Regulación, ética y AI Act

Gobernar lo que despliegas.

Conceptos clave

- **AI Act** (Reglamento UE 2024/1689): cuatro niveles de riesgo, cuatro paquetes de obligaciones.
- Cinco riesgos canónicos: alucinación, *prompt injection*, fuga de datos, sesgo, dependencia de proveedor.
- Cinco artículos clave: 10 (datos), 12 (logs), 13 (transparencia), **14** (HITL), **15** (robustez).
- Auditoría = ciclo continuo: diseño, datos, modelo, despliegue, auditoría externa.

Ejemplo a mano

Checklist AI Act para un scoring crediticio (sistema de alto riesgo, anexo III).

✓	Sistema de gestión de calidad documentado (art. 17).
×	Datos de entrenamiento documentados (art. 10).
✓	Registro automático de operaciones (logs, art. 12).
×	Transparencia hacia usuario (art. 13).
✓	Supervisión humana (HITL, art. 14).
×	Robustez y ciberseguridad (art. 15).
×	Marcado CE y declaración de conformidad (arts. 47-49).

3 OK de 7. Sistema **NO desplegable** hasta cerrar las 4 brechas.

Concepto → aplicación: cinco riesgos y su mitigación

Riesgo	Manifestación	Mitigación
Alucinación	Modelo inventa hechos.	RAG con cita verificable, fallback.
Prompt injection	Input reprograma el agente.	Sanitización, sandboxing.
Fuga de datos (PII)	Agente devuelve PII al cliente equivocado.	ABAC + redacción + audit logs.
Sesgo y discriminación	Penaliza grupos protegidos.	Eval por subgrupos, fairness.
Dependencia proveedor	El modelo cambia, el sistema se rompe.	Abstracción + plan B local.

Idea clave del Día 18. **Cada herramienta añade superficie de ataque.** El *principio de menor privilegio* se aplica también a agentes. En sectores regulados (banca, salud, seguros), nunca HITL nivel 4.

Para llevarte: lectura AI Act overview; entregable: checklist AI Act del proyecto AI-first del grupo.