Deep Learning para Business Analytics

Día 18 · Tema 7 — Regulación, ética y AI Act

Gobernar lo que despliegas.

Eduardo C. Garrido-Merchán

Profesor Propio Adjunto, Métodos Cuantitativos, ICADE

ecgarrido@comillas.edu

Madrid, Día 18

ejemplos del día

Lo que vamos a cubrir hoy.

1. Riesgos de los sistemas IA

2. El marco europeo AI Act

3. Auditoría: el ciclo

4. Caso BBVA: checklist completo

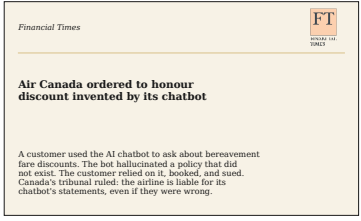
Al final del día: sabes clasificar un sistema en la pirámide AI Act y producir un checklist de auditoría.

Parte 1 / 4

Riesgos de los sistemas IA

Air Canada: el chatbot inventó un descuento y la empresa pagó.

Air Canada tuvo que honrar un descuento inventado por su chatbot.



Financial Times

FT

Air Canada ordered to honour discount invented by its chatbot

A customer used the AI chatbot to ask about bereavement fare discounts. The bot hallucinated a policy that did not exist. The customer relied on it, booked, and sued. Canada's tribunal ruled: the airline is liable for its chatbot's statements, even if they were wrong.

Lecciones

1. El chatbot ES la empresa.
2. Alucinacion = riesgo legal.
3. Sin guardrails, el coste es demanda + reputacion.

El caso. ⚠️ El chatbot de Air Canada alucinó una política de descuento por duelo que no existía. Un cliente la usó, voló, reclamó y ganó.

Tribunal. ⚖️ El tribunal canadiense dictaminó: la aerolínea es responsable de lo que dice su chatbot, aunque sea incorrecto.

Coste. Demanda + indemnización + daño reputacional. El coste real de no tener guardrails.

Idea clave. Sin validación de salida, el riesgo legal es real y medible.

Cinco riesgos ⚠️ típicos y su mitigación canónica.

Cinco riesgos típicos de un sistema IA y su mitigación canónica.

Riesgo	Manifestación	Mitigación
Alucinación	El modelo inventa hechos.	RAG con cita verificable, fallback low-score.
Prompt injection	El input del usuario reprograma el agente.	Sanitización, prompts firmados, sandboxing.
Fuga de datos	El agente devuelve PII al usuario equivocado.	ABAC + redacción + audit logs.
Sesgo y discriminación	Decisiones que penalizan grupos protegidos.	Evaluación por subgrupos, fairness toolkit.
Dependencia de proveedor	El modelo cambia, el sistema se rompe.	Abstracción + tests + plan B local.

Tres riesgos específicos de los agentes con tool use.

Riesgo 1 — Tool use no autorizado ⚠️. Agente con acceso a Bash ejecuta `rm -rf` sobre el repo. Mitigación: lista blanca de comandos y `PreToolUse` hooks.

Riesgo 2 — Exfiltración 🔒 a través de MCP. El agente filtra PII al CRM equivocado. Mitigación: ABAC en cada MCP y redacción de PII en logs.

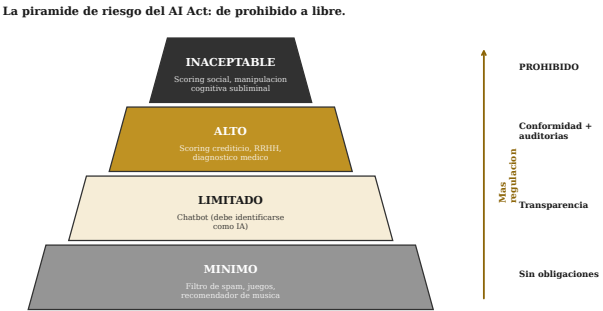
Riesgo 3 — Cadena de razonamiento manipulable 🛡️. Un input mete instrucciones que cambian el comportamiento. Mitigación: prompts firmados y filtrado de *prompt injection*.

Idea clave. Cada tool añade superficie de ataque. El *principio de menor privilegio* se aplica también a agentes.

Parte 2 / 4

El marco europeo AI Act

La pirámide de riesgo del AI Act: de prohibido a libre.



Cuatro niveles. ⚖️ No toda IA se regula igual. La clave es el riesgo sobre derechos fundamentales.

Mínimo. Filtro de spam, juegos. Sin obligaciones.

Limitado. Chatbot: debe identificarse como IA (transparencia).

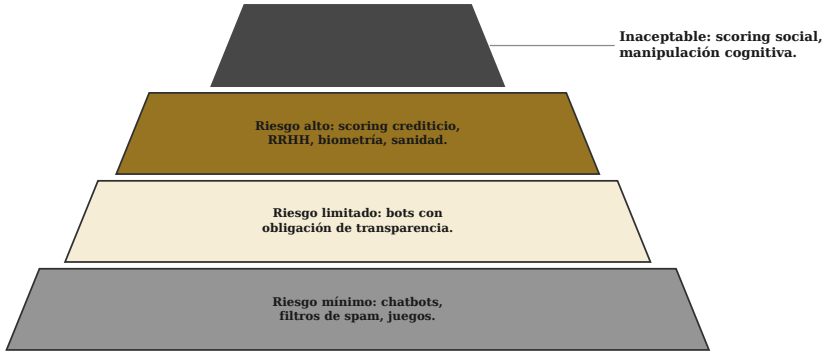
Alto. 🛡️ Scoring crediticio, RRHH, biometría. Auditorías, explicabilidad, logs.

Inaceptable. Scoring social, manipulación cognitiva. Prohibido en la UE.

Idea clave. Clasifica tu sistema antes de diseñarlo: el nivel determina tus obligaciones.

Pirámide de riesgo AI Act ⚖️: cuatro niveles, cuatro paquetes.

AI Act: cuatro niveles de riesgo, cuatro paquetes de obligaciones.



Cinco artículos del AI Act que conviene memorizar.

Art. 10 — Datos. Documentar conjunto de entrenamiento, validación, test. Trazabilidad.

Art. 12 — Registros. Cada decisión queda loggeada de forma auditable.

Art. 13 — Transparencia ⚖️. El usuario sabe que interactúa con IA y conoce capacidades y límites.

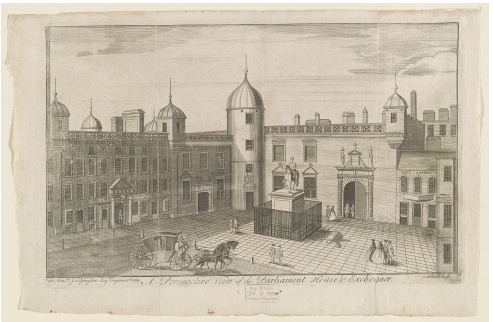
Art. 14 — Supervisión humana 🛡️. En riesgo alto hay HITL real, con autoridad para parar el sistema.

Art. 15 — Robustez. Tests de seguridad, ciberseguridad, gestión del riesgo.

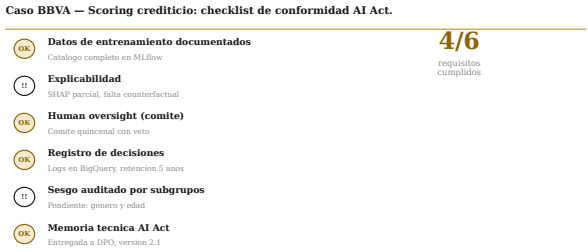
IDEA CLAVE. Conocer estos cinco artículos basta para defender el proyecto AI-first.

Parte 3 / 4

Auditoría: el ciclo



Checklist de conformidad para un sistema de alto riesgo.



Caso. 🏦 Sistema de scoring crediticio en BBVA, clasificado como alto riesgo (Anexo III AI Act).

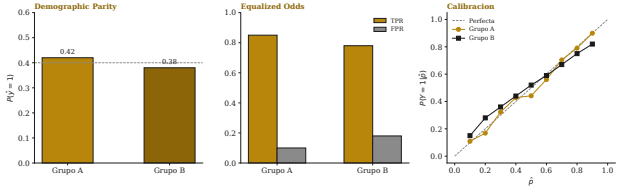
Estado. (1) Datos documentados ✅ (2) Explicabilidad: SHAP parcial, falta counterfactual ⚠️ (3) Human oversight: comité quincenal ✅ (4) Logs: BigQuery, 5 años ✅ (5) Sesgo: pendiente género y edad ⚠️ (6) Memoria técnica: entregada ✅.

Resultado. 4/6 cumplidos. Los ítems pendientes (explicabilidad, sesgo) son los que más retrasan la conformidad.

Idea clave. La auditoría real tiene matices: "parcialmente conforme" es la norma, no la excepción.

Métricas de sesgo: demographic parity, equalized odds, calibración.

No se pueden satisfacer las tres a la vez (Chouldechova, 2017).



Tres métricas. ⚖️

DP: $P(\hat{y}{=}1|A{=}0) = P(\hat{y}{=}1|A{=}1)$.

EO: $P(\hat{y}{=}1|Y{=}1, A{=}a)$ igual $\forall a$.

Cal: $P(Y{=}1|\hat{y}{=}p, A{=}a) = p \ \forall a$.

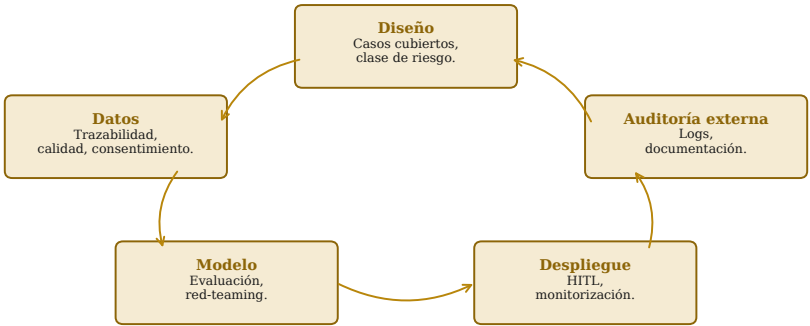
Imposibilidad. ⚠️ No se pueden satisfacer las tres a la vez (Chouldechova, 2017). Hay que elegir cuál priorizar según el contexto.

Práctica. En scoring crediticio, equalized odds suele dominar: mismo TPR y FPR por grupo protegido.

Idea clave. Medir sesgo sin elegir métrica es como medir error sin elegir loss.

Auditoría continua: cinco estaciones.

Auditoría de un sistema IA: ciclo continuo con cinco estaciones.



Concepto → aplicación: gobernar un sistema agéntico.

Pieza	Qué documenta	Quién la revisa
Memoria técnica ✅	Arquitectura, datos, evaluación.	DPO + Compliance.
Logs auditables 🛡️	Cada decisión y su justificación.	Auditoría interna anual.
Red-teaming ⚠️	Tests adversariales (prompt injection, jailbreak).	Seguridad.
Monitor en vivo	Latencia, faithfulness, fallbacks.	Producto + Ops.
Plan de retirada	Cómo apagar el sistema si falla.	Comité de IA.

Idea clave. Estos cinco artefactos son los entregables que pedirá un regulador en 2026-2027.

Parte 4 / 4

Caso BBVA: checklist completo

Caso BBVA 🏦: checklist ✅ AI Act para scoring crediticio.

Caso BBVA: checklist AI Act para scoring crediticio (5 de 8 OK).

- ✓ Sistema clasificado como ALTO RIESGO (anexo III, scoring crediticio).
- ✓ 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).

Lo que el alumno se lleva a casa de este día.

Concepto 1. El AI Act ⚖️ clasifica sistemas en cuatro niveles: mínimo, limitado, alto, inaceptable. Las obligaciones suben con el riesgo.

Concepto 2. Cinco riesgos ⚠️ típicos (alucinación, prompt injection, fuga, sesgo, dependencia) tienen mitigaciones canónicas.

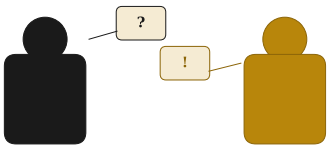
Concepto 3. La auditoría 🛡️ es un ciclo de cinco estaciones, no un evento puntual.

Concepto 4. Cinco artículos del AI Act (10, 12, 13, 14, 15) son la base de cualquier checklist ✅.

Para la práctica. Producir el checklist AI Act del proyecto AI-first del grupo.

Próximo día. Día 19 (opcional): repaso para examen + clínica final del proyecto.

Pausa para debatir: ¿Es lenta la AI Act para la velocidad de la tecnología?



La pregunta. El reglamento entra en vigor mientras los modelos evolucionan cada trimestre. ¿Llega tarde?

Posición A. Sí, regular tecnología obsoleta antes de su despliegue.

Posición B. No, regula principios estables (riesgo, transparencia, supervisión).

Reglas. 5 minutos, dos turnos de palabra por bando, móviles boca abajo. Yo modero, no participo.

Fin del curso.

Gracias por el semestre.

Deep Learning · **Agentes** · Regulación

Eduardo C. Garrido-Merchán

ecgarrido@comillas.edu