AI Guardian — Ficha técnica

Nov 14, 2026 · @Maria Camila

Track 5 (sin código). Requisitos mínimos que el Estado y los operadores de infraestructura crítica deben exigir antes de desplegar un agente basado en un modelo de frontera.

1. Vector de riesgo y modelo de amenazas

Caso de estudio: el operador del sistema eléctrico nacional despliega un agente de frontera. **Se protege:** los pesos, la continuidad del servicio, los sistemas de control y las llaves. **Atacan:** desde aficionados hasta Estados (niveles de RAND), personas internas y fallas del propio modelo.

- **Robo de pesos:** T1 leerlos de la memoria del servidor · T2 robar las llaves · T3 sacarlos poco a poco por la red · T4 persona interna sobornada · T5 proveedor o firmware comprometido.
- **Pérdida de control del agente:** T6 instrucciones ocultas en lo que lee · T7 permisos de más · T8 apagar su monitoreo o copiar sus pesos · T9 maniobra sin humano · T10 sin registros.

2. Mecanismo defensivo

32 requisitos de sí o no, cada uno con un documento de evidencia (auditoría, diagrama, configuración o prueba). **Si falla uno, el agente no se enciende.** Exigencia por nivel del agente: N1 apoyo (SL2), N2 operación supervisada (SL3), N3 operación crítica (SL3, con meta de SL4 en 3 años).

- **Pesos (8):** SL auditado; llaves en chip seguro; computación confidencial; red aislada; límite de salida; dos personas; firmware firmado; personal verificado (controles SL5: SC-7(10), SC-28(3), SC-49, AC-3(2)).
- **Contención (9):** mínimo privilegio; aprobación humana; lo leído no es orden; apagado en segundos; aislamiento; límite de acciones; evaluación de capacidades peligrosas; doble aprobación con datos fuente; operación manual sin el agente.
- **Auditoría (5):** registro inmodificable; responsable designado; revisión mensual; reporte de incidentes; auditoría externa en N3.
- **Líneas rojas (7):** ver sección 3.

- **Factor humano (3):** separación operar / aprobar / auditar; pruebas sorpresa con maniobras falsas; aprobaciones firmadas, sanciones y canal de denuncia.

3. Líneas rojas y gobernanza institucional

El agente nunca puede: copiar sus pesos; apagar su monitoreo o registros; cambiar sus permisos o conseguir recursos extra; atacar otros sistemas; hacer maniobras físicas sin aprobación; copiarse a sí mismo; obedecer órdenes escondidas en lo que lee. Si cruza una, se apaga y se reporta.

Gobernanza institucional. Lo **exige** el regulador del sector (en energía: la CREG regula y la Superservicios vigila). Lo **aprueba** un comité de la entidad (seguridad, operación y jurídica), con auditoría externa en N3. **Responden** el responsable designado, la entidad y, por el bloque 1, el proveedor. **Si no se cumple**, el agente no se enciende o se apaga. **Se revisa** cada 12 meses y con cada permiso nuevo.

4. Validación empírica

Prueba (en papel, con apoyo de IA)	Resultado
Caso aplicado: operador eléctrico, nivel N3 (hipotético)	Cumple 17 de 32: **no apto** (sin límite de salida, entre otros)
10 ataques · 5 acciones legítimas	9 frenados + 1 parcial · 0 bloqueadas por error
Cobertura de 10 amenazas	9 cubiertas + 1 parcial (persona interna)
Equipo rojo: 5 ataques nuevos	2 pasaron (engañar al operador, datos falsos) → corregidos en v2 y v3
Robo de 1 TB de pesos con límite de salida	1 Mbps: unos 93 días
Acuerdo entre dos evaluadoras humanas	84 % (27 de 32)

5. **Límites:** no alcanza contra los Estados más capaces (hoy SL3; SL4 en 3 años); depende de la evidencia del proveedor; el riesgo humano se reduce, no se elimina; la prueba es en papel con supuestos propios. **Doble uso:** solo restringe al agente y no describe cómo atacar.

Referencias: RAND – Securing AI Model Weights · Estándar SL5 · Documento completo