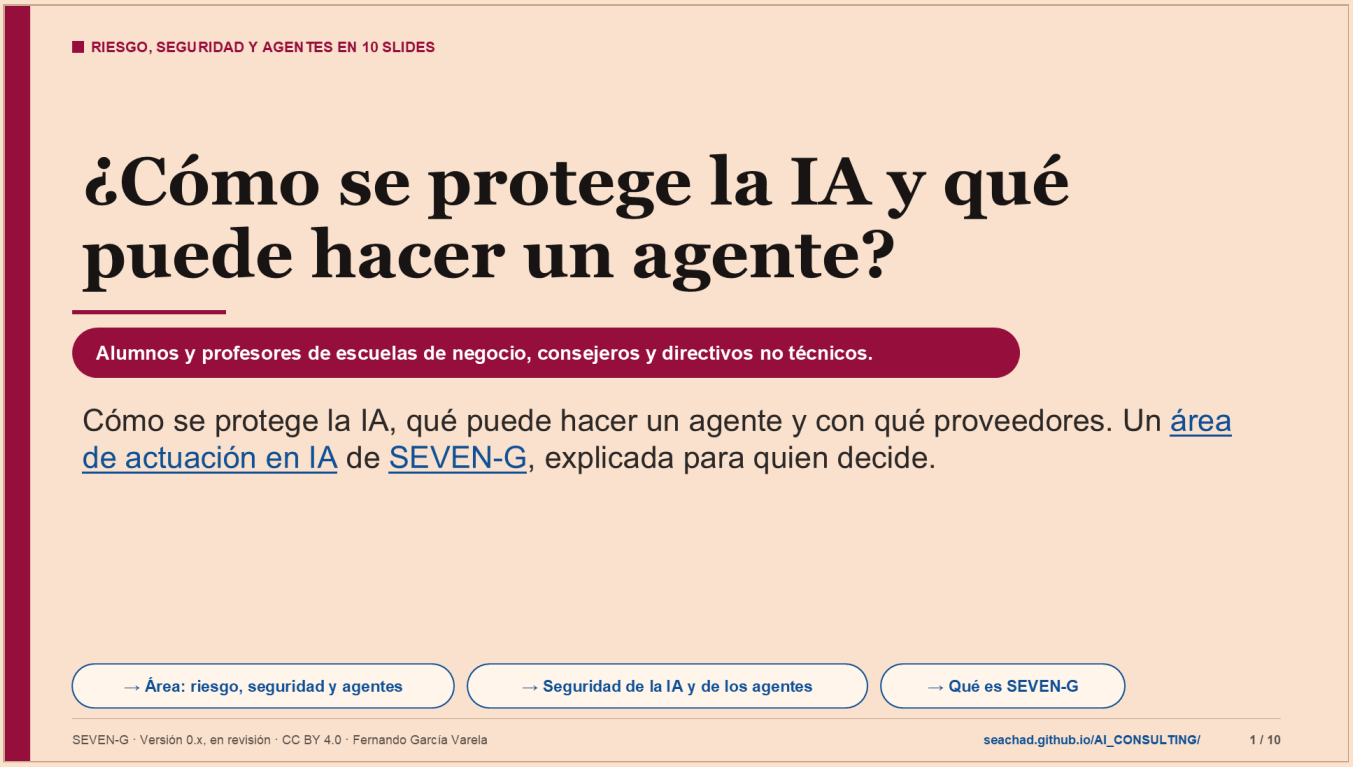


Notas

Esta presentación explica el área de actuación Riesgo, seguridad y agentes de SEVEN-G, el marco de gobierno, valor y medición de la IA. Responde a tres preguntas: cómo se protege la IA, qué puede hacer un agente y con qué proveedores se trabaja.

Para qué sirve. Hasta hace poco, la IA de una compañía respondía preguntas o resumía documentos. Hoy aparecen los agentes: sistemas que ejecutan acciones con herramientas y permisos. Pagan, escriben a clientes, cambian datos o tocan sistemas de producción. Cuando un sistema actúa, un error o un ataque deja de ser una respuesta mala y se convierte en una acción con efectos. Por eso la seguridad de la IA es un asunto de la dirección y del consejo, no solo del equipo técnico.

Qué va a encontrar en las diez slides:

- Por qué importa: qué cambia la IA en el riesgo de la compañía.
- Los conceptos clave, con las definiciones del glosario (documento 02).
- Cómo lo trata SEVEN-G: niveles de autonomía A0 a A3, qué se exige en cada uno y quién decide.
- En qué puerta del ciclo de vida se comprueba cada cosa y con qué plantillas.
- Un caso ficticio, los errores frecuentes, preguntas para la dirección o el aula, la relación con otras áreas y cómo empezar.

Por qué importa. La regla de fondo del área es sencilla: cuanta más autonomía tiene un agente, más controles se le exigen. Y los controles se deciden antes de producción, no después del primer incidente.

Pregunta para el aula: ¿cuántos sistemas de su organización ya ejecutan acciones por sí mismos, y quién lo sabe?

TÉRMINOS DE ESTA SLIDE

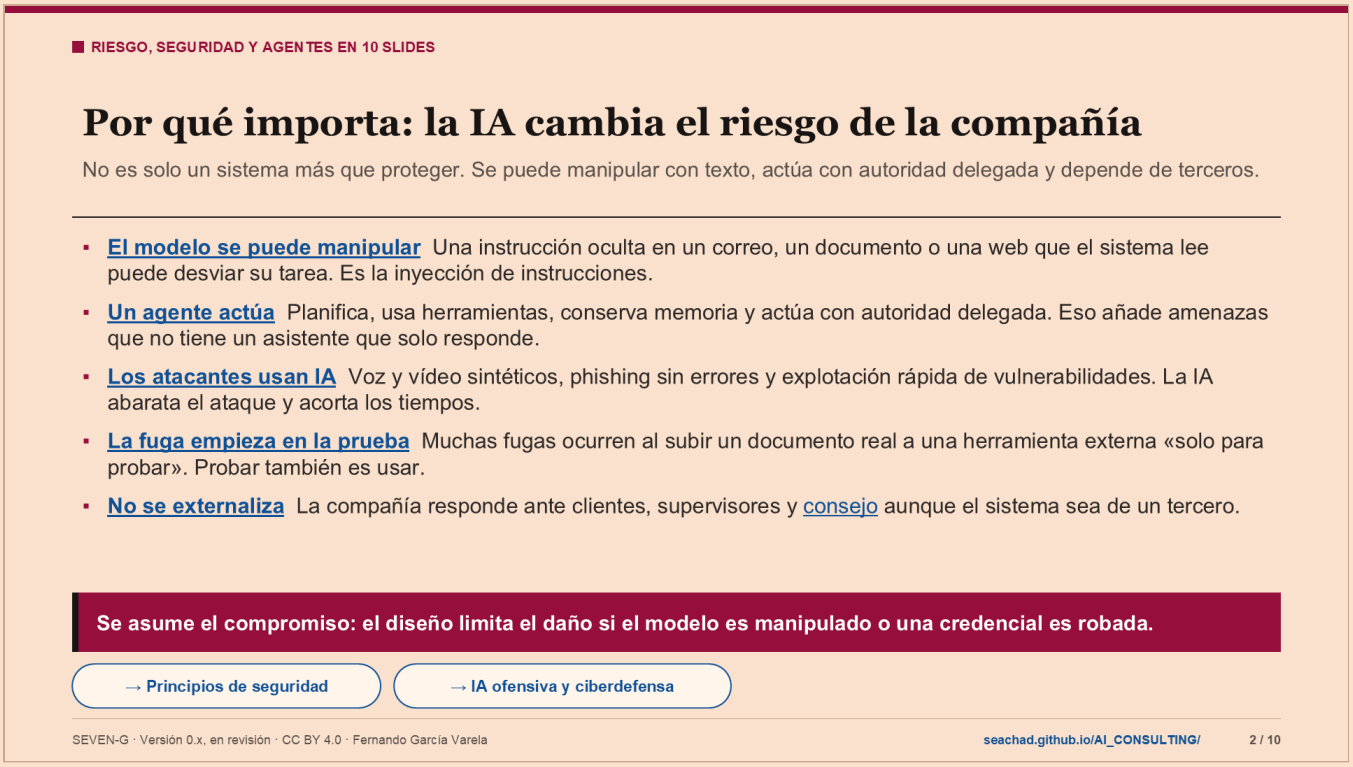
- **Áreas de actuación en IA**: Doce puertas de entrada al marco por temas: estrategia, casos de uso, valor, gobierno, organización, madurez, riesgo, seguridad, normas, consejo, datos y operación, e implantación. Cada área

responde a una pregunta y lleva a sus documentos, plantillas, herramientas y a una página propia. Sirve para llegar a un asunto sin recorrer todo el método. El enlace abre las áreas en la entrada de SEVEN-G. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/entrada/index.html#areas

- **SEVEN-G**: SEVEN-G (Seven-phase Enterprise Value & Governance) es un marco abierto para que una compañía implante, gobierne y mida la IA, y sepa si se transforma o solo gana eficiencia. Trata cada caso de uso como una inversión con puertas de decisión y da al consejo cifras que puede creerse. Lo usan la dirección, la oficina de IA, riesgos y el consejo. El enlace abre su página de entrada, con la historia, las preguntas y los ejemplos. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/entrada/index.html

FUENTES Y ENLACES

- **Área: riesgo, seguridad y agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/areas/seguridad.html
- **Documento 35 · Seguridad de IA y agentes: objeto y alcance** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#objeto-y-alcance
- **Documento 36 · Terceros y proveedores de IA** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html
- **Seguridad de la IA y de los agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html



Notas

La IA no es un sistema más que proteger con las reglas de siempre. Cambia el riesgo de la compañía por cinco vías.

- El modelo se puede manipular con texto. La inyección de instrucciones introduce órdenes que alteran el comportamiento esperado. Es directa si la escribe el usuario e indirecta si va oculta en lo que el sistema lee: documentos, correos, webs o respuestas de herramientas.
- Un agente actúa. Aparecen amenazas nuevas: que persiga un objetivo distinto del encomendado, que use más permisos de los que necesita su tarea o que induzca a una persona a aprobar algo inadecuado.
- Los atacantes también usan IA. La mayoría de los ataques no son nuevos, pero la IA reduce su coste, los hace más verosímiles y acorta los tiempos. Un ejemplo es el fraude del CEO con una videollamada falsa que ordena un pago urgente.
- La fuga empieza en la prueba. Por eso SEVEN-G dice que probar también es usar: ninguna herramienta externa se usa con datos internos, personales o confidenciales sin evaluar antes al proveedor.
- La responsabilidad no se externaliza. Aunque el modelo sea de un tercero, la compañía responde.

La respuesta de SEVEN-G empieza por ocho principios. El más importante para un directivo: el modelo no es una frontera de seguridad. Los límites que importan (permisos, importes, destinatarios y datos) se aplican fuera del modelo. Y se asume el compromiso: se diseña pensando que algún día el modelo será manipulado o una credencial será robada.

Por qué importa. Si la protección depende de que el modelo obedezca, basta un texto bien escrito para saltarla.

Pregunta para el aula: ¿alguna orden de pago de su organización depende solo de reconocer una voz o una cara?

TÉRMINOS DE ESTA SLIDE

- **Consejo de administración**: Máximo órgano de la compañía. En SEVEN-G aprueba la tesis de IA y el apetito de riesgo (C2), las apuestas de Transformar y su escalado, y supervisa la IA cada trimestre. La IA es un activo estratégico: sus grandes decisiones no son solo de tecnología. El enlace abre su papel en el modelo de gobierno (documento 30). https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/30_SEVEN-G_Modelo_de_gobierno.html#3-1-consejo-de-administracion

FUENTES Y ENLACES

- **Documento 35 §2 · Principios de seguridad** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#principios-de-seguridad
- **Documento 35 §3 · Amenazas a los sistemas de IA** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#amenazas-a-los-sistemas-de-ia
- **Documento 36 §2 · Principios (probar también es usar)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html#principios
- **OWASP Top 10 para aplicaciones con modelos de lenguaje (2025)** https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/
- **CCN-CERT BP/36 · Buenas prácticas frente a la IA ofensiva** https://www.ccn-cert.cni.es/es/informes/informes-de-buenas-practicas-bp/7469-ccn-cert-bp-36-buenas-practicas-ia-ofensiva/file.html
- **Principios de seguridad** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#principios-de-seguridad
- **IA ofensiva y ciberdefensa** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#exposicion-a-la-ia-ofensiva-y-uso-de-la-ia-en-la-ciberdefensa



Notas

Seis conceptos bastan para seguir una conversación sobre seguridad de la IA. Todos están definidos en el glosario de SEVEN-G (documento 02).

- Agente y nivel de autonomía. Un agente es un sistema de IA que ejecuta acciones sobre otros sistemas, datos o personas, o las prepara para validación humana, mediante herramientas y permisos. El nivel de autonomía mide cuánto actúa sin validación humana previa de cada acción. Se asigna el nivel más alto de las acciones que el sistema puede ejecutar, no el de su uso habitual. Un agente que casi nunca paga, pero puede pagar, se clasifica como si pagara.
- Inyección de instrucciones. Es el ataque típico de la IA generativa. Sus controles se demuestran con pruebas adversarias, no con declaraciones.
- Mínimo privilegio. El agente tiene solo lo imprescindible para su tarea, y los permisos se justifican en el diseño de seguridad (P18).
- Interruptor de parada. Al activarse revoca credenciales, bloquea nuevas acciones, conserva los registros y activa el proceso alternativo. Lo activan el responsable de operación de IA y seguridad.
- Pasarela corporativa de IA (control SEG-26). Aplica la misma configuración medida a todos los casos de uso. Falla cerrada en acciones sensibles y degradada en consultas.
- Banco de pruebas corporativo de agentes (control SEG-27). Sus escenarios salen del modelo de amenazas y de los incidentes propios, y cubren todos los canales: texto, documentos, correo, web, memoria, mensajes entre agentes y voz sintética.

Un séptimo concepto completa el área: el proveedor de IA, el tercero que suministra modelos, plataformas o software con IA. Se gestiona con tres niveles de exigencia, de N1 a N3 (documento 36).

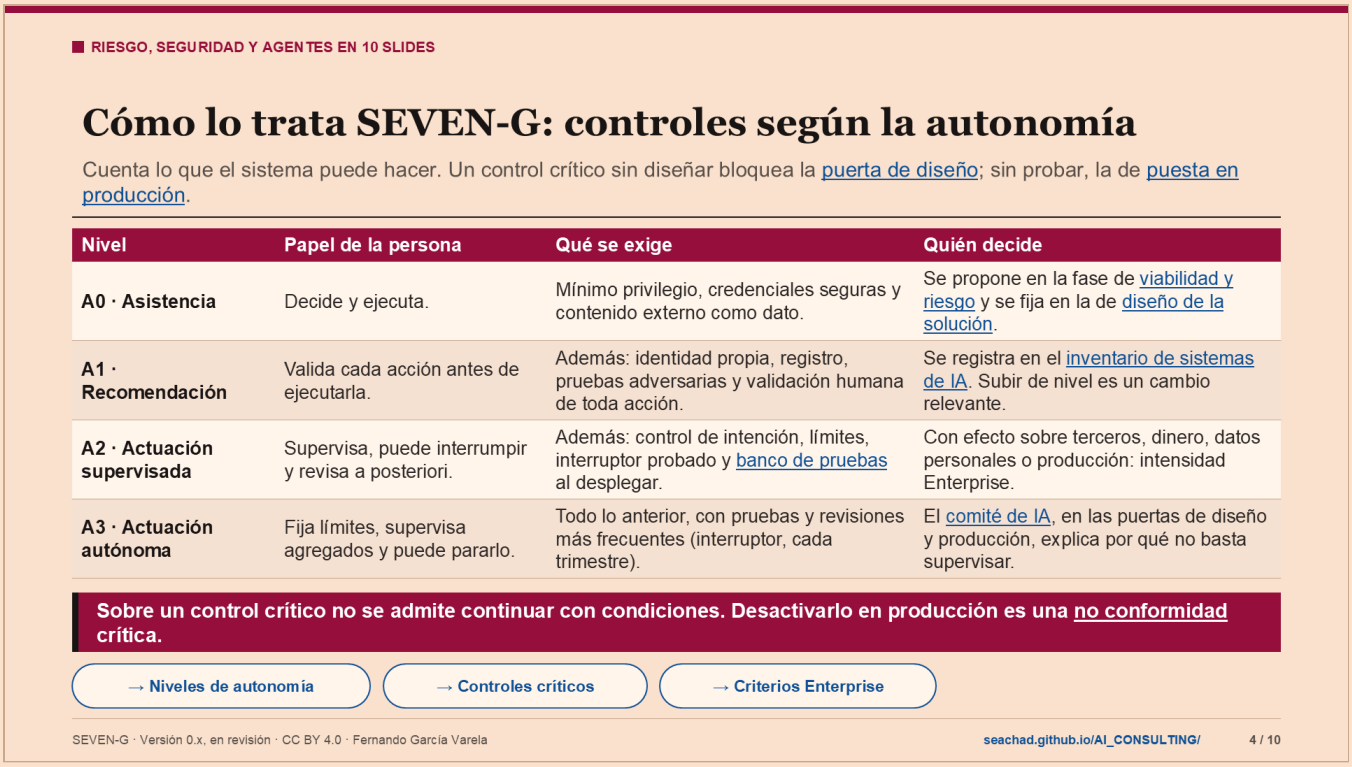
Por qué importa. Sin un vocabulario común, «el agente es seguro» no significa nada. Con él, la dirección puede preguntar: ¿qué nivel tiene?, ¿quién puede pararlo?, ¿pasa por la pasarela?

TÉRMINOS DE ESTA SLIDE

- **Glosario y convenciones** (documento 02): El glosario oficial de SEVEN-G: cada término con su definición, su equivalente en inglés y el documento que lo desarrolla, más las escalas y los códigos que se usan en todo el marco. Sirve para que todos llamen igual a lo mismo. El enlace abre el documento. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/02_SEVEN-G_Glosario_y_convenciones.html
- **SEVEN-G**: SEVEN-G (Seven-phase Enterprise Value & Governance) es un marco abierto para que una compañía implante, gobierne y mida la IA, y sepa si se transforma o solo gana eficiencia. Trata cada caso de uso como una inversión con puertas de decisión y da al consejo cifras que puede creerse. Lo usan la dirección, la oficina de IA, riesgos y el consejo. El enlace abre su página de entrada, con la historia, las preguntas y los ejemplos. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/entrada/index.html
- **Niveles de autonomía** (A0–A3): Escala de cuánto actúa un sistema de IA por su cuenta. A0: solo informa; A1: propone y una persona valida cada acción; A2: actúa dentro de límites y una persona supervisa; A3: actúa solo en un ámbito definido. Cuanta más autonomía, más controles: un agente A2 o A3 con efecto sobre terceros es Enterprise. El enlace abre la escala en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia
- **Pruebas adversarias** (P53): Pruebas en que un equipo intenta engañar o romper un sistema de IA como lo haría un atacante: inyección de instrucciones, fuga de datos, abuso de herramientas de un agente. Un escenario crítico superado por el atacante impide desplegar. Se planifican e informan con la plantilla P53. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P53_SEVEN-G_Plan_e_informe_de_pruebas_adversarias.html

FUENTES Y ENLACES

- **Glosario (documento 02) · Autonomía de sistemas y agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/02_SEVEN-G_Glosario_y_convenciones.html#4-11-autonomia-de-sistemas-y-agentes
- **Documento 35 §4 · Requisitos esenciales de un agente** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#requisitos-esenciales-de-un-agente
- **Documento 35 §6 · Controles de seguridad de IA (SEG)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#catalogo-de-controles-de-seguridad-de-ia-seg
- **Documento 36 §4 · Niveles de exigencia N1–N3** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html#niveles-de-exigencia-n1-n3
- **Requisitos de un agente** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#requisitos-esenciales-de-un-agente
- **Página del área** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/areas/seguridad.html



RIESGO, SEGURIDAD Y AGENTES EN 10 SLIDES

Cómo lo trata SEVEN-G: controles según la autonomía

Cuenta lo que el sistema puede hacer. Un control crítico sin diseñar bloquea la puerta de diseño; sin probar, la de puesta en producción.

Nivel	Papel de la persona	Qué se exige	Quién decide
A0 · Asistencia	Decide y ejecuta.	Mínimo privilegio, credenciales seguras y contenido externo como dato.	Se propone en la fase de viabilidad y riesgo y se fija en la de diseño de la solución.
A1 · Recomendación	Valida cada acción antes de ejecutarla.	Además: identidad propia, registro, pruebas adversarias y validación humana de toda acción.	Se registra en el inventario de sistemas de IA. Subir de nivel es un cambio relevante.
A2 · Actuación supervisada	Supervisa, puede interrumpir y revisa a posteriori.	Además: control de intención, límites, interruptor probado y banco de pruebas al desplegar.	Con efecto sobre terceros, dinero, datos personales o producción: intensidad Enterprise.
A3 · Actuación autónoma	Fija límites, supervisa agregados y puede pararlo.	Todo lo anterior, con pruebas y revisiones más frecuentes (interruptor, cada trimestre).	El comité de IA, en las puertas de diseño y producción, explica por qué no basta supervisar.

Sobre un control crítico no se admite continuar con condiciones. Desactivarlo en producción es una no conformidad crítica.

→ Niveles de autonomía · → Controles críticos · → Criterios Enterprise

SEVEN-G · Versión 0.x, en revisión · CC BY 4.0 · Fernando García Varela seachad.github.io/AI_CONSULTING/ 4 / 10

Notas

La tabla resume la regla central del área: cuanta más autonomía, más controles. Cada fila suma exigencias a la anterior.

- A0 · Asistencia. El sistema informa, resume o genera contenido; la persona decide y ejecuta. Son obligatorios el mínimo privilegio, la gestión de credenciales y tratar todo contenido externo como dato. Si accede a sistemas corporativos con credenciales propias, también la identidad propia y las herramientas aprobadas.
- A1 · Recomendación. El sistema propone una acción concreta y la persona valida cada una antes de ejecutarla. Se añaden la identidad propia, la trazabilidad entre acción e intención, el registro de acciones y las pruebas adversarias.
- A2 · Actuación supervisada. El sistema ejecuta acciones dentro de límites definidos. Se exigen todos los controles de agentes que apliquen: un punto de decisión fuera del modelo que compruebe cada acción, límites de importe y volumen, interruptor de parada probado cada semestre y banco de pruebas en cada despliegue.
- A3 · Actuación autónoma. El sistema ejecuta secuencias sin revisión individual. El interruptor se prueba cada trimestre y los permisos se revisan cada trimestre.

Quién decide y cuándo. El nivel se propone en la fase 3 (viabilidad), se fija en la fase 4 (diseño) y se registra en el inventario. Un sistema A2 o A3 que afecta a terceros, dinero, datos personales o producción cumple el criterio de «agentes con capacidad de actuar» y va con intensidad Enterprise. Un sistema A3 necesita la decisión del comité de IA en las puertas de diseño (G4) y de puesta en producción (G5). Tras un incidente grave relacionado con la autonomía, el sistema baja de nivel o se detiene hasta el análisis.

Por qué importa. Esta regla convierte «confiar en la IA» en una decisión concreta que se aprueba, se supervisa y se audita.

Pregunta para el aula: ¿por qué SEVEN-G clasifica por lo que el sistema puede hacer y no por lo que suele hacer?

TÉRMINOS DE ESTA SLIDE

- **Puerta de diseño** (G4): Puerta que aprueba el diseño completo antes de construir: arquitectura, datos, seguridad, supervisión humana, reversión y adopción. Comprueba que los controles están pensados desde el principio. El enlace abre sus criterios en el documento 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-5-g4-diseno
- **Puesta en producción** (G5): Puerta que autoriza a usar el sistema de IA con usuarios, clientes o datos reales. Exige la validación y el piloto superados, el plan de reversión probado y la firma de varios niveles (negocio, tecnología, riesgos). Nada entra en producción sin esta firma. El enlace abre sus criterios en el documento 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-6-g5-puesta-en-produccion
- **Viabilidad y riesgo** (fase 3): Fase en la que se comprueba con una prueba acotada si el caso es viable técnica y económicamente, se clasifica según la regulación y se registran sus riesgos. Termina en la puerta de viabilidad (G3), la principal puerta de parada antes del gran gasto. Es donde más dinero se ahorra parando a tiempo. El enlace abre el manual de la fase en el documento 20. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-3-viabilidad-y-riesgo
- **Diseño de la solución** (fase 4): Fase en la que se diseña la solución completa: arquitectura, datos, seguridad, supervisión humana, plan de reversión y plan de adopción. Termina en la puerta de diseño (G4). Sirve para que los controles se diseñen antes de construir y no se añadan después, cuando cuestan más. El enlace abre el manual de la fase en el documento 20. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-4-diseno-de-la-solucion
- **Inventario de sistemas de IA** (T02): Lista de todos los sistemas de IA de la compañía: los propios, la IA de terceros integrada en procesos, los asistentes que usan los empleados y el uso no autorizado detectado. Cada sistema tiene responsable, clasificación regulatoria, intensidad y autonomía. Es lo primero que se pide, porque no se puede gobernar lo que no se conoce. Es una vista del registro de casos de uso: al abrir el enlace, elija «Inventario» (datos ficticios). https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T01_registro_iniciativas/registro.html
- **Banco de pruebas de agentes** (SEG-27): Biblioteca común y versionada de ataques simulados, sacada del modelo de amenazas y de los incidentes propios, que prueba agentes y asistentes por todos sus canales, herramientas y conectores. En los agentes que actúan se ejecuta en cada despliegue, y si un escenario crítico tiene éxito no se despliega. El enlace abre el control en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#cod-seg-27
- **Comité de IA**: Órgano de dirección que decide la cartera y las puertas de su nivel, asigna presupuesto y eleva al consejo lo que le corresponde. Reúne a negocio, tecnología, finanzas y riesgos, y se rige por un reglamento (P38). Importa porque convierte decisiones dispersas en una sola cartera. El enlace abre su papel en el modelo de gobierno (documento 30). https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/30_SEVEN-G_Modelo_de_gobierno.html#3-3-comite-de-ia
- **No conformidad** (P50): Incumplimiento de una regla del marco o de un control, aunque no haya causado daño (por ejemplo, un caso en producción sin firma). Se clasifica en menor, mayor o crítica, se registra (P50), se analiza su causa (P52) y se corrige en plazo. Sirve para que el gobierno aprenda de sus fallos. El enlace abre el proceso en el documento 37. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/37_SEVEN-G_No_conformidades_e_incidentes.html#proceso-de-no-conformidades

FUENTES Y ENLACES

- **Documento 35 §5 · Niveles de autonomía** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia
- **Documento 35 §5.3 · Controles mínimos por nivel** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#5-3-controles-minimos-por-nivel
- **Documento 35 §7 · Controles de agentes (AG)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#catalogo-de-controles-de-agentes-ag

RIESGO, SEGURIDAD Y AGENTES EN 10 SLIDES

- **Documento 01 §9.2 · Criterios Enterprise** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/01_SEVEN-G_Metodologia_fundacional.html#9-2-criterios-enterprise
- **Niveles de autonomía** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia
- **Controles críticos** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#5-4-controles-criticos
- **Criterios Enterprise** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/01_SEVEN-G_Metodologia_fundacional.html#9-2-criterios-enterprise



Notas

En SEVEN-G cada caso de uso recorre fases separadas por puertas de decisión. La seguridad se comprueba en cinco de ellas, siempre con evidencia.

- Puerta de viabilidad (G3). Los riesgos de IA generativa y de seguridad se valoran en la matriz de riesgos (P12). Se propone el nivel de autonomía. Si hay proveedor, se evalúa su seguridad con el cuestionario de diligencia debida (P55).
- Puerta de diseño (G4). El diseño de seguridad (P18) recoge el modelo de amenazas, el nivel de autonomía, las identidades, la matriz de permisos, el control de intención, los límites, las acciones sensibles, el interruptor y el plan de pruebas. Las identidades de los agentes se inscriben, con su responsable humano, en el inventario de identidades no humanas (P54). El tráfico hacia modelos pasa por la pasarela corporativa.
- Puerta de puesta en producción (G5). Se hacen pruebas adversarias (red teaming) y se documentan en P53. Se prueba el interruptor. El banco de pruebas no puede tener escenarios críticos con éxito. Seguridad de la información firma dentro de la firma multinivel.
- Revisión de continuidad (R6). Se comprueba que los controles siguen operando y que las campañas y revisiones están en plazo. Los indicadores llegan al panel del consejo (T17).
- Puerta de escalado o retirada (G7). Al escalar, los controles se revisan para el nuevo alcance. Al retirar, no pueden quedar identidades ni credenciales huérfanas.

Quién hace qué. El responsable técnico de IA diseña e implanta los controles. Seguridad de la información los revisa. El responsable de riesgos de IA valora el riesgo residual. El auditor de IA verifica las evidencias con la lista LV-AG. La evaluación de seguridad de agentes (T10) se aplica con el diseño de seguridad (P18). El perfil de seguridad de IA (P72) da la visión de conjunto en el ciclo corporativo.

Por qué importa. Separar quien construye de quien prueba evita que nadie apruebe su propio trabajo.

TÉRMINOS DE ESTA SLIDE

- **Listas de verificación por puerta** (documento 22): Listas de comprobación prácticas para preparar y revisar cada puerta y algunos temas transversales (evidencias, agentes). Traducen los criterios del documento 21 en preguntas que se contestan sí o no antes de la sesión. El enlace abre el documento. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html
- **Matriz y registro de riesgos** (T06): Matriz de probabilidad por impacto (5 × 5) y registro de los riesgos de cada caso de uso: riesgo inherente y residual, controles y su eficacia, respuesta, contingencia y quién acepta el riesgo. Es obligatoria desde la fase 3. Se rellena con la plantilla P12 o en la vista «Riesgos» del registro de casos de uso, que además descarga un Excel del caso con planes de mitigación propuestos. Datos ficticios. https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T01_registro_iniciativas/registro.html
- **Diligencia debida a proveedores de IA** (P55): Cuestionario de doce bloques para evaluar a un proveedor de IA antes de contratarlo: seguridad, datos, modelos, continuidad, cumplimiento, subcontratistas y salida. Un resultado no conforme en los bloques críticos impide usarlo con datos de producción. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P55_SEVEN-G_Cuestionario_de_diligencia_debida_a_proveedores_de_IA.html
- **Diseño de seguridad** (P18): Plantilla para diseñar la seguridad de la solución antes de construirla, agentes incluidos: amenazas, controles, identidades, permisos, registros y pruebas previstas. Se revisa en la puerta de diseño. Añadir seguridad después cuesta más y protege menos. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P18_SEVEN-G_Diseno_de_seguridad.html
- **Inventario de identidades no humanas y componentes** (P54): Plantilla que lista las identidades que usan los agentes y servicios de IA (cuentas, claves, permisos) y los componentes de software de cada solución. Un agente con permisos de más es una puerta abierta; esta lista permite revisarlos y retirarlos. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P54_SEVEN-G_Inventario_de_identidades_no_humanas_y_componentes.html
- **Pruebas adversarias** (P53): Pruebas en que un equipo intenta engañar o romper un sistema de IA como lo haría un atacante: inyección de instrucciones, fuga de datos, abuso de herramientas de un agente. Un escenario crítico superado por el atacante impide desplegar. Se planifican e informan con la plantilla P53. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P53_SEVEN-G_Plan_e_informe_de_pruebas_adversarias.html
- **Panel del consejo** (T17): Panel de supervisión que se genera desde el registro de casos de uso. Muestra al consejo coste, valor validado frente a declarado, embudo, riesgos, incidentes, adopción, madurez y qué frena el escalado, con una ayuda «?» en cada tarjeta. Sirve para decidir, no solo para informar. El enlace abre el panel de ejemplo con datos ficticios; la interfaz está en español y las ayudas en los dos idiomas. https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T17_panel_consejo/ejemplo/salida/t01_Dashboard_Casos_Uso_IA_v8.html

FUENTES Y ENLACES

- **Documento 35 §11 · Integración con el ciclo de vida** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#integracion-con-el-ciclo-de-vida
- **Documento 35 §12 · Herramientas y plantillas asociadas** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#herramientas-y-plantillas-asociadas
- **Plantilla P72 · Perfil de seguridad de IA** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P72_SEVEN-G_Perfil_de_seguridad_de_IA_CSF_y_Cyber_AI_Profile.html
- **Panel del consejo (T17)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T17_panel_consejo/index.html
- **Lista de verificación de agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-ag-agentes



Notas

El caso es ficticio, como todos los del documento 92 (casos de aplicación): compañías, personas y cifras son ilustrativas. Un operador logístico quiere abrir un servicio nuevo: contratar capacidad de transporte bajo demanda para sus clientes con un agente que compra cargas en bolsas de transporte. Es una apuesta de Transformar, con intensidad Enterprise.

Cómo se aplica el área, paso a paso:

- Apetito de riesgo. En la etapa de dirección (C2) el consejo fija hasta dónde llega: agentes con actuación supervisada (A2), con un límite económico por operación aprobado por el comité, y ningún A3 en operaciones con dinero de terceros. La seguridad empieza en el consejo, antes de que exista el agente.
- Nivel de autonomía. El agente actúa solo hasta 2.500 € por carga. Por encima, decide una persona. Es un límite de actuación aplicado fuera del modelo, no una instrucción que el modelo pueda olvidar.
- Puerta de diseño (G4). El comité de IA comprueba que el límite por carga, el registro de acciones y el interruptor de parada están diseñados.
- Puerta de puesta en producción (G5). Un piloto con asignación aleatoria (520 cargas con el agente y 510 de control) da una mejora del margen de 0,6 puntos, por debajo del 1,0 fijado como criterio de parada. Errores: 0,8 %. El comité para el caso de uso. El equipo pidió otra iteración con un umbral más bajo y el consejo decidió no relajar los criterios.

Qué enseña. Los controles de seguridad no hacen bueno un caso: permiten que funcione sin daños si su valor se demuestra. Aquí el valor no se demostró, y la compañía paró con 165.000 € invertidos y un aprendizaje registrado. La imagen muestra el ciclo de vida, con sus fases y sus puertas.

Pregunta para el aula: si el piloto hubiera alcanzado el margen, ¿qué tendría que estar probado antes de poner el agente en producción?

TÉRMINOS DE ESTA SLIDE

- **Consejo de administración**: Máximo órgano de la compañía. En SEVEN-G aprueba la tesis de IA y el apetito de riesgo (C2), las apuestas de Transformar y su escalado, y supervisa la IA cada trimestre. La IA es un activo estratégico: sus grandes decisiones no son solo de tecnología. El enlace abre su papel en el modelo de gobierno (documento 30). https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/30_SEVEN-G_Modelo_de_gobierno.html#3-1-consejo-de-administracion
- **Puerta de diseño** (G4): Puerta que aprueba el diseño completo antes de construir: arquitectura, datos, seguridad, supervisión humana, reversión y adopción. Comprueba que los controles están pensados desde el principio. El enlace abre sus criterios en el documento 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-5-g4-diseno
- **Comité de IA**: Órgano de dirección que decide la cartera y las puertas de su nivel, asigna presupuesto y eleva al consejo lo que le corresponde. Reúne a negocio, tecnología, finanzas y riesgos, y se rige por un reglamento (P38). Importa porque convierte decisiones dispersas en una sola cartera. El enlace abre su papel en el modelo de gobierno (documento 30). https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/30_SEVEN-G_Modelo_de_gobierno.html#3-3-comite-de-ia

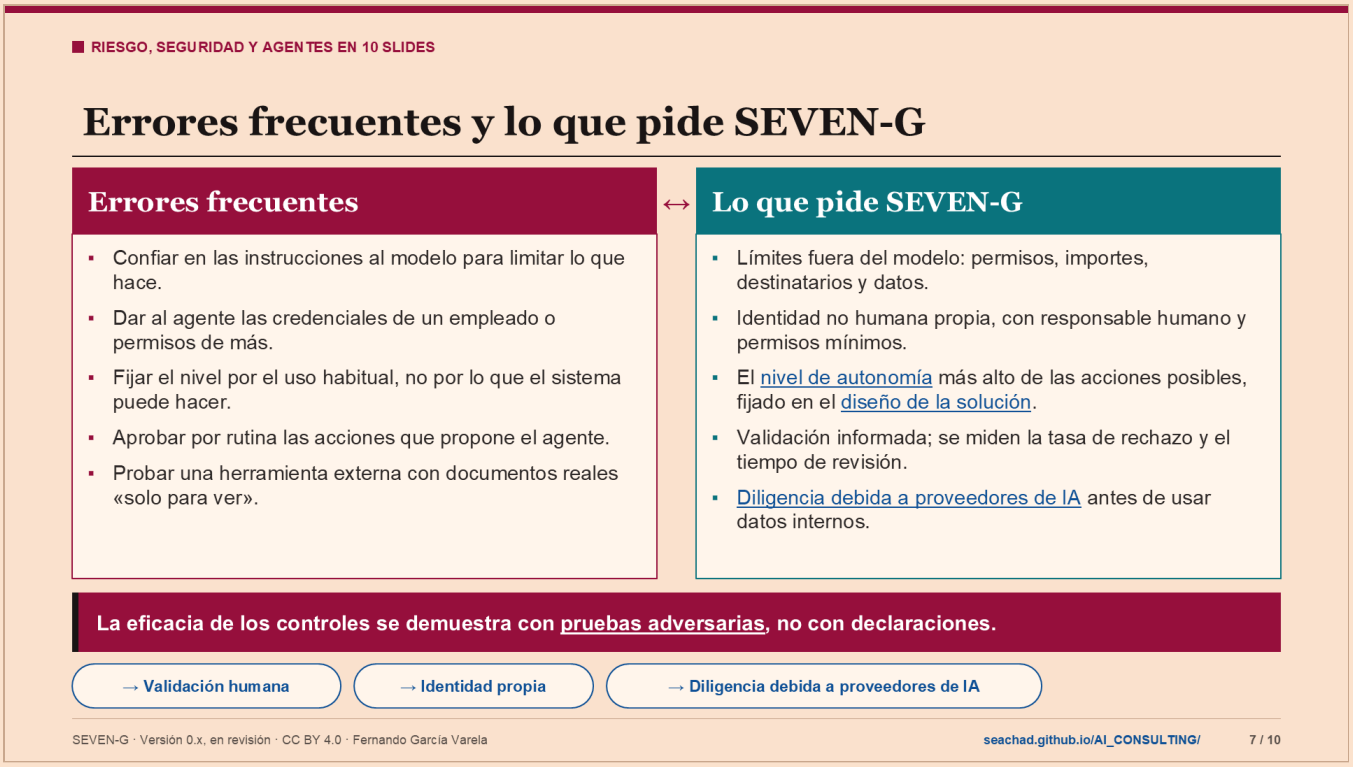
ENLACES DEL GRÁFICO

- Propuesto · fases 0 y 1: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/00_SEVEN-G_Que_es_y_para_que_sirve.html#4-3-nivel-caso-de-uso-ciclo-de-vida-con-puertas-de-decision
- Hipótesis de valor · fase 2: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/00_SEVEN-G_Que_es_y_para_que_sirve.html#4-3-nivel-caso-de-uso-ciclo-de-vida-con-puertas-de-decision
- Prueba de concepto · fase 3: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/00_SEVEN-G_Que_es_y_para_que_sirve.html#4-3-nivel-caso-de-uso-ciclo-de-vida-con-puertas-de-decision
- En desarrollo · fases 4 y 5: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/00_SEVEN-G_Que_es_y_para_que_sirve.html#4-3-nivel-caso-de-uso-ciclo-de-vida-con-puertas-de-decision
- En uso · fases 6 y 7 · o desenganchado: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/00_SEVEN-G_Que_es_y_para_que_sirve.html#4-3-nivel-caso-de-uso-ciclo-de-vida-con-puertas-de-decision
- Fase 0 · Contexto y restricciones: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-0-contexto-y-restricciones
- Fase 1 · Descubrimiento de oportunidades: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-1-descubrimiento-de-oportunidades
- Fase 2 · Hipótesis de valor: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-2-hipotesis-de-valor
- Fase 3 · Viabilidad y riesgo: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-3-viabilidad-y-riesgo
- Fase 4 · Diseño de la solución: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-4-diseno-de-la-solucion
- Fase 5 · Entrega y validación: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-5-entrega-y-validacion
- Fase 6 · Operación y gobierno: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-6-operacion-y-gobierno
- Fase 7 · Evolución o retirada: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-7-evolucion-o-retirada
- G0 · Autorización: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-1-g0-autorizacion
- Lista LV-G0: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-g0-autorizacion
- G1 · Oportunidad: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-2-g1-oportunidad

- Lista LV-G1: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-g1-oportunidad
- G2 · Hipótesis: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-3-g2-hipotesis
- Lista LV-G2: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-g2-hipotesis
- G3 · Viabilidad: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-4-g3-viabilidad
- Lista LV-G3: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-g3-viabilidad
- G4 · Diseño: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-5-g4-diseno
- Lista LV-G4: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-g4-diseno
- G5 · Puesta en producción: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-6-g5-puesta-en-produccion
- Lista LV-G5: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-g5-puesta-en-produccion
- Firma multinivel: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#firma-multinivel-de-puesta-en-produccion-g5
- R6 · Continuidad: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-7-r6-revision-de-continuidad
- Lista LV-R6: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-r6-revision-de-continuidad
- G7 · Escalado o retirada: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-8-g7-escalado-o-retirada
- Lista LV-G7: https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/22_SEVEN-G_Listas_de_verificacion_por_gate.html#lv-g7-escalado-o-retirada

FUENTES Y ENLACES

- **Documento 92 · Caso 2, operador logístico L (ficticio)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/92_SEVEN-G_Casos_de_aplicacion.html#caso-2-operador-logistico-l-ficticio
- **Documento 35 §4.5 · Límites de actuación** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#4-5-limites-de-actuacion
- **Documento 35 §4.7 · Interruptor de parada** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#4-7-interruptor-de-parada
- **Caso 2 · Operador logístico** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/92_SEVEN-G_Casos_de_aplicacion.html#caso-2-operador-logistico-l-ficticio
- **Límites de actuación** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#4-5-limites-de-actuacion



Notas

Cinco errores frecuentes, cada uno con la regla de SEVEN-G que lo evita.

- Confiar en las instrucciones al modelo. Decir al modelo «no hagas pagos de más de mil euros» ayuda, pero un atacante puede saltarlo con otra instrucción. SEVEN-G aplica los límites fuera del modelo, con un punto de decisión independiente que bloquea lo que no encaja.
- Credenciales de un empleado o permisos de más. Si el agente usa la cuenta de una persona, nadie sabe quién hizo qué. SEVEN-G exige una identidad propia, registrada con su responsable humano, y una lista cerrada de permisos. Cuando actúa por cuenta de un usuario, nunca tiene más permisos que él.
- Fijar el nivel por el uso habitual. Un sistema que casi siempre propone, pero puede ejecutar, es de actuación. Se clasifica por lo que puede hacer.
- Aprobar por rutina. Si una persona aprueba cien acciones al día sin leerlas, la validación humana no protege. SEVEN-G pide una validación informada: la persona ve qué acción se propone, con qué datos, por qué y con qué efectos, y puede rechazarla sin coste para ella. Se miden la tasa de rechazo y el tiempo de revisión para detectar la aprobación rutinaria.
- Probar con documentos reales. Ninguna herramienta externa se usa con datos internos, personales o confidenciales sin la evaluación del proveedor proporcionada a su nivel. Una prueba con datos sintéticos o públicos no la necesita.

Por qué importa. Todos estos errores tienen algo en común: dan una sensación de control que nadie ha comprobado. SEVEN-G pide demostrarlo con pruebas adversarias.

Pregunta para el aula: ¿cuál de estos cinco errores cree que es más frecuente en su organización?

TÉRMINOS DE ESTA SLIDE

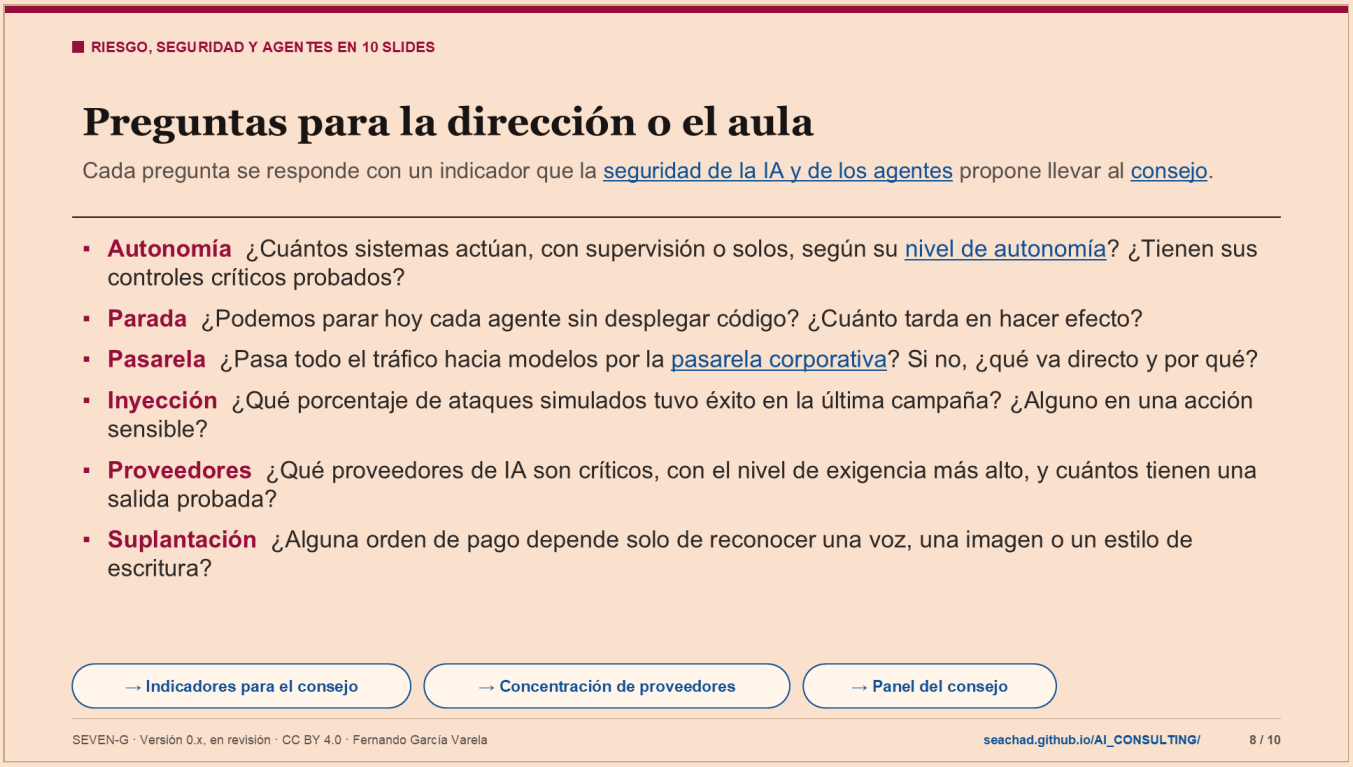
- **Niveles de autonomía** (A0–A3): Escala de cuánto actúa un sistema de IA por su cuenta. A0: solo informa; A1: propone y una persona valida cada acción; A2: actúa dentro de límites y una persona supervisa; A3: actúa solo en un ámbito definido. Cuanta más autonomía, más controles: un agente A2 o A3 con efecto sobre terceros es

Enterprise. El enlace abre la escala en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia

- **Diseño de la solución** (fase 4): Fase en la que se diseña la solución completa: arquitectura, datos, seguridad, supervisión humana, plan de reversión y plan de adopción. Termina en la puerta de diseño (G4). Sirve para que los controles se diseñen antes de construir y no se añadan después, cuando cuestan más. El enlace abre el manual de la fase en el documento 20. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/20_SEVEN-G_Manuales_de_fase.html#fase-4-diseno-de-la-solucion
- **Diligencia debida a proveedores de IA** (P55): Cuestionario de doce bloques para evaluar a un proveedor de IA antes de contratarlo: seguridad, datos, modelos, continuidad, cumplimiento, subcontratistas y salida. Un resultado no conforme en los bloques críticos impide usarlo con datos de producción. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P55_SEVEN-G_Cuestionario_de_diligencia_debida_a_proveedores_de_IA.html
- **Pruebas adversarias** (P53): Pruebas en que un equipo intenta engañar o romper un sistema de IA como lo haría un atacante: inyección de instrucciones, fuga de datos, abuso de herramientas de un agente. Un escenario crítico superado por el atacante impide desplegar. Se planifican e informan con la plantilla P53. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P53_SEVEN-G_Plan_e_informe_de_pruebas_adversarias.html

FUENTES Y ENLACES

- **Documento 35 §2 · Principios de seguridad** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#principios-de-seguridad
- **Documento 35 §4.4 · Control de intención** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#4-4-control-de-intencion
- **Documento 35 §4.6 · Validación humana en acciones sensibles** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#4-6-validacion-humana-en-acciones-sensibles
- **Documento 36 §2 · Principios (probar también es usar)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html#principios
- **Validación humana** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#4-6-validacion-humana-en-acciones-sensibles
- **Identidad propia** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#4-1-identidad-propia



Notas

Estas seis preguntas sirven para una sesión del consejo, un comité de dirección o un caso en el aula. Cada una tiene detrás un indicador del documento 35 (sección 10) o del documento 36 (sección 7), con su lectura esperada.

- Autonomía. Indicadores: sistemas por nivel de autonomía y controles críticos de agentes. Lectura: los sistemas A2 y A3 pueden crecer, pero con el 100 % de sus controles críticos probados y operando. Cualquier desviación se explica.
- Parada. Indicador: prueba del interruptor de parada. Lectura: el 100 % de los sistemas A2 y A3 con prueba vigente y tiempo de efecto medido.
- Pasarela. Indicador: tráfico por la pasarela corporativa. Lectura: 100 %; el tráfico directo se explica y se regulariza.
- Inyección. Indicador: éxito de la inyección de instrucciones en la última campaña de pruebas adversarias. Lectura: tendencia a la baja y cero en acciones sensibles.
- Proveedores. Indicadores de concentración: sistemas por proveedor, valor que depende de cada uno y proveedores críticos sin salida probada. Para los N3, la salida se prueba al menos una vez al año o antes de renovar.
- Suplantación. Medida de gobierno: ninguna orden de pago depende solo del reconocimiento de voz, imagen o estilo de escritura. Se hace al menos un simulacro anual de fraude con suplantación sintética que implique a finanzas y a dirección.

Cómo usarlas en el aula. Pida a los alumnos que elijan una compañía que conozcan y respondan con lo que saben. Lo normal es descubrir que varias respuestas son «no lo sabemos». Ese descubrimiento ya es el primer resultado.

Por qué importa. El consejo no necesita entender cómo funciona un modelo. Necesita hacer las preguntas correctas y exigir un dato detrás de cada respuesta. En SEVEN-G esos indicadores llegan al panel del consejo

(T17).

TÉRMINOS DE ESTA SLIDE

- **Seguridad de la IA y de los agentes** (documento 35): Las amenazas propias de la IA (inyección de instrucciones, fuga de datos, envenenamiento, abuso de agentes) y sus controles: niveles de autonomía, pasarela corporativa, banco de pruebas de agentes, identidades no humanas y uso de la IA en la ciberdefensa. El enlace abre el documento. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html
- **Consejo de administración**: Máximo órgano de la compañía. En SEVEN-G aprueba la tesis de IA y el apetito de riesgo (C2), las apuestas de Transformar y su escalado, y supervisa la IA cada trimestre. La IA es un activo estratégico: sus grandes decisiones no son solo de tecnología. El enlace abre su papel en el modelo de gobierno (documento 30). https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/30_SEVEN-G_Modelo_de_gobierno.html#3-1-consejo-de-administracion
- **Niveles de autonomía** (A0–A3): Escala de cuánto actúa un sistema de IA por su cuenta. A0: solo informa; A1: propone y una persona valida cada acción; A2: actúa dentro de límites y una persona supervisa; A3: actúa solo en un ámbito definido. Cuanta más autonomía, más controles: un agente A2 o A3 con efecto sobre terceros es Enterprise. El enlace abre la escala en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia
- **Pasarela corporativa de IA** (SEG-26): Punto de control común, fuera del modelo, por el que pasa todo el tráfico hacia modelos de IA, propios o de terceros: filtros de entrada, de salida y de contexto, prevención de fuga, cuotas, modelos aprobados y registro; para agentes, el control de cada acción frente a una intención autorizada. Tiene la misma configuración medida para todos los casos de uso. El enlace abre el control en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#cod-seg-26

FUENTES Y ENLACES

- **Documento 35 §10 · Indicadores para el consejo** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#indicadores-para-el-consejo
- **Documento 35 §9.2 · Medidas de gobierno frente a la IA ofensiva** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#9-2-medidas-de-gobierno
- **Documento 36 §7 · Dependencia tecnológica y concentración** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html#dependencia-tecnologica-y-concentracion
- **Indicadores para el consejo** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#indicadores-para-el-consejo
- **Concentración de proveedores** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html#7-2-concentracion
- **Panel del consejo** https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T17_panel_consejo/index.html



Notas

El área acoge dos subáreas, que conservan su propia página y sus propias slides: «Riesgo y regulación» y «Normas y estándares». Las dos primeras tarjetas son esas subáreas.

La seguridad de la IA no vive sola. Se apoya en otras áreas de SEVEN-G, en las reglas para construir la solución y en SPHERES.

- Riesgo y regulación. Las amenazas de seguridad tienen riesgos tipo propios (de IA generativa y de seguridad) que se valoran con la escala de probabilidad por impacto del documento 33. El Reglamento Europeo de IA exige a los sistemas de alto riesgo un nivel adecuado de precisión, solidez y ciberseguridad, incluida la resistencia a la manipulación de datos y de modelos.
- Normas y estándares. Cada control de seguridad de SEVEN-G indica la función del NIST CSF 2.0 a la que contribuye: gobernar, identificar, proteger, detectar, responder y recuperar. Con eso se construye el perfil de seguridad de la IA (P72). El perfil para la IA del NIST, el Cyber AI Profile, es todavía un borrador preliminar: se usa solo como orientación y nunca fundamenta un criterio de puerta.
- Datos, construcción y operación. El manual de operación (documento 52) opera la monitorización, el interruptor y las evaluaciones continuas, y mide la eficacia de los guardarraíles. Los incidentes de seguridad de IA siguen el circuito de no conformidades e incidentes (documento 37).
- Seguridad al construir. El documento 53 de SEVEN-G pide análisis de seguridad, de secretos y de dependencias sobre el código generado con IA antes de la puesta en producción (G5); en Enterprise, en cada cambio. Las dependencias que propone la IA se verifican antes de usarlas. Los agentes que construyen operan con mínimo privilegio y sin credenciales de producción, y su autonomía se clasifica con la escala A0–A3 del documento 35. Un hallazgo crítico o alto bloquea G5, salvo aceptación formal del riesgo.
- SPHERES. Es la capa explicativa de las esferas. En la esfera 07 (Decisión) usa la escala A0–A3 para decidir cuánto se delega en cada tipo de decisión. Recuerda que el nivel de autonomía no es lo mismo que el nivel de ambición.

Por qué importa. Un mismo vocabulario en SEVEN-G y en SPHERES evita que el consejo, el equipo de producto y el equipo de desarrollo hablen de cosas distintas.

TÉRMINOS DE ESTA SLIDE

- **Metodología de riesgos de IA** (documento 33): Cómo se identifican, valoran y aceptan los riesgos de la IA: escalas de probabilidad e impacto, matriz de 5 × 5, diez categorías, catálogo de riesgos tipo y quién puede aceptar cada nivel. El registro de riesgos es obligatorio desde la fase de viabilidad. El enlace abre la matriz en el documento 33. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/33_SEVEN-G_Metodologia_de_riesgos_de_IA.html#4-3-nivel-de-riesgo-y-matriz-5-5
- **Perfil de seguridad de IA (NIST CSF)** (P72): Plantilla para las compañías que usan el marco de ciberseguridad del NIST: perfil actual y objetivo de la seguridad de la IA, subcategoría a subcategoría, con la brecha y las acciones. El perfil de IA del NIST en que se apoya aún es un borrador. Es condicional: solo si la compañía usa ese marco. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P72_SEVEN-G_Perfil_de_seguridad_de_IA_CSF_y_Cyber_AI_Profile.html
- **Incidente de IA y su severidad** (S1–S4): Un hecho en que un sistema de IA causa o puede causar daño: error grave, sesgo, fuga de datos, acción indebida de un agente. Se clasifica de S1 (el más grave) a S4. Los S1 y S2 llegan al consejo por la vía rápida y pueden exigir notificar a autoridades (P51). El enlace abre la escala de severidad en el documento 37. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/37_SEVEN-G_No_conformidades_e_incidentes.html#4-2-severidad
- **Puesta en producción** (G5): Puerta que autoriza a usar el sistema de IA con usuarios, clientes o datos reales. Exige la validación y el piloto superados, el plan de reversión probado y la firma de varios niveles (negocio, tecnología, riesgos). Nada entra en producción sin esta firma. El enlace abre sus criterios en el documento 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-6-g5-puesta-en-produccion
- **SPHERES**: Metodología que explica dónde puede crear valor la IA en una compañía y con qué ambición, con tres preguntas: ¿dónde jugamos?, ¿con qué ambición? y ¿qué preguntamos? Es la capa explicativa de las esferas y los niveles de ambición; las reglas operativas están en SEVEN-G. Sirve para conversar con la dirección y el consejo. El enlace abre su documento de presentación. https://seachad.github.io/AI_CONSULTING/SPHERES/html/es/00_SPHERES_Que_es_y_para_que_sirve.html#que-es-spheres
- **Niveles de autonomía** (A0–A3): Escala de cuánto actúa un sistema de IA por su cuenta. A0: solo informa; A1: propone y una persona valida cada acción; A2: actúa dentro de límites y una persona supervisa; A3: actúa solo en un ámbito definido. Cuanta más autonomía, más controles: un agente A2 o A3 con efecto sobre terceros es Enterprise. El enlace abre la escala en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia

FUENTES Y ENLACES

- **Documento 34 §5.3 · NIST CSF 2.0 y Cyber AI Profile (borrador)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/34_SEVEN-G_Mapeo_regulatorio.html#5-3-nist-csf-2-0-y-cyber-ai-profile
- **Documento 53, sección 7.4 · Seguridad** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/53_SEVEN-G_Construccion_de_soluciones_con_IA.html#7-4-seguridad
- **SPHERES 03 · Los cuatro niveles de autonomía** https://seachad.github.io/AI_CONSULTING/SPHERES/html/es/03_SPHERES_Esferas_habilitadoras.html#los-cuatro-niveles-de-autonomia
- **Marco de ciberseguridad del NIST, versión 2.0** https://doi.org/10.6028/NIST.CSWP.29.spa
- **NIST IR 8596 · Cyber AI Profile (borrador preliminar)** https://csrc.nist.gov/pubs/ir/8596/iprd
- **Reglamento (UE) 2024/1689 (Reglamento Europeo de IA)** https://eur-lex.europa.eu/eli/reg/2024/1689/oj/spa
- **Construcción de soluciones con IA** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/53_SEVEN-G_Construccion_de_soluciones_con_IA.html
- **Qué es SPHERES** https://seachad.github.io/AI_CONSULTING/SPHERES/html/es/00_SPHERES_Que_es_y_para_que_sirve.html
- **Página del área** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/areas/seguridad.html



Notas

No hace falta montar todo a la vez. El recorrido de implantación de SEVEN-G (documento 96) ordena los hitos según el punto de partida de cada compañía. Para esta área, cuatro pasos marcan el camino.

- Saber qué IA se usa (hito HI-02). Lo hacen la oficina de IA y seguridad. Incluye el uso no autorizado: asistentes externos con datos de la compañía, agentes locales con credenciales en los equipos y extensiones del navegador con IA. Todo se inventaría y se autoriza como cualquier otra herramienta. Si hay constancia de uso no autorizado, o no se sabe, este hito pasa a la máxima prioridad.
- Política de uso aceptable (hito HI-03). Fija qué herramientas están aprobadas e incluye una formación básica. Aplica el principio de que probar también es usar.
- Nivel y diseño (fase 4 de cada caso de uso). Cada sistema recibe su nivel de autonomía y su diseño de seguridad (P18). Los controles críticos se diseñan antes de la puerta de diseño (G4).
- Agentes y proveedores (hito HI-16). Lo llevan seguridad, compras y riesgos: los requisitos de los agentes del documento 35 y la exigencia N1, N2 o N3 a cada proveedor según el documento 36. Debe estar listo antes de la puesta en producción (G5).

La herramienta del recorrido de implantación (T23) clasifica el punto de partida de la compañía con un cuestionario y muestra el recorrido priorizado. Funciona en el navegador con datos de ejemplo ficticios.

Por qué importa. Lo que no está inventariado no se puede proteger. Empezar por el inventario y la política es barato y da al consejo una primera respuesta honesta: qué IA usamos y quién la controla.

Pregunta para el aula: si tuviera que hacer mañana el inventario de la IA de su organización, ¿a quién preguntaría primero?

TÉRMINOS DE ESTA SLIDE

- **Hitos del recorrido de implantación** (HI-01–HI-22): Veintidós pasos para implantar SEVEN-G en cinco etapas, desde nombrar patrocinador e inventariar la IA hasta la revisión anual y la declaración. Cada hito dice qué hacer, quién lo hace, con qué plantilla o herramienta y qué pregunta de madurez lo acredita; su prioridad depende

del punto de partida. El enlace abre los hitos en el documento 96. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/96_SEVEN-G_Puntos_de_partida_y_recorrido_de_implantacion.html#los-hitos-del-recorrido

- **Puntos de partida** (PP-A–PP-F): Seis tipos de compañía según su punto de partida con la IA: punto cero, usuaria de IA de terceros, automatizadora con reglas o RPA, con IA propia en producción, con muchos pilotos y a escala. El tipo cambia el orden, el énfasis y el calendario de la implantación, nunca el destino ni las reglas. El enlace abre los seis tipos en el documento 96. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/96_SEVEN-G_Puntos_de_partida_y_recorrido_de_implantacion.html#2-1-los-seis-arquetipos
- **Inventario de sistemas de IA** (T02): Lista de todos los sistemas de IA de la compañía: los propios, la IA de terceros integrada en procesos, los asistentes que usan los empleados y el uso no autorizado detectado. Cada sistema tiene responsable, clasificación regulatoria, intensidad y autonomía. Es lo primero que se pide, porque no se puede gobernar lo que no se conoce. Es una vista del registro de casos de uso: al abrir el enlace, elija «Inventario» (datos ficticios). https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T01_registro_iniciativas/registro.html
- **Niveles de autonomía** (A0–A3): Escala de cuánto actúa un sistema de IA por su cuenta. A0: solo informa; A1: propone y una persona valida cada acción; A2: actúa dentro de límites y una persona supervisa; A3: actúa solo en un ámbito definido. Cuanta más autonomía, más controles: un agente A2 o A3 con efecto sobre terceros es Enterprise. El enlace abre la escala en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia
- **Diseño de seguridad** (P18): Plantilla para diseñar la seguridad de la solución antes de construirla, agentes incluidos: amenazas, controles, identidades, permisos, registros y pruebas previstas. Se revisa en la puerta de diseño. Añadir seguridad después cuesta más y protege menos. El enlace abre la plantilla. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/plantillas/P18_SEVEN-G_Diseno_de_seguridad.html
- **Puesta en producción** (G5): Puerta que autoriza a usar el sistema de IA con usuarios, clientes o datos reales. Exige la validación y el piloto superados, el plan de reversión probado y la firma de varios niveles (negocio, tecnología, riesgos). Nada entra en producción sin esta firma. El enlace abre sus criterios en el documento 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-6-g5-puesta-en-produccion

FUENTES Y ENLACES

- **Documento 31 §4.7 · Agentes personales y automatizaciones** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/31_SEVEN-G_Politica_corporativa_y_uso_aceptable.html#4-7-agentes-personales-y-automatizaciones
- **Documento 35 §6 · Control del uso corporativo de IA (SEG-20)** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#catalogo-de-controles-de-seguridad-de-ia-seg
- **Herramienta T23 · Recorrido de implantación** https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T23_recorrido_implantacion/recorrido.html
- **Recorrido de implantación** https://seachad.github.io/AI_CONSULTING/SEVEN-G/herramientas/T23_recorrido_implantacion/recorrido.html
- **Hitos del recorrido** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/96_SEVEN-G_Puntos_de_partida_y_recorrido_de_implantacion.html#4-1-que-es-cada-hito-quien-lo-hace-y-como-se-acredita
- **Página del área** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/areas/seguridad.html

AVISO LEGAL

Aviso legal: ayuda metodológica genérica y gratuita, «tal cual» y con fines informativos; no es asesoramiento jurídico ni está revisada jurídicamente para ningún caso. Cada organización la adapta y es responsable de su revisión legal y de su cumplimiento regulatorio; puede no recoger cambios normativos recientes. En la máxima medida permitida por la ley, el autor no asume responsabilidad alguna por los efectos de su aplicación. No otorga certificación de ningún tipo. Detalle en el documento 93.

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