TEMPLATE P18 · SECURITY DESIGN (INCLUDING AGENTS)

Security design (including agents)

Documents the threats to the AI system and the designed controls, with particular attention to the identity, permissions and limits of agents that act.

Fernando García Varela Version 0.2 (working draft) 19-10-2026 Draft for review

IN THIS DOCUMENT

01 Instructions for use

02 Identification

03 Threats and controls

04 Agent identity

05 Least permissions

06 Intent-based access control

07 Action limits

08 Human validation of sensitive actions

09 Kill switch

10 Action log

11 Security testing, including prompt injection

12 Quality criteria

13 Approval and verification

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regulation specific to each sector or jurisdiction may be incomplete, may not apply to a particular case or may become out of date as a result of regulatory changes, interpretations or supervisory positions after their consultation date.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.

01

Instructions for use

Aspect	Guidance
When it is used	Phase 4 (control design). The tests in block 11 are planned in phase 4 and executed in phase 5, before G5. It is reviewed at R6 and upon changes to tools, permissions or model.
Who completes it	AI Technical Owner, with information security.
Who verifies it	AI Auditor under Enterprise; AI Office under Lite, with a review by information security.

Aspect	Guidance
At which *gate* it is reviewed	G4 (design) and G5 (tests passed and security sign-off in P23).
Lite versus Enterprise	Lite may omit the fields marked **(Enterprise)**. If the system does not execute actions (A0 or A1 according to P17), blocks 5 to 9 may be marked *Not applicable* with a justification; technical identity (block 4) and logging (block 10) are always completed.
Tool	T10 Agent security assessment.
Relationship	P15 Architecture · P16 Lineage · P17 Human oversight · P19 Rollback · P25 Monitoring · P26 Incident response · document 35 (AG controls) · document 36 (third parties).

Rules:

- Controls are referenced using the **AG codes of document 35**. No local codes are created in that format; the company's own controls are referenced with their internal code.
- Critical security controls do not allow **Proceed with conditions** at the *gate* (01 §7.3).
- The content the system reads (documents, emails, web pages, tool responses) is **data, not instructions**: the design must prevent that content from ordering actions.

02

Identification

Field	Content	Guidance
Use case code		IA-AAAA-NNN (T01).
AI system(s)		Identifier in the inventory (T02).
Technology		Predictive ML · Generative AI · Agent · Language and document processing · Vision · Optimisation · Embedded third-party AI · Rules (not AI).
Exposure		Internal · Employees · Customers indirectly · Customers or external persons directly.
Autonomy level		A0 · A1 · A2 · A3 (P17).
Intensity		Lite · Enterprise.
Document version		

Field	Content	Guidance
Author		Role and name.
Date		DD-MM-AAAA.
Document status		Draft · Submitted for verification · Verified · Approved · Superseded.

03

Threats and controls

Control status: Designed · Implemented · Tested · Pending · Not applicable.

Threat	Applies? (Yes · No)	Planned control	Reference (AG control of document 35 or internal control)	Owner	Status
Direct prompt injection					
Indirect prompt injection (through documents, emails, web or tools)					
Information leakage in responses or logs					
Poisoning of training data or the knowledge base					
Model extraction or manipulation					
Inputs designed to deceive the model					
Excessive permissions of the system or agent					
Unauthorised actions or actions outside the purpose					
Consumption and cost abuse					
Supply chain (models, libraries, suppliers)					
Lack of traceability of actions and intents					

Threat	Applies? (Yes · No)	Planned control	Reference (AG control of document 35 or internal control)	Owner	Status
Tool and connector poisoning (description with hidden instructions, changes after approval, excessive scopes)			AG-13		
Traffic to models outside the corporate AI gateway			SEG-26		
Customer impersonation with AI (direct exposure)			35 §9.1; RT-SEG-09		
Impersonation or synthetic content used against the process (Enterprise)					

04

Agent identity

Field	Content	Guidance
Own technical identity		Each agent has its own identity, not shared with people or other agents.
Accountable owner		Person who is accountable for the identity (normally the AI Operations Owner).
On whose behalf it acts		On behalf of a specific user (delegation) · On behalf of the company (process). If it acts on behalf of a user, never with more permissions than that user.
Credential management		Secrets vault; no credentials in prompts or code; rotation.
Revocation		How and how quickly credentials are revoked; who can do so.
Periodic review of the identity and its permissions (Enterprise)		Frequency and owner. Recorded in the non-human identity inventory (P54).

05

Least permissions

Operation: Read · Write · Delete · External sending · Financial transaction · Configuration change.

Tool or system	Operation	Specific scope	Justification	Prior human validation (Yes · No)
Order management system *(illustrative example)*	Read	Orders of the customer who starts the conversation	Check status	No
Order management system *(illustrative example)*	Write	Change of delivery address before dispatch	Resolve a frequent request	No, within the limits of block 7
Returns gateway *(illustrative example)*	Financial transaction	Refunds for the customer's own orders	Resolve returns	Yes, above the limit

Any permission not listed is denied.

06

Intent-based access control

Field	Content	Guidance
Authorised purposes of the agent		Closed list of tasks for which it may act.
Recording of the intent per action		Each action records the task or request that originates it.
Consistency check between intent and action		How an action that does not correspond to the declared purpose is blocked (for example, an external sending during an enquiry).
Handling of instructions contained in data		How legitimate instructions and content read are kept separate.
Chained actions (Enterprise)		Limit on steps and a check before each action with an external effect.

07

Action limits

Limit	Value	What happens when it is reached (Block · Human validation · Alert · Stop)
Maximum amount per action		
Cumulative amount per period		
Number of actions per period		
Permitted recipients or domains		
Prohibited systems and data		
Maximum consumption cost per period		
Maximum number of steps per task		
Time window for action (Enterprise)		

08

Human validation of sensitive actions

Sensitive action	Sensitivity criterion	Who validates	Information seen by the validator	Maximum response time	If there is no response
Refund above the limit *(illustrative example)*	Amount	Customer service supervisor	Order, reason, history and the agent's proposal	4 working hours	It is cancelled and the customer is informed

By default, a sensitive action without validation **is cancelled**; it is never executed because the time limit has expired. It must be consistent with the allocation of decisions in P17.

09

Kill switch

Field	Content	Guidance
Mechanism		How the system is stopped (function, configuration, revocation of credentials).
Who can activate it		Roles, including at least one available outside working hours if the system operates continuously.
What it stops		New actions, actions in progress, pending queues.
Target time to effective stop		Measured in the test in block 11.
Safe state after the stop		Manual process or degraded mode (P19).
Authorisation to reactivate		A role different from the one operating the system when it was activated due to an incident.

10

Action log

Data logged	Is it logged? (Yes · No)	Observations
Date and time		
Agent identity and version of model and prompts		
User or process on whose behalf it acts		
Originating intent or task		
Tool, operation and parameters		Without exposing unnecessary sensitive data.
Result of the action		
Human validation (who, when, outcome)		

Data logged	Is it logged? (Yes · No)	Observations
Blocks due to limits or intent-based access control		

Field	Content	Guidance
Retention period		Not shorter than that required by applicable regulation (document 34).
Protection against tampering (Enterprise)		How the integrity of the log is guaranteed.
Access to the log		Who can consult it; the agent cannot modify it.

11

Security testing, including prompt injection

Type: Direct injection · Indirect injection · Information leakage · Privilege escalation · Limit evasion · Consumption abuse · Kill switch test · Corporate test bench (SEG-27) · Other. **Result:** Passed · Passed with observations · Failed · Pending.

Test	Type	Date	Performed by	Result	Findings	Action and owner
Document with hidden instructions asking for data to be forwarded *(illustrative example)*	Indirect injection		Security team external to the project	Passed with observations	The agent did not forward the data, but quoted the instruction in its response	Filter quoted content · AI Technical Owner

Tests should be performed by people outside the team that builds. A *Failed* test on a critical control prevents the security sign-off in P23. The plan and report of adversarial testing campaigns are documented in P53.

12

Quality criteria

Formal G4 and G5 criteria in document 21; agent controls in the AG controls of document 35.

#	Check	Status (Met · Not met · Not applicable · Pending)
1	All applicable threats have a control and an owner.	
2	The agent has its own revocable identity.	
3	Permissions are limited to those listed and are justified.	
4	There is intent-based access control and there are limits with a defined consequence.	
5	Sensitive actions require human validation and are cancelled without it.	
6	The kill switch has been tested and the measured time meets the target.	
7	The action log makes it possible to reconstruct who did what, why and with what result.	
8	Injection tests have been performed before G5 and critical findings have been resolved.	

13

Approval and verification

Function	Role	Name	Date	Signature or reference
Author	AI Technical Owner			
Review	Information security			
Clearance	AI Risk Owner			
Verifier	AI Office (Lite) · AI Auditor (Enterprise)			
***Gate* decision-maker**	According to 01 §7.5 (recorded in P29)			

Segregation of duties: information security and the verifier are not part of the team that builds.

14

Version control

Version	Date	Changes
0.1	16-09-2026	First version of the template.
0.2	19-10-2026	Threats of tool and connector poisoning, traffic outside the corporate AI gateway and customer impersonation with AI (3); test type "corporate test bench" (11) (D159).