

1. Riesgos estructurales de los sistemas IA

Cuatro familias de riesgo que el ingeniero debe conocer:

- **Sesgo:** el modelo aprende correlaciones del conjunto de entrenamiento; si los datos están sesgados, las decisiones también.
- **Alucinación:** el LLM produce afirmaciones plausibles pero falsas con confianza alta.
- **Privacidad:** información personal puede memorizarse durante el entrenamiento y filtrarse en la generación.
- **Manipulación:** deepfakes, voz clonada, suplantación de identidad, generación de desinformación a escala.

2. Sesgo, formalmente

Sea $\mathcal{A}$ un atributo protegido (género, etnia, edad). Tres definiciones formales:

$$\begin{aligned} \text{Demographic parity: } & p(\hat{y} = 1 \mid \mathcal{A} = 0) = p(\hat{y} = 1 \mid \mathcal{A} = 1), \\ \text{Equalized odds: } & p(\hat{y} = 1 \mid Y = y, \mathcal{A} = 0) = p(\hat{y} = 1 \mid Y = y, \mathcal{A} = 1) \quad \forall y, \\ \text{Calibración: } & p(Y = 1 \mid \hat{p} = p, \mathcal{A} = 0) = p(Y = 1 \mid \hat{p} = p, \mathcal{A} = 1) \quad \forall p. \end{aligned} \tag{1}$$

Y etiqueta verdadera
 $\hat{y}, \hat{p}$ predicción de clase y de probabilidad
 $\mathcal{A}$ atributo protegido

Resultado de imposibilidad (Kleinberg, Chouldechova): *las tres no pueden satisfacerse simultáneamente excepto en casos degenerados*. El diseñador elige cuál garantiza y lo justifica al regulador.

3. El AI Act europeo (Reglamento 2024/1689)

Aplicable progresivamente desde agosto 2024, plenamente vigente en 2026. Cuatro categorías de riesgo:

- **Inadmisibles:** scoring social estilo China, manipulación subliminal, biometría en tiempo real en espacios públicos. Prohibido.
- **Alto riesgo:** educación, empleo, crédito, justicia, infraestructura crítica. Requiere evaluación de conformidad, sistema de gestión de riesgos, datos de entrenamiento documentados, supervisión humana y registro.
- **Riesgo limitado:** chatbots, deepfakes. Obligación de transparencia (el usuario debe saber que está interactuando con IA o que el contenido es generado).
- **Riesgo mínimo:** filtros antispam, recomendadores. Sin obligaciones adicionales.

Modelos de propósito general ($\geq 10^{25}$ FLOPs de entrenamiento): obligaciones adicionales (*model card*, datos de entrenamiento, evaluaciones de riesgo sistémico).

4. Auditoría: el ciclo

Una auditoría de un sistema IA cubre cinco bloques:

1. **Documentación:** model card, datasheet del dataset, *purpose*, restricciones.
2. **Trazabilidad:** logs por inferencia, retención ≥ 6 meses.
3. **Métricas de equidad:** las medidas en (1) anteriores, calculadas sobre el dataset de validación estratificado por $\mathcal{A}$.
4. **Pruebas de robustez:** adversarial, red-team, drift de datos.
5. **Supervisión humana:** roles, vías de apelación, criterios de override.

$$\text{Conformidad} = \bigwedge_{b=1}^5 (\text{bloque } b \text{ aprobado}). \tag{2}$$

Falta cualquier bloque $\Rightarrow$ no conformidad $\Rightarrow$ multa hasta el 7 % de la facturación global o 35 M€.

5. Sesgo en la práctica: caso scoring de crédito

Un modelo de crédito que aprende sobre histórico replicará el sesgo histórico. Tres mitigaciones formales:

- **Pre-proceso:** reponderar el dataset para que $p(Y|\mathcal{A})$ sea balanceada antes de entrenar.
- **In-proceso:** añadir un término de regularización $\lambda \Delta_{\text{eq}}$ a la pérdida, donde Δ_{eq} es la violación de la métrica elegida.

- **Post-proceso:** ajustar el umbral τ_a por subgrupo a para satisfacer equalized odds tras entrenar.

IDEA CLAVE. Ninguna técnica elimina el sesgo sin coste. El diseñador hace una elección política (qué métrica priorizar) que debe documentar y defender.

6. Compromiso del egresado

El alumno de este grado va a tomar decisiones que afectan a personas físicas. La ética no es un capítulo del último día: es la condición que hace defensible cualquier decisión técnica que tome. Tres principios operativos:

- **Trazabilidad.** Toda decisión IA debe poder explicarse a un humano afectado.
- **Reversibilidad.** Las decisiones de alto impacto deben admitir apelación y rectificación.
- **Proporcionalidad.** No usar un sistema más invasivo de lo necesario para el caso de uso.

Cuando dudes: pregunta al DPO de tu empresa antes de desplegar.