Risk, security and agents

How AI is protected, what an agent may do and with which suppliers.

> *"The model is not a security boundary. An agent that acts needs its own identity, minimum permissions and a tested kill switch; and accountability is not outsourced to the supplier."*

WHAT IT IS

What changes with AI

AI opens new attacks: hidden instructions in the content it reads, tampered tools, data leakage. Agents also act. SEVEN-G assumes the model can be deceived: external content is not trusted, every action needs an authorised intent and it can always be stopped.

Implication: How many of your company's agents move money, data or systems without a person validating it?

Risk: A deceived agent carries out real actions with the permissions it was given.

WHAT IT PROPOSES

Autonomy with limits

Autonomy is set at four levels, from only suggesting to acting alone. Every agent has an identity, minimum permissions, limits, an action log and a kill switch tested without deploying code. A person validates what cannot be undone. Traffic to the models goes through a corporate gateway.

Implication: Has the kill switch of your agents been tested, or does it only exist in the design?

Risk: Without a tested kill switch, a failing agent keeps acting until someone notices.

HOW IT IS CHECKED

Test like the attacker

A critical control not designed blocks the build; not tested, it blocks go-live. The builder does not run its own adversarial tests. With agents that act, every deployment goes through a corporate test bench. With suppliers: plan the exit before entering, and testing with internal data is also use.

Implication: Which AI supplier holds company data today without a prior assessment, even in a test?

Risk: An unassessed supplier may use your data or leave the company with no way out.

FOUR QUESTIONS FOR THE BOARD

1. **What can each agent do without a person in the loop?**
   An inventory of agents with their autonomy level, permissions and owner. Letting an agent act alone requires an AI committee decision.
2. **Are our critical AI controls tested?**
   All critical controls tested, the kill switch included, and no successful injection attack on sensitive actions.
3. **Does all traffic to the models go through one control point?**
   The corporate AI gateway: filters, approved models and logging; if it fails, it closes on sensitive actions.
4. **Can we leave our most critical AI supplier?**
   Requirements on the supplier by criticality, data clauses and, for the most critical ones, a tested exit plan.

WHAT FAILS IF IT IS NOT GOVERNED

WHAT FAILS	LEVEL	WHAT TO DO
Hidden instructions that deceive an agent	HIGH	Test each agent against prompt injection
Agents with more permissions than they need	HIGH	Review the identity and permissions of each agent
Impersonation fraud with synthetic voice or video	HIGH	Approve no payment on voice or image alone
A kill switch that nobody has tested	HIGH	Test the kill switch of each agent
Unassessed AI embedded in third-party products	MEDIUM	Keep embedded AI switched off until assessed
Internal data in tests with suppliers	MEDIUM	Assess the supplier before any test
Dependence on a single model supplier	MEDIUM	Prepare and test the supplier exit plan

FURTHER READING

AI and agent security
Document 35 · threats, autonomy and controls

Security design
Editable template

AI third parties and suppliers
Document 36 · requirements, data and exit

OWASP Top 10 for Agentic Applications 2026
OWASP GenAI Security Project · 09-12-2025

Fernando García Varela · CC BY 4.0 · Version 0.x, under review

Generic help offered free of charge, with no legal review for any specific case: each organisation adapts it, identifies the regulation that applies to it and is responsible for its own compliance. It is neither advice nor certification. To the maximum extent permitted by law, the author accepts no liability for the effects of its application.