HACKATHON DEFACC 2026 · FICHA TÉCNICA

GRIZZLY 360 · PIASG

Diagnóstico de seguridad de sistemas de IA a lo largo del Blueprint

Equipo: José José Molina · Andrea María Bernal | Track 5: Seguridad de infraestructura de IA y gobernanza defensiva (No-Code)

Demo: claude.ai/artifact/V2NfaRWZiWQWC92E1AxnFp · Contacto: molinajjose@gmail.com | andre2707_90@hotmail.com

El problema y la propuesta

Las organizaciones llevan sistemas de IA a producción sin demostrar que su protección corresponde a la capacidad de un atacante plausible. Las matrices tradicionales de probabilidad × impacto o los porcentajes de cumplimiento pueden ocultar esa brecha. GRIZZLY 360 · PIASG es una herramienta No-Code que determina el nivel de seguridad (SL) requerido a partir de la capacidad del atacante (OC1–OC5, RAND), contrasta requisito por requisito los controles y evidencias, identifica brechas por etapa y prioriza acciones. No reemplaza la decisión humana.

En cifras: 89 preguntas · 266 controles · 20 marcos de referencia · 13 reglas de brecha · repositorio de 99 hojas con fuentes versionadas.

Estructura

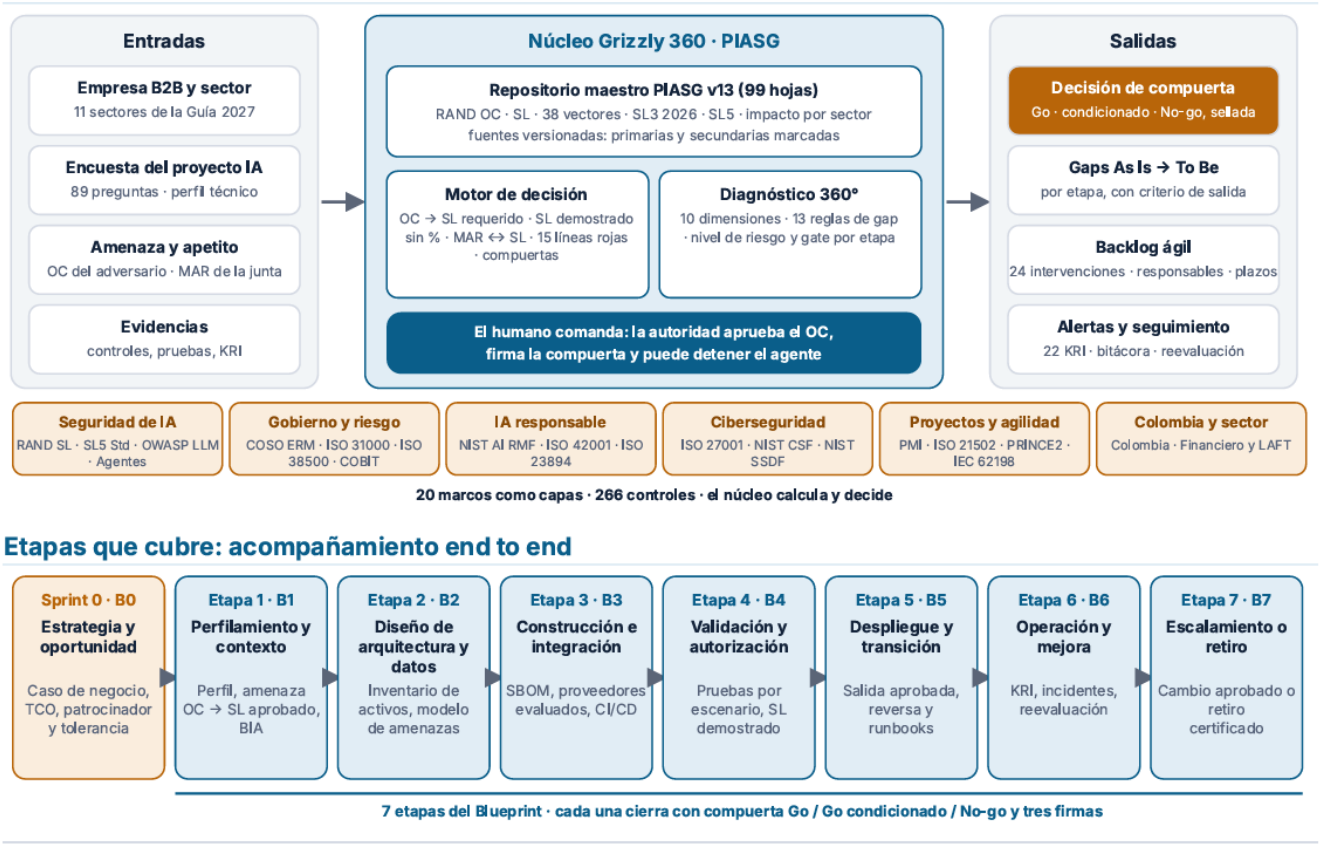


Figura 1. Recorrido de uso reconstruido a partir del flujo funcional descrito en la ficha: perfilamiento, aprobación del atacante, diagnóstico, priorización y decisión humana.

Etapas: B0 Estrategia y oportunidad · B1 Perfilamiento y contexto · B2 Diseño de arquitectura, datos y seguridad · B3 Construcción e integración · B4 Validación y autorización · B5 Despliegue y transición · B6 Operación y mejora · B7 Escalamiento, cambio o retiro.

Evaluación y resultados

Se construyeron 15 escenarios sintéticos sobre un agente de trámites ciudadanos (OC3 → SL3), cada uno con una perturbación controlada. La línea base fue una matriz de probabilidad × impacto aplicada al mismo registro y apetito de riesgo.

Métrica	Matriz clásica	GRIZZLY 360 · PIASG
Despliegues inseguros que pasan (de 10)	6	0
Decisiones correctas (de 15)	9	15

La herramienta bloqueó los 10 escenarios que requerían bloqueo y acertó las 15 decisiones esperadas. Se reportan 103 pruebas automáticas y 0,8 ms por decisión. En un escenario de riesgo mínimo con todo cumplido (E13), GRIZZLY fue más estricto de lo necesario.

Alcance y límites. Los resultados provienen de escenarios sintéticos sobre un solo caso; falta validación con una organización real. La herramienta verifica que la evidencia exista, no que sea cierta ni técnicamente eficaz. El reporte de brechas puede revelar superficies sensibles: acceso privado por defecto, sin técnicas de ataque. No es una certificación ni una garantía de seguridad.

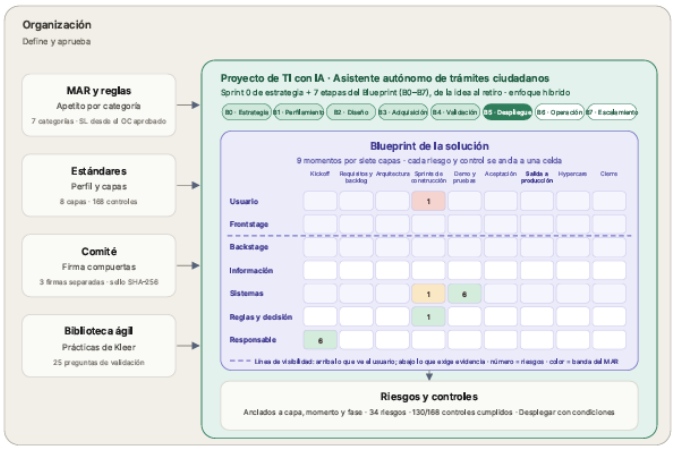
Referencias principales:

1. RAND Corporation (2024). *Securing AI Model Weights*, RR-A2849-1; (2026). *Achieving AI Model Weight Security Level 3*, RR-A4704-1.
2. SL5 Standard for AI Security, v0.1 (2026).
3. ISACA (2018). COBIT 2019; ISO/IEC 38500; ISO/IEC 27001:2022; NIST (2024). CSF 2.0.
4. Reglamento (UE) 2024/1689 de IA; NIST (2023). AI RMF 1.0; ISO/IEC 42001:2023; COSO (2017). ERM; ISO 31000:2018.
5. Grupo Cibest (2026). *Guía 2027*, taxonomía de sectores. https://media.ffycdn.net/us/grupo-bancolombia-arquitectura-y-contenidos/nzNhkuBXg6UVE6njeqzz.pdf

Cobertura del Blueprint

COBERTURA DEL MOTOR · 266 CONTROLES DISTRIBUIDOS POR ETAPA Y FAMILIA

	B0	B1	B2	B3	B4	B5	B6	B7
Seguridad de IA		1	11	31	7	10	1	
Seguridad de la información		3	9	13	3	7	4	1
Gobierno de TI		3	3	2	1	2	3	1
IA responsable	1	11	14	3	3	3	5	
Riesgo y apetito		7	4				3	1
Proyectos	1	6	7	5				3
Colombia y sector	1	4	4	3		2	6	1
Núcleo ERM	4	7	13	7	4	7	7	3
TOTAL	7	42	65	64	18	31	29	10



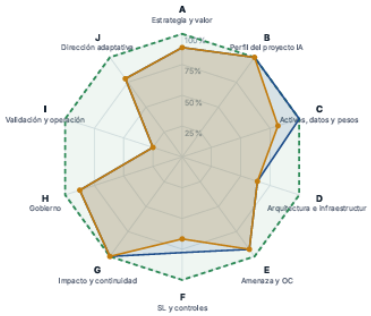
La junta aprueba el MAR y los estándares; el proyecto recorre el sprint 0 y las 7 etapas (ágil, predictivo o híbrido); el Blueprint cruza momentos con 7 capas y ancla cada riesgo en tres planos: IA, Solución y Marco.

Del atacante al nivel de seguridad (RAND)

- **OC1 · Aficionado** — 1 persona · días · ≤ USD 1.000 → **SL1**
- **OC2 · Profesional oportunista** — 1 persona · semanas · ≤ USD 10.000 → **SL2**
- **OC3 · Crimen organizado o insider** — ~10 personas · meses · ≤ USD 1 M → **SL3**
- **OC4 · Operación estándar de un Estado** — ~100 personas · 1 año · ≤ USD 10 M → **SL4**
- **OC5 · Prioridad de los Estados más capaces** — ~1.000 personas · años · ≤ USD 1.000 M → **SL5**

El OC lo aprueba la autoridad con evidencia. El sector y el impacto activan escenarios y apetito, pero no asignan el SL.

Diagnóstico 360° y tablero de seguimiento



Radar 360° del caso demo (Fintech Andina): **según las respuestas**, **hoy con lo cerrado** y **meta del plan**, en las diez dimensiones de la encuesta:

A · Estrategia y valor	B · Perfil del proyecto IA	C · Activos, datos y pesos	D · Arquitectura e infraestructura	E · Amenaza y capacidad OC
F · SL y controles	G · Impacto y continuidad	H · Gobierno y responsabilidad	I · Validación y operación	J · Dirección adaptativa

- **Tablero de actividades:** avance planeado frente a real, cronograma tipo Gantt por sprint, vencidas, cumplimiento del cronograma y alertas KRI configuradas.
- **Informe ejecutivo automático:** decisión, atacante y SL, dimensiones más débiles y las 5 prioridades para el próximo comité.
- **Modo demo guiado y de solo lectura** con un caso completo y las 89 respuestas a la vista.

Sectores cubiertos y datos

Agroindustria · Consumo y comercio · Servicios · Construcción e inmobiliario · Infraestructura · Transporte · Energía · Recursos naturales · Salud, educación y gobierno · Sector financiero · Telecomunicaciones