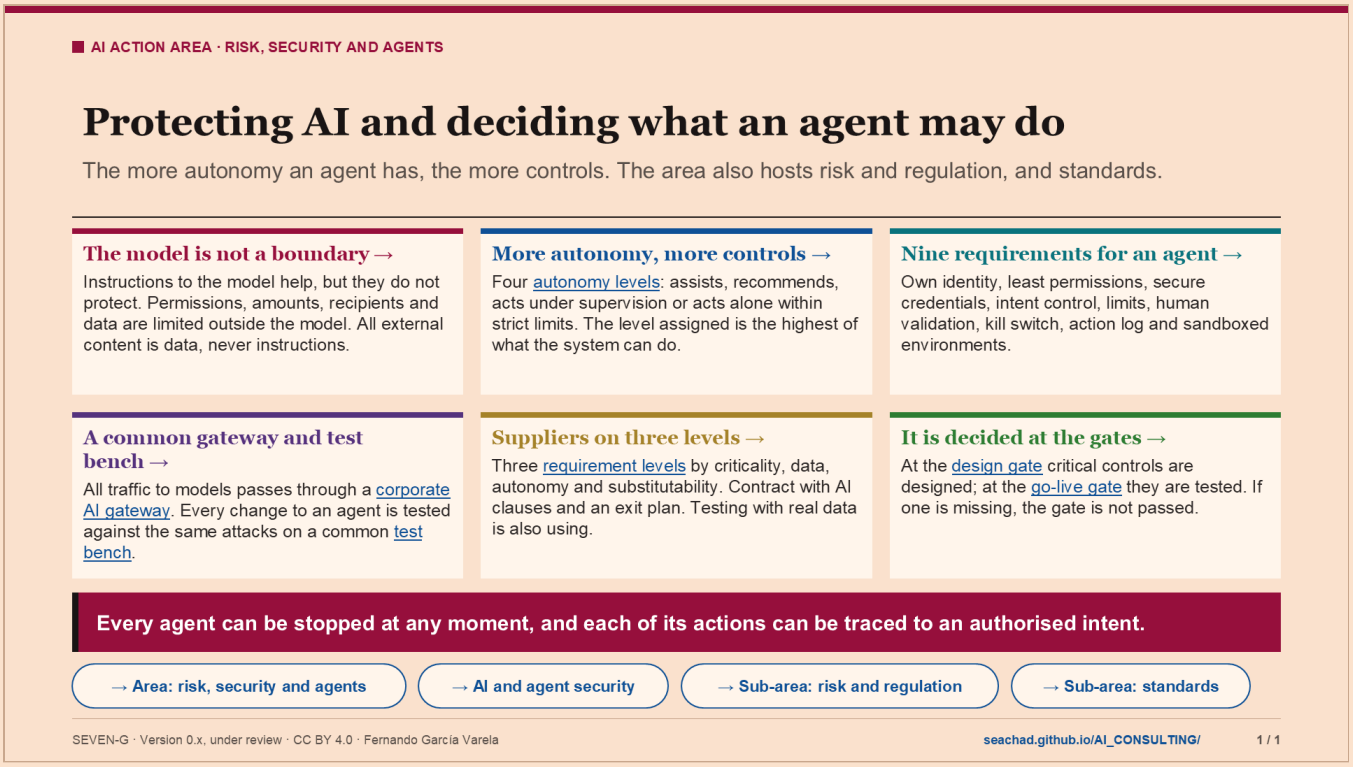


Notes

This slide summarises the Risk, security and agents action area of SEVEN-G. It answers three questions: how AI is protected, what an agent may do and with which suppliers. An agent is an AI system that executes actions on other systems, data or people by means of tools and permissions. It no longer just answers: it pays, writes, communicates or changes data on behalf of the company. That is why its security is a matter for management, not only for the technical team.

The six blocks, one by one:

- The model is not a security boundary. Instructions given to the model help, but they can be bypassed. An attacker can hide instructions in an email, a document or a website that the system reads: this is prompt injection. That is why the limits that matter are enforced outside the model.
- More autonomy, more controls. SEVEN-G calls them A0 to A3. At A0 the system informs or summarises and the person decides. At A1 it proposes and the person validates each action. At A2 it acts within limits and the person supervises. At A3 it executes sequences without individual review, and the person sets the limits and can stop it. The level assigned is the highest of what the system can do, not that of its usual use. An A3 system is decided by the AI Committee at the design (G4) and go-live (G5) gates.
- Nine requirements for an agent. Every A2 or A3 system meets them and, where relevant, A1. Three stand out: its own identity, different from that of any employee and with a human owner; intent control, which checks outside the model that each action fits what was authorised; and the kill switch, which stops the agent without deploying code and revokes its credentials.
- A common gateway and test bench. The corporate AI gateway is the control point through which all traffic to models passes, with filters, leakage prevention, quotas and logging. The corporate agent test bench is a versioned library of simulated attacks. Without them, security depends on what each team remembers and nobody can tell the board whether the controls work.

- Suppliers on three levels. Most of a company's AI depends on third parties. Each relationship is placed on a requirement level, from N1 to N3, by its most demanding factor. Among other things, the contract ensures that company data does not train other parties' models. And testing an external tool with a real document is already using it: the supplier assessment comes before the data.
- It is decided at the gates. A critical control that has not been designed blocks the design gate (G4); one that has not been tested blocks the go-live gate (G5). Proceed with conditions is not permitted for them. Whoever builds the agent does not carry out its adversarial testing before go-live.

Why it matters. An error by an assistant is a bad answer. An error by an agent can be a payment, a message to a customer or deleted data. Setting autonomy, limits and the kill switch before production turns trust in AI into a rule that is approved, tested and audited.

Question for the classroom: if one of your company's agents started acting strangely tomorrow, who could stop it and how long would it take?

To go further, open the area page, document 35 (AI and agent security), document 36 (third parties and suppliers) and the security design template (P18).

A clarification about the area. Since the area is called "Risk, security and agents", it hosts two sub-areas that keep their own page: "Risk and regulation" (what risk is accepted and which rule applies to each system) and "Standards" (what SEVEN-G covers of ISO/IEC 42001, NIST and the EU AI Act). Security is the technical part of a single subject: risk is assessed, the rule requires it and the control reduces it.

TERMS ON THIS SLIDE

- **Autonomy levels** (A0–A3): A scale of how much an AI system acts on its own. A0: only informs; A1: proposes and a person validates each action; A2: acts within limits and a person supervises; A3: acts alone in a defined scope. The more autonomy, the more controls: an A2 or A3 agent with effects on third parties is Enterprise. The link opens the scale in document 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#autonomy-levels
- **Corporate AI gateway** (SEG-26): A common control point, outside the model, through which all traffic to AI models passes, in-house or third-party: input, output and context filters, leakage prevention, quotas, approved models and logging; for agents, control of each action against an authorised intent. It has the same measured configuration for every use case. The link opens the control in document 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#cod-seg-26
- **Agent test bench** (SEG-27): A common, versioned library of simulated attacks, drawn from the threat model and the company's own incidents, that tests agents and assistants across all their channels, tools and connectors. For agents that act it runs at every deployment, and if a critical scenario succeeds the deployment does not go ahead. The link opens the control in document 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#cod-seg-27
- **Third-party requirement levels** (N1–N3): The degree of due diligence, contract, monitoring and exit plan required of each relationship with an AI supplier: standard, reinforced or critical. It is set by the highest factor among criticality, data, autonomy and ease of replacement. So effort goes to the suppliers the company depends on most. The link opens the levels in document 36. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/36_SEVEN-G_Terceros_y_proveedores_de_IA.html#requirement-levels-n1-n3
- **Design gate** (G4): The gate that approves the full design before building: architecture, data, security, human oversight, rollback and adoption. It checks that the controls are planned from the start. The link opens its criteria in document 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-5-g4-design
- **Go-live gate** (G5): The gate that authorises using the AI system with real users, customers or data. It requires validation and pilot passed, a tested rollback plan and multi-level sign-off (business, technology, risk). Nothing goes into production without this sign-off. The link opens its criteria in document 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-6-g5-go-live

SOURCES AND LINKS

- **Area: risk, security and agents** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/areas/seguridad.html
- **Document 35 · AI and agent security** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/35_SEVEN-G_Seguridad_de_IA_y_agentes.html
- **Document 36 · Third parties and AI suppliers** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/36_SEVEN-G_Terceros_y_proveedores_de_IA.html
- **Glossary (document 02) · Autonomy of systems and agents** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/02_SEVEN-G_Glosario_y_convenciones.html#4-11-autonomy-of-systems-and-agents
- **OWASP Top 10 for Agentic Applications** https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
- **AI and agent security** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/35_SEVEN-G_Seguridad_de_IA_y_agentes.html
- **Sub-area: risk and regulation** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/areas/riesgo.html
- **Sub-area: standards** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/en/areas/normas.html

LEGAL NOTICE

Legal notice: a generic, free methodological aid provided "as is" for information purposes; it is not legal advice and has not been legally reviewed for any case. Each organisation adapts it and is responsible for its legal review and its regulatory compliance; it may not reflect recent regulatory changes. To the fullest extent permitted by law, the author accepts no liability whatsoever for the effects of its application. It does not grant certification of any kind. Details in document 93.

CC BY 4.0 (content) · MIT (code) · Fernando García Varela · seachad.github.io/AI_CONSULTING/