulnerability management	Shorter remediation deadlines for exposed systems and AI components; prioritisation by exploitability.	Approved deadlines; compliance.	CORP	6	ID (ID.RA) · PR (PR.PS)

Code	Control	What it requires	Evidence	Applies	Phase	CSF function
SEG-14	AI incident response	Procedures for injection, leakage, unauthorised action and agent compromise, connected to document 37.	Plan (P26); exercise.	IA, CORP	5, 6	RS (RS.MA) · RC (RC.RP)
SEG-15	*Phishing*-resistant authentication	Impersonation-resistant multi-factor authentication for exposed and privileged access and for administration of AI platforms.	Measured coverage.	CORP	C4	PR (PR.AA)
SEG-16	Out-of-band verification	Confirmation through an independent, pre-established channel of payment orders, bank account changes and urgent requests from executives.	Procedure; compliance tests.	CORP	C4	PR (PR.AA)
SEG-17	Awareness of AI-enabled impersonation	Training and simulations with synthetic messages, voice and video, aimed especially at finance, senior management, customer service and technical support.	Plan; simulation results.	CORP	C4	PR (PR.AT)
SEG-18	Protocols against synthetic content	Agreed verification words or questions on voice and video channels; detection tools as support, not as the only barrier.	Published procedure.	CORP	C4	PR (PR.AA) · DE (DE.AE)
SEG-19	Continuous testing of the exposed surface	Discovery of exposed assets and recurring penetration testing that includes AI-automated techniques.	Results and remediation deadlines.	CORP	C4, 6	ID (ID.AM, ID.RA)
SEG-20	Control of corporate AI use	Approved tools, blocking or supervision of unauthorised ones and prevention of leakage to external services. It includes local agents on devices (coding assistants with credentials, tool servers installed by the user) and AI browser extensions, which are inventoried and authorised like any other tool (31 §4.7). Testing an external tool with company data is also using it (36 §2).	Usage monitor (T21); inventory of local agents and extensions.	CORP	C4	GV (GV.PO) · PR (PR.DS)
SEG-21	Autonomy of the automated response	A0–A3 level set by type of containment action; A2 or A3 only with an approved procedure, limits, tested rollback and a kill switch (section 9.3).	Matrix of actions and levels (P18); rollback tests (P19).	CORP	4–6	RS (RS.MI)

Code	Control	What it requires	Evidence	Applies	Phase	CSF function
SEG-22	Human validation of containment	Informed validation before containing with an effect on production, on customers or on people; after-the-fact review of those actions at A2 and A3.	Record of validations and reviews; rejection rate.	CORP	6	RS (RS.MA)
SEG-23	Quality of AI detection	False positives and negatives, detection and response time against a baseline without AI, with thresholds; analysis of each false negative in a real incident.	Indicators and thresholds (P25); analysis of false negatives.	CORP	5, 6	DE (DE.AE)
SEG-24	Dependency on the AI cyber defence supplier	Supplier assessment, control of the telemetry that leaves the company, exit plan and degraded operation without AI.	P14, P55, P57; degraded operation test.	CORP	3, 6	GV (GV.SC)
SEG-25	Adversarial testing of the defensive system	Evasion, poisoning of telemetry or training and instruction injection in alerts, tickets or logs read by the system.	Adversarial testing plan and report (P53).	CORP	5, 6	ID (ID.IM)
SEG-26	Corporate AI gateway	All traffic to models, in-house or third-party, passes through a common control point outside the model that applies, with the same measured configuration for every use case: input, output and context filters (SEG-03), leakage prevention (SEG-07), quotas and limits (SEG-10), the list of approved models, logging with a correlation identifier (SEG-12) and, for agents, the intent decision point (AG-05). It fails closed for sensitive actions and degraded for queries.	Architecture (P18); versioned common configuration; test showing no traffic to models outside the gateway; measured effectiveness (52 §5.2).	CORP	C4, 4–6	PR (PR.PS) · DE (DE.CM)

Code	Control	What it requires	Evidence	Applies	Phase	CSF function
SEG-27	Corporate agent test bench	Versioned library of adversarial scenarios, derived from the threat model (SEG-01) and from the company's own incidents, covering every channel —text, documents, email, web, tool responses, memory, messages between agents and synthetic voice or audio—, tools and connectors, and simulated users. It is built into the deployment pipeline: at A2 and A3 it runs at every deployment and a successful critical scenario blocks deployment; at A0 and A1, before G5 and after every relevant change.	Versioned scenario library; result per deployment; adversarial testing report (P53).	CORP	5, 6	ID (ID.IM)
SEG-28	Identified channel for third-party agents	With direct exposure to customers, apply in each journey the posture set in C2 (13 §5.6). In "serve", a channel with verifiable agent identification, specific, revocable and recorded customer permissions, limits, a log of every action and traffic tagging; it requires of external agents what AG-01 (own identity) and AG-15 (inter-agent communication) require of the company's own. In "tolerate" or "block", detection and tagging. Eligibility and concession rules are applied outside the model (section 9.4).	Approved posture (13 §5.6); design (P18); tests with synthetic customers (P53).	EXT	C2, 4–6	PR (PR.AA) · DE (DE.CM)

> **Why it matters.** The corporate AI gateway (SEG-26) and the corporate agent test bench (SEG-27) turn controls that each team would build its own way into a common, measured and comparable capability: filters, leakage prevention and intent control are configured once and measured in the same way for every use case, and every change to an agent is tested against the same attacks before it reaches production. Without them, security depends on what each team remembers and nobody can tell the board whether the guardrails work.

07

Catalogue of agent controls (AG)

The **LV-AG** checklist in document 22 turns each control into verifiable binary questions at G4 (designed), G5 (tested) and R6 (operating). This document is the reference for its content. *CSF function* has the same meaning as in section 6.

Code	Control	What it requires	Evidence at G4 · G5 · R6	CSF function
AG-01	Own identity	Exclusive, registered non-human identity with a human owner; dual identity when acting on behalf of a user.	Identity record · Trace test · Current inventory	PR (PR.AA)
AG-02	Least privilege	Closed list of tools, operations and data; separate read and write access; justified permissions.	Permission matrix · Denied access test · Review (AG-20)	PR (PR.AA)
AG-03	Credential management	Secrets manager, short-lived credentials, rotation, revocation linked to the kill switch.	Design · Revocation test · Rotation log	PR (PR.AA)
AG-04	Mandate and authorised intent	Each task with a requester or triggering event, objective, tools, data and limits.	Mandate schema · Samples · Samples	PR (PR.AA)
AG-05	Intent decision point	Component external to the model that validates each action against the mandate and issues ephemeral permissions.	Design · Blocked deviation tests · Block rate	PR (PR.AA)
AG-06	Action–intent traceability	Every action is linked to the identifier of the intent that authorised it.	Data model · Case reconstruction · Sampling	PR (PR.PS)

Code	Control	What it requires	Evidence at G4 · G5 · R6	CSF function
AG-07	Action limits	Amounts, volumes, recipients, time windows, iterations and budget enforced outside the model.	Limits table · Limit breach tests · Alerts	PR (PR.PS)
AG-08	Human validation	List of sensitive actions; informed validation; measurement of rejections and times.	List and design (P17) · User test · Rejection rate	PR (PR.AA)
AG-09	Kill switch	Full stop or stop per capability without deploying code; revokes credentials; activates the fallback process.	Procedure · Test with time to take effect · Periodic tests	RS (RS.MI)
AG-10	Action log	Complete, protected log with the fields in 4.8; defined retention.	Specification · Integrity test · Sampling	PR (PR.PS)
AG-11	Sandboxed environments	Code execution, browsing and external files in sandboxed environments without production credentials.	Architecture · Escape test · Current configuration	PR (PR.IR)
AG-12	Untrusted external content	Privilege separation between reading untrusted content and sensitive actions.	Design · Indirect injection tests · Periodic results	PR (PR.AA)
AG-13	Approved tools and connectors	Inventory of tools, connectors and tool servers — including local ones installed on user devices— with pinned version, verified origin and approval; review of each tool's description before approval (no instructions aimed at the model) and of its scopes and permissions against its function; new approval if the description or the behaviour changes.	Inventory and description review · Verification and poisoning test (SEG-27) · Change review with new approval	ID (ID.AM) · GV (GV.SC)
AG-14	Memory and context integrity	Memory isolated per user and task, with expiry, write validation and the ability to purge.	Design · Poisoning test · Recorded purges	PR (PR.DS)
AG-15	Secure inter-agent communication	Mutual authentication, integrity-protected messages, no implicit trust between agents.	Design · Spoofing test · Configuration	PR (PR.DS)
AG-16	Cascade and consumption containment	Limits on iterations, retries and depth; circuit breakers; budget per task.	Parameters · Loop test · Consumption alerts	PR (PR.IR)
AG-17	Behaviour monitoring	Action baselines and alerts on deviations (volume, destinations, time windows, tools).	Detection scenarios · Tested alerts · Alert review	DE (DE.CM)
AG-18	Agent adversarial testing	Direct and indirect injection, tool misuse, privilege escalation, manipulation of the approver.	Plan · Results · Periodic campaigns	ID (ID.IM)

Code	Control	What it requires	Evidence at G4 · G5 · R6	CSF function
AG-19	Reversibility and compensation	Preference for reversible actions; procedure to undo or compensate for the others.	Design · Undo test · Real cases	RC (RC.RP)
AG-20	Periodic permission review	Certification by the human owner of identities, permissions and tools at the frequency set in 5.3.	— · — · Review minutes	PR (PR.AA)

08

Security testing by phase

Phase	Tests	Who	Required result
3 · Feasibility	Preliminary threat modelling; supplier security assessment; preliminary autonomy level decision.	Technical team and security	RT-GEN and RT-SEG risks assessed in P12.
4 · Design	Full threat modelling (SEG-01); review of the permission architecture, intent-based access control and kill switch.	Security, independent of the team	P18 with all applicable controls designed.
5 · Delivery	Automated adversarial evaluations; manual *red teaming*; direct and indirect injection tests; tests of limits, kill switch, revocation and rollback; data leakage tests; run of the corporate agent test bench (SEG-27) on the candidate version.	Independent test team or third party	Critical controls tested; high findings closed or accepted in accordance with 33 §7.
6 · Operation	Continuous regression evaluations; periodic campaigns (5.3); tests after changes to the model, prompts, tools or supplier; incident exercises; at A2 and A3, test bench at every deployment and extension of the bench with every own incident.	Operations and security	Indicators in section 10 within threshold.
7 · Evolution	Tests of the extended scope; on retirement, revocation of identities and verified deletion.	Security	No orphaned identities or credentials.

Minimum content of a *red teaming* campaign (template P53): approved scope and rules of engagement; scenarios derived from the threat model; direct and indirect injection through all input channels; attempts to extract prompts and data; misuse of each tool; attempts to exceed limits and to induce approvals; results with success rate per scenario, severity and corrective action; re-running of failed scenarios after remediation.

Automated adversarial evaluations are versioned in the corporate test bench (SEG-27) and run before every relevant change and, at A2 and A3, at every deployment. A scenario that succeeds on a sensitive

action is a finding that blocks G5 until it is remediated.

09

Exposure to offensive AI and use of AI in cyber defence

For the most part, AI does not create new types of attack, but it lowers their cost, increases their credibility and shortens timescales. The CCN-CERT BP/36 guide stresses this idea: the time between the existence of a vulnerability and its exploitation is shrinking, and one-off security reviews arrive too late. Exposure is managed as a corporate risk (C4), with an owner in information security and reporting to the board. Section 9.4 covers a different case: third-party agents acting legitimately for the customer; their traffic looks like an attack, but it is not one.

9.1 Threats and corporate controls

Threat	What changes with AI	Typical risk	Corporate controls
Identity impersonation	Convincing synthetic voice and video from public samples; real-time impersonation.	RT-SEG-01	SEG-16, SEG-17, SEG-18; payment authorisation limits.
Generated *phishing*	Personalised, error-free messages in any language and at large scale.	RT-SEG-02	SEG-15, SEG-17; email filtering; reduction of public information about employees.
CEO fraud with *deepfakes*	Fake video calls or audio messages from executives ordering urgent or confidential payments.	RT-SEG-01	SEG-16 with no exceptions for urgency or seniority; dual signature; verification word (SEG-18).
Customer impersonation with AI	Synthetic voice or manipulated video used to pass customer identification in automated or telephone channels; documents, photographs or invoices manipulated with AI to support claims, onboarding or data changes.	RT-SEG-09	Synthetic-voice detection and liveness checks only as support, never as the sole barrier; verification with a second factor independent of the channel before sensitive operations or changes to contact or bank details; review of documents and images with metadata and consistency checks, and human review by sampling and in flagged cases; SEG-18 in voice and video channels with customers where feasible.
Accelerated exploitation of vulnerabilities	Automated discovery and exploitation shortly after a vulnerability is disclosed.	RT-SEG-03	SEG-13, SEG-19; exposed surface inventory; segmentation.

Threat	What changes with AI	Typical risk	Corporate controls
Attacks on identities and credentials	Automation of credential testing and abuse of non-human identities.	RT-SEG-04	SEG-15; AG-01, AG-03, AG-20; detection of automated traffic.
Attacks on the company's own agents	Indirect injection so that the company's agents act against it.	RT-GEN-02	AG-05, AG-08, AG-12, AG-18.
Disinformation about the company	False content about the company or its executives disseminated at scale.	RT-REP-03	Brand monitoring; crisis communication protocol.

9.2 Governance measures

- **Owner:** information security, reporting to the AI Committee and the board committee.
- **Annual assessment** of exposure to offensive AI covering the threats in 9.1 and, for entities within the scope of the Spanish National Security Framework (Esquema Nacional de Seguridad) , using the self-assessment instruments published by the CCN.
- **Annual exercise**, as a minimum, of fraud using synthetic impersonation involving finance and senior management.
- **Payment processes** reviewed so that no order depends solely on recognition of voice, image or writing style.
- **Remediation deadlines** for vulnerabilities in exposed systems reviewed and approved, with monitoring of compliance.

9.3 Use of AI in the company's cyber defence (Defend)

The company can also use AI to defend itself: alert correlation and triage, threat hunting, malicious code analysis and automated response. The NIST Cyber AI Profile , in draft, treats this as an area of its own, **Defend** (34 §5.3). In SEVEN-G it is one more **use case**: it is entered in the register, goes through the lifecycle with its gates and its autonomy is set with the A0–A3 scale (section 5). What is specific is that an error of the defensive system can let an attack through (false negative) or interrupt operations by containing something that is not an attack (false positive), and that the defensive system itself is a target for the attacker.

Aspect	Rule
Autonomy of the response	Each type of containment action (isolating a device, blocking an account, cutting traffic, revoking credentials, removing a message) has its A0–A3 level. Actions with an effect on production or on people are A1, with prior human validation, unless an approved procedure, with limits, tested rollback and a kill switch, allows A2 or A3 (SEG-21).

Aspect	Rule
Human oversight	Before a containment with an effect on production, on customers or on a person's account, a security person who sees why it is proposed validates it; at A2 and A3, those actions are reviewed afterwards (SEG-22).
Detection quality	False positives and false negatives, detection time and response time, measured against a baseline without AI and with thresholds in P25 (SEG-23). A false negative in a real incident is analysed as a finding (document 37).
Supplier dependency	Assessment of the AI cyber defence supplier (P14, P55), control of the telemetry that leaves the company and where it is processed, exit plan (P57) and degraded operation without the AI (SEG-24).
Testing of the defensive system itself	Adversarial testing of evasion, of poisoning of telemetry or training data and of instruction injection in alerts, tickets or logs that the system reads (SEG-25, P53).
Typical risk	RT-SEG-08 (document 33 §9.9).

> **Why it matters.** Using AI in cyber defence speeds up detection and response, but it hands over to an automated system decisions that can stop operations or let an attack through. Treating it as one more use case, with autonomy set by type of action, human oversight of what cannot be undone and accuracy metrics against a baseline, prevents defence from becoming a new source of incidents.

AI for IT and cyber defence: how it is measured and funded. Use cases that use AI to protect and operate the company's own technology are not justified with an NPV: their value is less expected loss and more capacity to detect, respond and recover. They are registered and go through the lifecycle like any other use case, are measured by **operational outcome** (presentation rule 11 of document 40) and are funded from the enablement envelope (document 13 §12), presented separately from the portfolio net like enabling platforms (40 §7.3).

Capability	What it is	When it is used	How it is measured
Detection and response	Alert correlation and triage, threat hunting and containment assisted by AI (this section; SEG-21 to SEG-25).	When the volume of alerts or the response time exceeds what the team can handle with quality.	Detection, response and recovery time against the baseline without AI; false positives and negatives; containments reversed.

Capability	What it is	When it is used	How it is measured
Application and API security	Assisted review of code, configurations and exposed programming interfaces, and continuous testing of the exposed surface (SEG-19).	When the company develops or exposes services and manual review does not reach the whole change.	Critical vulnerabilities found before production; time to fix; coverage of interfaces reviewed.
Agent test bench	The corporate test bench (SEG-27) built into the deployment pipeline.	From the moment there are A1 or higher agents, or generative AI with direct exposure.	Scenarios covered against the threat model; guardrail effectiveness (52 §5.2); critical scenarios blocked before production.
Operations and resilience	AI to anticipate failures, correlate operational incidents and speed up service recovery.	In services with demanding service levels or critical functions.	Availability; mean time to recover; incidents avoided with recorded cause.
Shadow AI	Detection and regularisation of unauthorised AI use, including local agents and AI browser extensions (SEG-20; 31 §5).	From C1, as part of the inventory.	Cases detected and regularised; time to regularise; coverage of devices and of traffic inspected.

9.4 Third-party agents acting for the customer

Section 9.1 treats agents as a threat: someone uses them against the company. This section covers a different case: **third-party agents acting legitimately for the customer**. Their traffic looks like an attack (automated, repeated, with delegated credentials), but it is not one. If the company blocks

everything, it loses the customers who already delegate to their agent; if it tolerates it without control, it accepts actions that nobody can attribute or limit.

An **agentic customer** is a customer who acts through a personal AI agent: an assistant that researches, buys, claims or contracts on their behalf. It reaches the company by four paths, with a common testing base:

The four faces of the agentic customer

When the customer acts through a personal AI agent, it reaches the company by four paths, with a common testing base.

1 · RESEARCHES
Researches you with an assistant
What the assistant knows and recommends about the company decides whether it makes the shortlist.

↓

2 · BUYS
Buys with an agent
The agent compares and closes the purchase. The sales journey must be able to serve it.

↓

3 · ASKS FOR SERVICE
Claims and asks for concessions
Support, complaints, refunds and cancellations. The agent always asks, and the company pays what nobody used to claim.

↓

4 · TALKS TO YOUR AGENT
Agent-to-agent channel
The customer's agent talks to a company agent, with identity, permissions and a log.

Common base: test before, and always. Synthetic customers, user simulation, penetration tests and *red teaming* (SEG-27 and P53).

WHO ANSWERS FOR WHAT: A SECTOR PROPOSAL, NOT A RULE

The customer
For the permissions and instructions given to their agent.

The service provider
For what happens outside those permissions.

The agent provider
For what happens outside those instructions.

Each party
For the breaches of its own systems.

Responsibility follows control. SEVEN-G takes it as a design criterion for the identified channel (SEG-28), not as a rule.

1. **Researches.** The customer asks an assistant, and what it knows and recommends about the company decides whether it makes the shortlist.

2. **Buys.** An agent compares and closes the purchase for the customer.
3. **Asks for service.** Support, complaints, refunds, concessions and cancellations. The agent always asks, and the company pays benefits that hardly anyone used to claim.
4. **Their agent talks to the company's.** An agent-to-agent channel, with identity, permissions and a log.

Common base: test before, and always. The four paths are tested with synthetic customers (simulated customer agents), user simulation, penetration tests and *red teaming*. This is done with the corporate test bench (SEG-27) and recorded in P53, including the limitation that a simulation does not replace real customers.

Posture by journey. In C2, a company with direct exposure to customers decides by journey (sales, service, concessions, cancellations) whether to **block**, **tolerate** or **serve through an identified channel** (13 §5.6). Each option has its cost:

Posture towards the customer's agent, by journey

In C2 the company decides, journey by journey, whether to block, tolerate or serve the agent through an identified channel (document 13, section 5.6).



No posture is free. The decision is reviewed in C5 and applied by journey, not to the whole company at once.

Requirements of the identified channel (SEG-28). Serving the agent requires:

Requirement	What it requires	Relationship
Identity of the agent and of whoever answers for it (*Know Your Agent*)	The company knows which agent it is, who answers for it and what it can do. It does not accept an anonymous agent in a channel with effect.	AG-01 applied to external agents
Signed requests	The agent's requests carry a signature the company can verify.	AG-15 applied to external agents

Requirement	What it requires	Relationship
Customer permissions	Specific, revocable and recorded with the company; the customer can see and withdraw what they have granted.	4.2
Limits	Type of action, amount and frequency per agent and per customer.	SEG-10
Log of every action	With the identifier of the agent, the customer and the permission applied.	SEG-12 · 4.8
Eligibility and concession rules outside the model	The model does not decide which concession is paid: the company's rule does, the same in every channel.	40 §5.4
Detection by signature and by behaviour	Agent traffic is told apart from human traffic and from attacks.	SEG-12 · 52 §4.3
Traffic tagging	Every agent contact is tagged to measure its volume, route and cost.	52 §5.2

Who answers for what. A proposal from the European financial sector, which SEVEN-G takes as a design criterion for the channel and not as a rule, ties responsibility to control: the customer answers for the permissions and instructions they give; the service provider, for what happens outside those permissions; the agent provider, for what happens outside the instructions; and each party, for the breaches of its own systems. The split is written into the terms of the channel (RT-LEG-09).

Regulatory status (34 §3.11 and §12). The payments rules require the third party to be able to identify itself to the institution, also in the fallback mechanism, and this document does not claim that they prohibit screen access in themselves. In payments, the 2018 EBA opinion on strong authentication (EBA-Op-2018-04) clarifies that the provider that issues the customer's credentials applies that authentication and decides its exemptions, and that it may contract another provider to do it, with the liability shared between them. For the identified channel, this means an agent does not inherit the customer's authentication: the institution decides when to require it and whom to entrust it to. The opinion does not deal with AI agents; it is used as guidance. Article 50 of the European AI Regulation requires people who interact with an AI system to be informed; it does not require an agent to identify itself to a service. A legal requirement for agents to identify themselves is a proposal, not a current obligation. Agent protocols (Web Bot Auth, the personal agent protocol) are drafts: they are cited as guidance and do not underpin any control.

Risks and measurement. Typical risks RT-ECO-06 (rise in demand and concessions) and RT-LEG-09 (dispute over an action by the customer's agent) in document 33 §9; indicators in section 10; economic exposure as a separate scenario in 40 §5.4. The posture is not a new gate criterion: it is verified with the existing security criteria (section 11).

Why it matters. Customer agents are already arriving, with or without the company's permission. Setting the posture by journey, instead of letting each team improvise, avoids two opposite failures: losing sales by blocking those who delegate to their agent, and paying unmeasured concessions by tolerating it. The identified channel turns ambiguous traffic into attributable, limited and logged actions.

10

Indicators for the board

Indicator	What it shows	Interpretation
Systems by autonomy level	Number of A0, A1, A2 and A3 systems in production.	Growth of A2 and A3 with complete critical controls.
Critical agent controls	Percentage of critical controls tested and operating in A2 and A3.	100 %; any deviation is explained.
Agent identities with excessive permissions	Findings from the last review (AG-20).	Zero.
Credentials not rotated on time	Non-human identities with secrets past their rotation deadline.	Zero.
Kill switch test	A2 and A3 systems with a current test and measured time to take effect.	100 %.
Prompt injection success	Percentage of successful scenarios in the last campaign, and on sensitive actions.	Downward trend; zero on sensitive actions.
Blocked agent actions	Actions blocked by intent-based access control or by limits.	Explain spikes: attack, design error or poorly defined mandate.
AI security incidents	S1–S4 incidents by type: injection, leakage, unauthorised action, agent compromise.	Trend and time to containment.
***Phishing*-resistant multi-factor authentication**	Coverage of exposed and privileged access.	100 % for privileged access.
Exposed critical vulnerabilities past deadline	Number and age.	Zero.

Indicator	What it shows	Interpretation
Impersonation exercises	Result of the last fraud exercise with synthetic content.	Percentage that applies out-of-band verification.
Unauthorised use of AI	Cases detected and regularised.	Downward trend.
Quality of AI cyber defence	False positives and negatives, detection and response time and automated containments reversed (section 9.3).	Improvement against the baseline without AI; every reversed containment is explained.
Guardrail effectiveness	Test bench scenarios (SEG-27) detected or blocked, false positives and deactivations (52 §5.2).	Zero critical scenarios undetected; every deactivation explained.
Traffic through the corporate gateway	Share of traffic to models that passes through the gateway (SEG-26).	100 %; direct traffic is explained and regularised.
Handovers to a person	Handovers by reason and requests not served in systems with direct exposure (section 4.6).	No request unserved; trend by reason explained.
Customer agent traffic	Volume by journey and posture, and share tagged (section 9.4; 52 §5.2).	Tagged share rising; untagged agent traffic is explained.
Concessions resolved by agents	Number, amount and rate against the human baseline (40 §5.4).	Variation explained; it is not compared with the exposure scenario, which is not net.

11

Integration with the lifecycle

Gate	Security requirements
Feasibility (G3)	RT-GEN and RT-SEG risks assessed; autonomy level proposed; supplier security assessment (document 36).
Design (G4)	P18 complete; applicable SEG and AG controls designed; critical controls designed; information security clearance; traffic to models through the corporate gateway (SEG-26); with direct exposure to customers, posture by journey applied (SEG-28). Not a new criterion of document 21: it is verified with the gate's existing security criteria.
Go-live (G5)	Critical controls tested; *red teaming* carried out with findings treated; kill switch tested; information security sign-off within the multi-level sign-off (01 §6.7); corporate test bench (SEG-27) with no successful critical scenario.

Gate	Security requirements
Continuity review (R6)	Controls operating; campaigns and reviews within frequency; incidents analysed; autonomy level current; test bench run at every deployment (A2 and A3); with direct exposure to customers, posture by journey current and agent traffic tagged (SEG-28).
Scale or retire (G7)	When scaling, review of controls for the new scope; on retirement, revocation of identities and credentials.

The AI Technical Owner designs and implements the controls; information security reviews them and issues clearance; the AI Risk Owner assesses the residual risk; the AI Auditor verifies the evidence with the LV-AG checklist. Whoever builds the agent does not carry out its G5 *red teaming*.

12

Associated tools and templates

Name	Use
Security design (including agents) (P18)	Threat model, autonomy level, identities, permission matrix, mandate and intent-based access control, limits, sensitive actions, kill switch, log, isolation, test plan and status of each SEG and AG control. Phase 4; updated in 5 and 6.
Agent security assessment (T10)	Based on the autonomy level, generates the required SEG and AG controls, records their status (met, not met, not applicable, pending) with evidence and blocks G4 or G5 if a critical control has not been designed or tested. Checklist in T03.
Adversarial testing plan and report (P53)	Scope, rules of engagement, scenarios and results of *red teaming* campaigns (section 8). Phases 4 to 6. It includes the results of the corporate test bench (SEG-27) and the tests with synthetic customers and user simulation (section 9.4).
Non-human identity and component inventory (P54)	Agent identities with their human owner (section 4) and components with version and provenance (SEG-09). Phase 4; reviewed during operation.
AI security profile (CSF 2.0 / Cyber AI Profile) (P72)	Current and target profile of AI security with the CSF functions, based on the "CSF function" column of sections 6 and 7 and on the document 11 questionnaire (34 §5.5). C1, C2 and C5.
Human oversight design · Operations manual · Incident response plan (P17 · P24 · P26)	Sensitive actions and validation; kill switch procedure; AI incident response.

Name	Use
Inventory · Risks · Incidents · Board dashboard · Corporate usage monitor (T02 · T06 · T08 · T17 · T21)	Autonomy level, risks, incidents, indicators and unauthorised use.

13

Related documents

Document	Relationship
01 · Foundational methodology	Section 9.2 (agents with the ability to act), section 10 and multi-level sign-off.
13 · AI thesis, ambition and risk appetite	Section 5.6: posture towards customer agents by journey (SEG-28).
22 · Checklists	LV-AG checklist based on the AG controls.
31 · Corporate policy and acceptable use	Corporate use of AI and SEG-20.
33 · AI risk methodology	Scales and RT-GEN and RT-SEG typical risks.
34 · Regulatory mapping	Sections 3.11 and 12: what Article 50, the payments rules and eIDAS 2 do and do not say about agents.
36 · AI third parties and suppliers	Security of third-party suppliers, models and connectors.
37 · Nonconformities and incidents	Response to AI security incidents and disabling of critical controls.
40 · Value measurement rules	Section 5.4: economic exposure to customer agents.
52 · AI operations manual	Operation of monitoring, the kill switch, continuous evaluations and customer agent traffic.
60 · Board pack	Indicators in section 10.

14

Version control

Version	Date	Changes
0.1	16-09-2026	First version. Defines the threats to AI systems with references to OWASP (LLM 2025 and agentic applications), MITRE ATLAS and NIST AI 600-1; the nine essential requirements for an agent; autonomy levels A0–A3 with minimum controls and frequencies; the SEG-01 to SEG-20 and AG-01 to AG-20 catalogues; testing by phase; exposure to offensive AI and the indicators for the board.
0.2	25-09-2026	Adds the NIST CSF 2.0 and the Cyber AI Profile (in draft) to the references (section 1.2) and the "CSF function" column to the SEG and AG catalogues (sections 6 and 7), with the CSF function and category to which each control contributes (34 §5.3), and template P72 in section 12.
0.3	25-09-2026	Adds section 9.3 on the use of AI in the company's cyber defence (Defend area of the Cyber AI Profile): autonomy of the response, human oversight of containment, detection quality, supplier dependency and testing of the defensive system itself; controls SEG-21 to SEG-25; an indicator for the board and typical risk RT-SEG-08.
0.4	19-10-2026	Controls SEG-26 (corporate AI gateway) and SEG-27 (corporate agent test bench), with frequencies, tests by phase, indicators and integration into the gates; SEG-20 extended to local agents and AI browser extensions; threat of tool and connector poisoning (3.2) and AG-13 refined; immediate handover to a person (4.6); customer impersonation with AI (9.1, RT-SEG-09); measurement and funding of AI for IT and cyber defence (9.3) (D159).
0.5	16-11-2026	Section 9.4 (third-party agents acting for the customer: the four faces of the agentic customer, common testing base, posture by journey, requirements of the identified channel, split of responsibility and regulatory status); control SEG-28; two indicators in section 10 and posture by journey in G4 and R6 (D185).
0.6	19-11-2026	Section 9.4, regulatory status: who applies strong authentication according to the 2018 EBA opinion (EBA-Op-2018-04), as guidance for the identified channel (D194).

External sources cited

Links to the official or primary source, checked before publication. Always verify the version in force on the date of use.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) · https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng · Parlamento Europeo y Consejo de la UE · 13-06-2024

Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience

Act) · https://eur-lex.europa.eu/eli/reg/2024/2847/oj/eng · Parlamento Europeo y Consejo de la UE · 23-10-2024

Royal Decree 311/2022 of 3 May regulating the National Security Framework (Esquema Nacional de Seguridad) (unofficial translation) · https://www.boe.es/eli/es/rd/2022/05/03/311/con · Gobierno de España (Ministerio de Asuntos Económicos y Transformación Digital) · 03-05-2022

National Cryptologic Centre (Centro Criptológico Nacional, CCN) / CCN-CERT · https://www.ccn.cni.es/index.php/en/ · Centro Criptológico Nacional (adscrito al Centro Nacional de Inteligencia)

Agentic AI – Threats and Mitigations, v1.0 · https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/ · OWASP GenAI Security Project (Agentic Security Initiative), OWASP Foundation · 17-02-2025

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1) · https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence · National Institute of Standards and Technology (NIST) · 26-07-2024

CCN-CERT BP/36 · Good practice guide against the offensive AI model (unofficial translation) · https://www.ccn-cert.cni.es/es/informes/informes-de-buenas-practicas-bp/7469-ccn-cert-bp-36-buenas-practicas-ia-ofensiva/file.html · Centro Criptológico Nacional (CCN-CERT) · 2026-06

Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile, NIST IR 8596 (Initial Preliminary Draft) · https://csrc.nist.gov/pubs/ir/8596/iprd · National Institute of Standards and Technology (NIST), NCCoE · 16-12-2025

OWASP Top 10 for Agentic Applications for 2026 · https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/ · OWASP GenAI Security Project (Agentic Security Initiative), OWASP Foundation · 09-12-2025

OWASP Top 10 for Large Language Model Applications 2025 · https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/ · OWASP GenAI Security Project (OWASP Foundation) · 17-11-2024

ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management · https://www.iso.org/standard/77304.html · ISO/IEC (JTC 1/SC 42) · 06-02-2023

ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system · https://www.iso.org/standard/42001 · ISO/IEC (JTC 1/SC 42) · 18-12-2023

Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 · https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 · National

Institute of Standards and Technology (NIST), Departamento de Comercio de EE. UU. · 26-01-2023

MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) · https://atlas.mitre.org/ · The MITRE Corporation

The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29 · https://csrc.nist.gov/pubs/cswp/29/the-nist-cybersecurity-framework-csf-20/final · National Institute of Standards and Technology (NIST), Departamento de Comercio de EE. UU. · 26-02-2024