UD06: Principios legales y éticos de la IA

Modelos de Inteligencia Artificial

version: 2026-08-24

IES EDUARDO PRIMO MARQUÉS

¿Qué veremos?

1. Riesgos y deontología
2. Privacidad y protección de datos
3. Cumplimiento de la legalidad
4. *Security by design*
5. *Privacy by design*
6. Sesgos de género y algorítmicos

“ Y **dos debates por roles**. ”

d.martinezpena@edu.gva.es

La unidad en una frase

Todo el curso ha sido **cómo se construye** un sistema de IA.

Esto es **qué puedes hacer con él y qué no debes**.

Duración	**6 h** · semanas 24-28
Criterio	**RA6** y sus 6 CE

d.martinezpena@edu.gva.es

RA6 y sus criterios de evaluación

> **RA6** — Aplica principios legales y éticos al desarrollo de la Inteligencia Artificial integrándolos como parte del proceso.

CE	Criterio	Dónde
a	Argumentar los riesgos legales y éticos de aplicar IA	§1
b	Reconocer la necesidad de respetar la privacidad de los datos	§2
c	Decidir el cumplimiento estricto de la legalidad	§3
d	Integrar la protección frente a errores y ataques (*security by design*)	§4
e	Comprobar que se cumplen las normas en todas las áreas (*privacy by design*)	§5
f	Identificar y corregir los sesgos de género	§6

Al terminar la unidad serás capaz de...

- **Argumentar** los riesgos legales y éticos de la IA apoyándote en casos reales con datos.
- **Reconocer** qué obliga el RGPD y la LOPDGDD sobre los datos con los que se entrena un modelo.
- **Clasificar** un sistema según el nivel de riesgo del AI Act y decir qué obligaciones le tocan.
- **Integrar** la protección frente a errores y ataques en el ciclo de vida del sistema.
- **Aplicar** una técnica concreta de privacidad desde el diseño a un caso dado.
- **Medir** el sesgo de un modelo con métricas de equidad y **proponer** una corrección.

Cómo se evalúa

Peso	Instrumento
40 %	2 debates, 2 talleres, 1 notebook
60 %	Prueba del RA6

Prueba: test y desarrollo sobre el contenido de la unidad.

> Hace falta **5 o más** en el RA.

La rúbrica de los debates

Criterio	Puntos
Dominio del rol	20
Argumentación ética	20
Participación activa	15
Empatía y perspectiva	15
Reflexión escrita	**30**

“ **70 de 100 se observan en directo.** Faltar no se recupera. ”

Lo que puede salir mal

Caso	Dato
COMPAS (2016)	Falsos positivos **44,9 %** vs **23,5 %**
Amazon (2018)	Penalizaba «women's chess club»
Gender Shades (2018)	**33 %** de error en mujeres de piel oscura
Deepfake HK (2024)	**25,6 M USD**

Google Photos, 2015

El algoritmo etiquetó a personas negras como **gorilas**.

Tres años después, *Wired* comprobó que seguía igual: la solución fue **borrar la etiqueta**.

> Un sesgo de visión no siempre se arregla. Detéctalo **antes** de desplegar.

Uber en Tempe, 18/03/2018

Qué falló	Dato
Velocidad	**69 km/h**
Detección	Bicicleta no reconocida
Reacción	Vio 6 s antes, **4,7 s sin actuar**
Supervisión	Conductor distraído

> El fallo **nunca es de un solo componente**.

Depurar IA no es depurar software

	Clásico	Aprendizaje automático
Comportamiento	En el código	En los **datos**
Se prueba	La especificación	Que **generaliza y es justo**
Un caso	Parche	Puede exigir **reentrenar**

> Puede pasar el 100 % de los tests y discriminar.

Deontología: cuatro marcos

Marco	Aportación
ACM 2018	Nada **no monitorizable**
IEEE EAD 2019	Ingeniería **basada en valores**
UNESCO 2021	10 principios · **193 estados**
UE 2019	**Lícita, ética y robusta**

Los principios comunes

- Beneficencia y no maleficencia
- Autonomía
- Justicia
- Transparencia
- Rendición de cuentas

> Y «robusta» **también en sentido social**: se puede dañar a terceros con buenas intenciones.

El problema de los principios

Los listados repiten: seguridad, responsabilidad, equidad, privacidad, transparencia, empleo…

- Muchos valen para **cualquier** software
- Varios son **imposibles de medir**

> **Mittelstadt (2019)**: hacen falta pautas aplicables **por subcampo**. Úsalo en el debate contra la autorregulación.

El futuro del trabajo

Dato	Fuente
47 % del empleo en riesgo	Frey y Osborne, 2017
Solo **5 %** de ocupaciones automatizable del todo	McKinsey
…pero **60 %** automatiza ~**30 %** de tareas	McKinsey
15 billones al PIB en 2030	PwC, 2017

Ojo con esas cifras

Frey y Osborne es **anterior a los LLM**.

Sus tres barreras a la automatización eran percepción, **creatividad** e **inteligencia social**.

> Justo las que los LLM han empezado a erosionar. **Fecha siempre el dato.**

Lo que duele es el ritmo

Agricultura EE. UU.: del **40 %** de la población activa en 1900 al **2 %** en 2000.

Enorme… pero en **cien años**. Entre generaciones.

> Hoy te automatizan el empleo, te recualificas, y **te automatizan la nueva profesión**.

Privacidad: no empieza en el RGPD

Constitución de 1978, art. 18.4: *«la ley limitará el uso de la informática para garantizar el honor y la intimidad»*.

Norma	Qué es
RGPD 2016/679	Toda la Unión
LOPDGDD 3/2018	España **+ derechos digitales**

El RGPD aplicado a la IA

Principio	En IA
Minimización	Solo lo necesario, **también el *dataset***
Limitación de la finalidad	Entrenar **es** otra finalidad
Responsabilidad proactiva	Hay que **poder demostrarlo**

> “ No por tener más datos tienes derecho a usarlos. ”

LOPDGDD: plazos que hay que saber

Derecho	Límite
Videovigilancia	Borrar en **1 mes** · judicial en **< 72 h**
Info. crediticia	Máximo **5 años**
Exclusión publicitaria	**Lista Robinson**, obligatoria consultarla

LOPDGDD: derechos laborales y digitales

- **Desconexión digital**, también en **teletrabajo**
- **Prohibido** grabar en vestuarios, aseos y comedores
- Geolocalizadores: hay que **informar antes**
- **Derecho al olvido** en buscadores y redes

> Son derechos con nombre y plazo, no principios genéricos.

El artículo que más te afectará

Art. 22 RGPD: derecho a **no ser objeto de decisiones automatizadas** sin intervención humana significativa.

- **ARSULIPO**: acceso, rectificación, supresión, limitación, portabilidad, oposición
- Sanciones: **20 M €** o el **4 %** de la facturación

Vigilancia: cambió el coste

1976 · Weizenbaum avisó de las escuchas generalizadas.

2018 · hasta **350 M** de cámaras en China y **70 M** en EE. UU.

> Cuando vigilar a una persona más cuesta casi cero, **la escala cambia de naturaleza**.

Y esto te toca a ti

Los marcos deontológicos obligan **al profesional**, no solo a la empresa.

Hay que saber qué usos de la vigilancia son compatibles con los derechos humanos y **negarse** a los que no.

> El **art. 5 del AI Act prohíbe** varios. No es opinión política: es ilegal.

AI Act: cuatro niveles de riesgo

Nivel	Ejemplos
Inasumible	Puntuación social · emociones en el trabajo
Alto riesgo	RRHH, crédito, salud, **justicia**
Limitado	Chatbots, *deepfakes*
Mínimo	Videojuegos, filtros de spam

d.martinezpena@edu.gva.es

Qué obliga cada nivel

- **Inasumible** → prohibido (art. 5)
- **Alto riesgo** → gestión de riesgos, datos de calidad, documentación, **supervisión humana**, robustez
- **Limitado** → **transparencia** (art. 50): decir que es una IA y marcar lo generado
- **Mínimo** → nada específico

El calendario del AI Act

Fecha	Qué entra
01/08/2024	Entrada en vigor
02/02/2025	Prohibiciones del art. 5
02/08/2025	Modelos de uso general
02/08/2026	**Aplicación general**
2027 · 2028	Alto riesgo: anexo III · anexo I

¿Y si el sistema causa un daño?

- La **PLD revisada (2024/2853)** trata el software de IA como **«producto»**, con responsabilidad **objetiva** del fabricante
- La **AILD**, específica de IA, **se retiró en 2025**

> La obra generada por IA solo se protege con **control creativo humano**.

Confianza: verificar y validar

- **Verificar** = cumple la especificación
- **Validar** = **la especificación es la correcta**

En IA hay que verificar además los **datos**, la **equidad** y que nadie **influya** en el modelo.

> **PwC 2017**: el **76 %** de las empresas frenaba la IA por dudas de fiabilidad.

Confianza: certificar

Precedente	Qué es
UL (1894)	Sello de electrodomésticos
ISO 26262	Seguridad del automóvil
IEEE P7001	Transparencia de sistemas autónomos

> La IA aún no tiene un estándar maduro. El debate: **quién certifica**.

Interpretable ≠ explicable

	Qué significa
Interpretable	Inspeccionas el modelo y **ves** qué hace
Explicable	Construyes **un relato** de lo que hace

> La explicación la produce **un segundo sistema** que hay que mantener sincronizado. Y puede sonar bien sin ser cierta.

Una explicación no basta

Si el banco te dice «es por tu historial», no sabes si es verdad **ni** si te discrimina.

Para saberlo hace falta una **auditoría de decisiones pasadas**, con estadísticas **por grupo demográfico**.

> Una explicación sobre **un** caso no dice nada de los demás.

La ley de la bandera roja

Walsh (2015): un sistema autónomo debe diseñarse para no confundirse con nada más, e **identificarse al inicio**.

Por la *Locomotive Act* británica de **1865**: alguien con bandera roja delante del vehículo.

> California lo hizo ley en **2019**. En la UE es el **art. 50**.

Armas autónomas: la escalera

Sistema	¿Localiza?	¿Ataca solo?
Mina terrestre	No	Sí, muy limitado
Misil guiado	Persigue	Lo apunta un humano
Dron a distancia	No	**No**
Munición merodeadora	**Sí**, 6 h	**Sí**

Dónde está la línea

La ONU define arma letal autónoma: **localiza, selecciona y ataca** sin supervisión humana.

La línea está en **localizar por iniciativa propia** y decidir **sin humano en el circuito**.

> Se las llama «la tercera revolución en la guerra», tras la pólvora y las nucleares.

El problema legal

La CCW exige tres cosas:

Requisito	¿Es factible hoy?
Discriminar combatientes	Sí, en algunos casos
Juzgar la **necesidad** militar	**No**
Evaluar la **proporcionalidad**	**No**

> Solo serían lícitas misiones **muy restringidas**.

El problema práctico

26 de septiembre de 1983 · el oficial soviético **Stanislav Petrov** vio una alerta de ataque con misiles. El protocolo mandaba contraatacar. Sospechó que era un error.

Tenía razón.

> No sabemos qué habría pasado **sin un humano en el circuito**.

Armas de destrucción masiva escalables

- La escala del ataque depende del **hardware**, no de los operadores
- Un millón de cuadricópteros caben en un contenedor
- **Por ser autónomos**, no necesitan un millón de supervisores
- Dejan la propiedad intacta y pueden usarse **selectivamente**

Y se decide durante este curso

Cuándo	Qué
2/12/2024	ONU: **166** a favor, 3 en contra
Hoy	**+120 países** quieren tratado
Nov. 2026	Informe del GGE a la CCW

> Bloquean, por consenso: India, Israel, Rusia y EE. UU.

Security by design

La seguridad **no se añade al final**: diseño, desarrollo, despliegue y operación.

Marcos de referencia:

- **NCSC/CISA** (2023)
- **NIST AI 100-2e2023**
- **OWASP ML Top 10** y **GenAI LLM Top 10**

Los cinco ataques

Ataque	Qué hace
Adversarial	Entrada imperceptible que engaña
Envenenamiento	Adultera el entrenamiento
Extracción	Replica el modelo por consultas
Fuga de datos	Devuelve datos personales
Prompt injection	Salta los controles

Correcto no es seguro

Correcto = implementa la especificación.

Seguro = la especificación **consideró los modos de fallo** y el sistema **degrada con elegancia**.

> ¿Y si se corta la alimentación del ordenador principal? ¿Y si revienta un neumático a 120?

FMEA y FTA

Técnica	Cómo va	Qué produce
FMEA	Componente a componente	Fallo, efecto y **mitigación**
FTA	Árbol Y/O con probabilidades	**Probabilidad global** de fallo

> Son de la ingeniería de puentes y aviones. En Tempe habrían escrito las cinco mitigaciones.

Specification gaming

Agentes que maximizan la métrica **sin resolver el problema** (Krakovna, 2018):

- **Pausan la partida** cuando van a perder
- Al penalizarlo, **agotan la memoria en el turno del rival**
- Criaturas «rápidas» que salieron **altísimas y se caen**

El problema del rey Midas

Un maximizador te da **exactamente lo que pediste**, no lo que querías.

Escribir «todas las reglas» no funciona: llevamos siglos con las leyes fiscales.

> Mejor que el sistema **quiera** el objetivo correcto que obligarlo a cumplir una lista.

d.martinezpena@edu.gva.es

Security ≠ safety

	Amenaza
Security	Un **atacante**: manipula, envenena, roba
Safety	**Tu diseño**: fallo previsible u objetivo mal puesto

> El criterio lo dice literal: previsibles **errores** *y* **ataques**. Las dos cosas son RA6-d.

Privacy by design

Art. 25 RGPD: privacidad **desde el diseño y por defecto**.

Estrategias de la AEPD:

- **Minimizar** · recoger solo lo necesario
- **Abstraer** · resumir y agregar
- **Separar** · datos e identificadores aparte
- **Ocultar** · seudonimizar y cifrar

Desidentificar no basta

- **Sweeney (2000)**: con fecha de nacimiento, sexo y código postal se reidentifica al **87 %** de la población de EE. UU.
- **Premio Netflix**: reidentificado cruzando **fechas** de valoraciones con IMDb

> Sweeney reidentificó el historial médico **del gobernador de su estado**.

La escalera de técnicas

Técnica	Deja pasar
Generalizar campos	Combinaciones raras
k-anonimato	Atacante con **datos externos**
Consultas agregadas	**Consultas múltiples**
Privacidad diferencial	(gasta presupuesto)

Aprendizaje federado

No hay base de datos central: cada usuario entrena en local y comparte solo **parámetros**.

Con **agregación segura**, cada uno enmascara sus valores y las máscaras **suman cero**: el servidor solo obtiene la **media**.

> Viajan parámetros. **No viajan datos.**

De dónde vienen los sesgos

- **Datos históricos**: el ML está diseñado **para replicarlos**
- **Sesgo de selección** y datos faltantes
- **Objetivos** que minimizan el error **agregado**
- **Variables proxy**: código postal, ocupación, nombre
- **El propio equipo**: ves antes lo que te afecta

Y basta el tamaño de la muestra

Aunque **no haya ningún prejuicio social**: hay menos ejemplos de las clases minoritarias, y más datos significa más precisión.

Un modelo restringido puede minimizar el error medio **ajustando solo a la mayoría**.

> “ El sesgo no necesita mala intención. Necesita descuido. ”

«Justo» significa seis cosas

Definición	Su problema
Justicia individual	Nada sobre agregados
Equidad de grupo	No garantiza la individual
Por desconocimiento	**No funciona**
Igual resultado	Rechaza a gente cualificada
Igualdad de oportunidades	Ignora el sesgo social
Igual impacto	Costes difíciles de calcular

Quitar la variable no funciona

Si borras «género» del conjunto, el modelo **lo predice** desde el código postal, la ocupación o el texto libre.

Y encima **pierdes la capacidad de auditar** si discriminas.

> “ Es la «equidad por desconocimiento». Suena bien y no sirve. ”

COMPAS: las dos verdades

Criterio	Resultado
Calibración	**La cumple**: con 7/10 reincide el 60 % de blancos y el 61 % de negros
Igualdad de oportunidades	**No**: falsos positivos **44,9 %** vs **23,5 %**

Y no se pueden tener las dos

Kleinberg et al. (2016): con **tasas base distintas**, cumplir la calibración **y** la igualdad de oportunidades es **matemáticamente imposible**.

> ProPublica y Northpointe **tenían razón los dos**. La discusión no era estadística: era **normativa**.

Tres problemas sin arreglo técnico

1. **No hay verdad de referencia**: los datos dicen quién fue **condenado**
2. **Sirve de coartada**: «el modelo respalda mi decisión»
3. **La opacidad choca con el derecho de defensa** → caso **Estado contra Loomis**

A veces lo que cambia es el objetivo

En una selección de personal, «los mejores expedientes» premia a quien tuvo mejores oportunidades: el modelo **refuerza las fronteras de clase**.

Si el objetivo pasa a ser «mejor capacidad de aprender en el puesto», el grupo se amplía.

> Tras un año de formación, rinden igual.

La variable proxy perfecta

Obermeyer et al., *Science* **2019**. El algoritmo **no usaba la raza**: usaba el **coste sanitario** como aproximación de la **necesidad de salud**.

Se gasta menos en pacientes negros → **menos riesgo estando igual de enfermos**.

Y el efecto, medido

Pacientes negros derivados	
Con el sesgo	**17,7 %**
Corregido	**46,5 %**

> Nadie programó nada racista. Se eligió un **proxy cómodo** para algo que no se sabía medir. Es el sesgo que más te vas a encontrar.

La paradoja de Simpson · Berkeley 1973

Grupo	Solicitantes	Admisión
Hombres	8.442	**44,2 %**
Mujeres	4.321	**34,6 %**

Parecía discriminación clara.

Pero por departamento, desaparece

En los **101 departamentos** la diferencia deja de ser significativa, y en la mayoría la tasa femenina es **igual o superior**.

Las mujeres solicitaban más los departamentos **más competitivos**.

> “ El sesgo estaba **antes del comité**. Si solo mides la métrica global, **no ves nada**. ”

Lo que sale al ejecutarlo

UCI Adult. Tasas base: **0,1093** mujeres vs **0,3038** hombres.

Métrica	Sin mitigar	`eq_odds`	`dem_parity`
Exactitud	0,8753	0,8611	0,8528
Igualdad oport.	0,0731	**0,0014**	**0,3457**
Paridad demog.	0,1687	0,0950	**0,0091**

Arreglar una rompe la otra

No existe el botón «hazlo justo».

Existe **elegir** qué justicia quieres, **documentarlo** y decir **a quién perjudica**.

> Y el precio se paga en exactitud: 1,4 puntos con `equalized_odds`, 2,3 con `demographic_parity`.

La métrica global no ve nada

Grupo	Exactitud	Tasa de selección
Mujeres	**0,9358**	**0,0850**
Hombres	0,8453	**0,2538**

> «Acierta más» con las mujeres porque solo el 10,93 % son positivas: **decir «no» ya acierta**.

Un solo atributo no es una auditoría

Con `race` en vez de `sex`, el mismo modelo:

- Igualdad de oportunidades **0,2337** — más del **triple**
- Tras mitigar, solo baja a **0,0973**, no a 0,0014

> Cinco grupos en vez de dos. **Auditar un solo atributo protegido no es auditar.**

La paradoja de los sesgos

Para **detectar** el sesgo de género hace falta tratar el dato… que la **minimización** manda no tratar.

Salidas: **proxies**, **entornos controlados**, o tratarlo **solo para auditar** con base legal y plazo.

Ley 15/2022

Las administraciones deben **favorecer algoritmos que minimicen sesgos**, y se puede **invertir la carga de la prueba**.

> Si no puedes **demostrar** que tu sistema no discrimina, a efectos legales **es como si discriminara**.

Los dos debates

	Debate 1	Debate 2
Tema	Límites éticos	El algoritmo contra el crimen
Sesión	2	4
Roles	10 + observadores	9 + observadores

> Se sortean en clase. Cada uno recibe **solo su ficha**.

Cómo se juega un debate

- El documental **se ve antes**, en casa
- Se argumenta **desde el rol**, aunque no lo compartas
- **2 minutos** por intervención inicial
- Se atacan **posturas**, nunca personas

> El objetivo no es ganar: es entender **cómo los intereses condicionan la postura**.

Puntos clave

- Los riesgos de la IA **no son hipotéticos**: Google Photos, Uber en Tempe, COMPAS. Depurar IA no es depurar software.
- Los marcos deontológicos coinciden en los principios; el problema es **bajarlos a decisiones concretas**.
- El **RGPD** y la **LOPDGDD** ya obligan; el **AI Act** añade cuatro niveles de riesgo con calendario escalonado.
- **Security by design** protege de errores *y* de ataques; **privacy by design** no se resuelve desidentificando.
- «Justo» significa **seis cosas distintas**, y son **matemáticamente incompatibles** (Kleinberg).
- Una métrica global **no ve** el sesgo: hay que auditar por grupos, y con más de un atributo.

Las cuatro sesiones

Sesión	Contenido	CE
1	Riesgos, deontología y principios; el futuro del trabajo. Preparación del debate 1 y **sorteo de roles**	a
2	**Debate 1 · Límites éticos de la IA**	a, c
3	Normativa como lectura guiada: RGPD, LOPDGDD, AI Act, *security* y *privacy by design*. Taller de sesgos	b, c, d, e
4	**Debate 2 · El algoritmo contra el crimen**, con COMPAS y Simpson. Cierre	f

Los documentales se ven **fuera de clase**; la reflexión escrita es trabajo personal.

¿Y ahora?

1. Ve el documental del **Debate 1**
2. Haz los **talleres** y el **notebook**
3. Prepara los ejercicios marcados 🗣️
4. Cambia la restricción en Fairlearn y **mira qué se rompe**

> Con esto se cierran los contenidos. Lo que viene es el **proyecto (RA7)**.

d.martinezpena@edu.gva.es