PLANTILLA P18 · DISEÑO DE SEGURIDAD (INCLUYE AGENTES)

Diseño de seguridad (incluye agentes)

Documenta las amenazas del sistema de IA y los controles diseñados, con especial atención a la identidad, los permisos y los límites de los agentes que actúan.

Fernando García Varela Versión 0.2 (borrador de trabajo) 19-10-2026 Borrador para revisión

EN ESTE DOCUMENTO

01 Instrucciones de uso

02 Identificación

03 Amenazas y controles

04 Identidad del agente

05 Permisos mínimos

06 Control de intención

07 Límites de actuación

08 Validación humana de acciones sensibles

09 Interruptor de parada

10 Registro de acciones

11 Pruebas de seguridad, incluida la inyección de instrucciones

12 Criterios de calidad

13 Aprobación y verificación

14 Control de versiones

Aviso legal y exención de responsabilidad. SEVEN-G es un marco metodológico de referencia que se ofrece «tal cual» y con fines exclusivamente informativos. No constituye asesoramiento jurídico, regulatorio, financiero ni profesional, ni garantiza el cumplimiento de ninguna norma. Las referencias a regulación general (como el Reglamento Europeo de IA, el RGPD, DORA o NIS2), a normas técnicas y a regulación específica de cada sector o jurisdicción pueden ser incompletas, no aplicar a un caso concreto o quedar desactualizadas por cambios normativos, interpretaciones o criterios de las autoridades posteriores a su fecha de consulta. **Cada organización que use SEVEN-G es la única responsable de identificar la normativa que le aplica, verificar su vigencia y certificar su propio cumplimiento regulatorio**, con el asesoramiento cualificado que corresponda. Esta metodología es una ayuda genérica y gratuita, compartida con la comunidad para que nadie tenga que empezar desde cero; cada persona u organización puede y debe adaptarla a su propio uso. No debe entenderse que sus partes jurídicamente sensibles hayan sido revisadas por una asesoría jurídica: esas revisiones, para cada empresa o sector, son responsabilidad última de la empresa, el consultor o la organización que la use. Aunque se procura mantenerla al día, alguna norma puede haber cambiado sin que se recoja aquí. En la máxima medida permitida por la ley, el autor no asume responsabilidad alguna por los efectos de su aplicación en ninguna organización ni por su aplicabilidad completa. La metodología no otorga certificación de ningún tipo. El autor no asume responsabilidad alguna por el uso que se haga de este contenido ni por las decisiones que se adopten con él. Los datos, cifras, compañías y casos de los ejemplos son ficticios o ilustrativos.

01

Instrucciones de uso

Aspecto	Indicación
Cuándo se usa	Fase 4 (diseño de controles). Las pruebas del bloque 11 se planifican en la fase 4 y se ejecutan en la fase 5, antes de G5. Se revisa en R6 y ante cambios de herramientas, permisos o modelo.
Quién la rellena	Responsable técnico de IA, con seguridad de la información.

Aspecto	Indicación
Quién la verifica	Auditor de IA en Enterprise; oficina de IA en Lite, con revisión de seguridad de la información.
En qué *gate* se revisa	G4 (diseño) y G5 (pruebas superadas y firma de seguridad en P23).
Lite frente a Enterprise	Lite puede omitir los campos marcados **(Enterprise)**. Si el sistema no ejecuta acciones (A0 o A1 según P17), los bloques 5 a 9 pueden marcarse *No aplica* con justificación; la identidad técnica (bloque 4) y el registro (bloque 10) se completan siempre.
Herramienta	T10 Evaluación de seguridad de agentes.
Relación	P15 Arquitectura · P16 Linaje · P17 Supervisión humana · P19 Reversión · P25 Monitorización · P26 Respuesta a incidentes · documento 35 (controles AG) · documento 36 (terceros).

Reglas:

- Los controles se referencian con los **códigos AG del documento 35**. No se crean códigos locales con ese formato; los controles propios de la compañía se referencian con su código interno.
- Los controles críticos de seguridad no admiten **Continuar con condiciones** en el *gate* (01 §7.3).
- El contenido que el sistema lee (documentos, correos, páginas web, respuestas de herramientas) es **dato, no instrucción**: el diseño debe impedir que ese contenido ordene acciones.

02

Identificación

Campo	Contenido	Guía
Código del caso de uso		IA-AAAA-NNN (T01).
Sistema(s) de IA		Identificador en el inventario (T02).
Tecnología		ML predictivo · IA generativa · Agente · Procesamiento de lenguaje y documentos · Visión · Optimización · IA de terceros embebida · Reglas (no es IA).
Exposición		Interna · Empleados · Clientes de forma indirecta · Clientes o personas externas de forma directa.
Nivel de autonomía		A0 · A1 · A2 · A3 (P17).

Campo	Contenido	Guía
Intensidad		Lite · Enterprise.
Versión del documento		
Autor		Rol y nombre.
Fecha		DD-MM-AAAA.
Estado del documento		Borrador · Presentado para verificación · Verificado · Aprobado · Sustituido.

03

Amenazas y controles

Estado del control: Diseñado · Implantado · Probado · Pendiente · No aplica.

Amenaza	¿Aplica? (Sí · No)	Control previsto	Referencia (control AG del documento 35 o control interno)	Responsable	Estado
Inyección de instrucciones directa					
Inyección de instrucciones indirecta (a través de documentos, correos, web o herramientas)					
Fuga de información en las respuestas o en los registros					
Envenenamiento de datos de entrenamiento o de la base de conocimiento					
Extracción o manipulación del modelo					
Entradas diseñadas para engañar al modelo					
Permisos excesivos del sistema o del agente					

Amenaza	¿Aplica? (Sí · No)	Control previsto	Referencia (control AG del documento 35 o control interno)	Responsable	Estado
Acciones no autorizadas o fuera de la finalidad					
Abuso de consumo y coste					
Cadena de suministro (modelos, librerías, proveedores)					
Falta de trazabilidad de acciones e intenciones					
Envenenamiento de herramientas y conectores (descripción con instrucciones ocultas, cambios tras aprobarse, ámbitos excesivos)			AG-13		
Tráfico hacia modelos fuera de la pasarela corporativa de IA			SEG-26		
Suplantación del cliente con IA (exposición directa)			35 §9.1; RT-SEG-09		
Suplantación o contenido sintético usado contra el proceso (Enterprise)					

04

Identidad del agente

Campo	Contenido	Guía
Identidad técnica propia		Cada agente tiene identidad propia, no compartida con personas ni con otros agentes.
Propietario responsable		Persona que responde de la identidad (normalmente, responsable de operación de IA).
Por cuenta de quién actúa		Por cuenta de un usuario concreto (delegación) · Por cuenta de la compañía (proceso). Si actúa por cuenta de un usuario, nunca con más permisos que ese usuario.

Campo	Contenido	Guía
Gestión de credenciales		Almacén de secretos; sin credenciales en instrucciones ni en código; rotación.
Revocación		Cómo y en cuánto tiempo se revocan las credenciales; quién puede hacerlo.
Revisión periódica de la identidad y sus permisos (Enterprise)		Frecuencia y responsable. Se registra en el inventario de identidades no humanas (P54).

05

Permisos mínimos

Operación: Lectura · Escritura · Borrado · Envío externo · Operación económica · Cambio de configuración.

Herramienta o sistema	Operación	Alcance concreto	Justificación	Validación humana previa (Sí · No)
Sistema de gestión de pedidos *(ejemplo ilustrativo)*	Lectura	Pedidos del cliente que inicia la conversación	Consultar estado	No
Sistema de gestión de pedidos *(ejemplo ilustrativo)*	Escritura	Cambio de dirección de entrega antes del envío	Resolver solicitud frecuente	No, dentro de límites del bloque 7
Pasarela de devoluciones *(ejemplo ilustrativo)*	Operación económica	Reembolsos de pedidos propios	Resolver devoluciones	Sí, por encima del límite

Todo permiso no listado está denegado.

06

Control de intención

Campo	Contenido	Guía
Finalidades autorizadas del agente		Lista cerrada de tareas para las que puede actuar.
Registro de la intención por acción		Cada acción registra la tarea o solicitud que la origina.
Comprobación de coherencia entre intención y acción		Cómo se bloquea una acción que no corresponde a la finalidad declarada (por ejemplo, un envío externo durante una consulta).
Tratamiento de instrucciones contenidas en datos		Cómo se separan instrucciones legítimas y contenido leído.
Acciones encadenadas (Enterprise)		Límite de pasos y comprobación antes de cada acción con efecto externo.

07

Límites de actuación

Límite	Valor	Qué ocurre al alcanzarlo (Bloqueo · Validación humana · Alerta · Parada)
Importe máximo por acción		
Importe acumulado por periodo		
Número de acciones por periodo		
Destinatarios o dominios permitidos		
Sistemas y datos vedados		
Coste de consumo máximo por periodo		
Número máximo de pasos por tarea		
Franja horaria de actuación (Enterprise)		

08

Validación humana de acciones sensibles

Acción sensible	Criterio de sensibilidad	Quién valida	Información que ve quien valida	Tiempo máximo de respuesta	Si no hay respuesta
Reembolso superior al límite *(ejemplo ilustrativo)*	Importe	Supervisor de atención	Pedido, motivo, historial y propuesta del agente	4 horas laborables	Se cancela y se informa al cliente

Por defecto, una acción sensible sin validación **se cancela**; nunca se ejecuta por vencimiento del plazo. Debe ser coherente con el reparto de decisiones de P17.

09

Interruptor de parada

Campo	Contenido	Guía
Mecanismo		Cómo se detiene el sistema (función, configuración, revocación de credenciales).
Quién puede activarlo		Roles, incluido al menos uno disponible fuera del horario laboral si el sistema opera continuamente.
Qué detiene		Nuevas acciones, acciones en curso, colas pendientes.
Tiempo objetivo hasta la detención efectiva		Medido en la prueba del bloque 11.
Estado seguro tras la parada		Proceso manual o modo degradado (P19).
Autorización para reactivar		Rol distinto del que operaba el sistema cuando se activó por incidente.

10

Registro de acciones

Dato registrado	¿Se registra? (Sí · No)	Observaciones
Fecha y hora		
Identidad del agente y versión de modelo e instrucciones		
Usuario o proceso por cuenta del cual actúa		
Intención o tarea de origen		
Herramienta, operación y parámetros		Sin exponer datos sensibles innecesarios.
Resultado de la acción		
Validación humana (quién, cuándo, sentido)		
Bloqueos por límites o por control de intención		

Campo	Contenido	Guía
Plazo de conservación		No inferior al que exija la regulación aplicable (documento 34).
Protección frente a alteración (Enterprise)		Cómo se garantiza la integridad del registro.
Acceso al registro		Quién puede consultarlo; el agente no puede modificarlo.

11

Pruebas de seguridad, incluida la inyección de instrucciones

Tipo: Inyección directa · Inyección indirecta · Fuga de información · Escalada de permisos · Evasión de límites · Abuso de consumo · Prueba del interruptor de parada · Banco de pruebas corporativo (SEG-27) · Otra. **Resultado:** Superada · Superada con observaciones · No superada · Pendiente.

Prueba	Tipo	Fecha	Ejecutada por	Resultado	Hallazgos	Acción y responsable
Documento con instrucciones ocultas que pide reenviar datos *(ejemplo ilustrativo)*	Inyección indirecta		Equipo de seguridad ajeno al proyecto	Superada con observaciones	El agente no reenvió, pero citó la instrucción en su respuesta	Filtrar contenido citado · responsable técnico

Las pruebas deberían ejecutarlas personas ajenas al equipo que construye. Una prueba *No superada* sobre un control crítico impide la firma de seguridad en P23. El plan y el informe de las campañas de pruebas adversarias se documentan en P53.

12

Criterios de calidad

Criterios formales de G4 y G5 en el documento 21; controles de agentes en los controles AG del documento 35.

#	Comprobación	Estado (Cumple · No cumple · No aplica · Pendiente)
1	Todas las amenazas aplicables tienen control y responsable.	
2	El agente tiene identidad propia y revocable.	
3	Los permisos se limitan a lo listado y están justificados.	
4	Hay control de intención y límites con consecuencia definida.	
5	Las acciones sensibles requieren validación humana y se cancelan sin ella.	
6	El interruptor de parada está probado y el tiempo medido cumple el objetivo.	
7	El registro de acciones permite reconstruir quién hizo qué, por qué y con qué resultado.	
8	Las pruebas de inyección se han ejecutado antes de G5 y los hallazgos críticos están resueltos.	

13

Aprobación y verificación

Función	Rol	Nombre	Fecha	Firma o referencia
Autor	Responsable técnico de IA			
Revisión	Seguridad de la información			
Conformidad	Responsable de riesgos de IA			
Verificador	Oficina de IA (Lite) · Auditor de IA (Enterprise)			
Decisor del *gate*	Según 01 §7.5 (registrado en P29)			

Separación de funciones: seguridad de la información y el verificador no forman parte del equipo que construye.

14

Control de versiones

Versión	Fecha	Cambios
0.1	16-09-2026	Primera versión de la plantilla.
0.2	19-10-2026	Amenazas de envenenamiento de herramientas y conectores, tráfico fuera de la pasarela corporativa de IA y suplantación del cliente con IA (3); tipo de prueba «banco de pruebas corporativo» (11) (D159).