IA generativa y uso responsable

Pregúntale a un chatbot quién fundó el club de fútbol de tu pueblo, y te responde en un segundo: un nombre, un año, una frase bonita sobre los fundadores. Suena seguro. ¿Es verdad?

A veces sí. A veces el nombre, el año e incluso la «fuente» que cita se han inventado en el momento, escritos con la misma soltura que la verdad. Empresas de las Islas Baleares ya usan estas herramientas para atender a los huéspedes, escribir descripciones y hacer borradores de código. Un chatbot de hotel que se inventa una política de cancelación no es un error gracioso: es una reclamación, quizá una demanda. Hoy aprendes cómo funcionan estos modelos, cómo pillarlos cuando se equivocan y qué normas les ha puesto Europa.

Qué aprenderás

- Explicar qué es la **IA generativa** y cómo escribe un **gran modelo de lenguaje**, sin magia
- Pillar una **alucinación** contrastando las respuestas de un modelo con fuentes reales
- Escribir normas para el **uso responsable** de la IA en una empresa
- Clasificar sistemas de IA según los cuatro niveles de riesgo del **Reglamento de IA**, y saber qué partes ya se aplican
- Evitar la trampa del día: un modelo que escribe bien no es un modelo que sabe

Modelos que crean

> La **IA generativa** es la IA que **produce contenido nuevo** (texto, código, imágenes, audio, vídeo) a partir de una instrucción, en lugar de solo clasificar o predecir un valor.

Tu modelo de Teachable Machine de la lección anterior **clasificaba**: correcta o defectuosa. Un modelo generativo **crea**: un párrafo, una imagen, una función.

> Un **gran modelo de lenguaje (LLM)** es una red neuronal muy grande entrenada con cantidades enormes de texto para **predecir el siguiente fragmento de texto**. Repitiendo esa predicción una y otra vez, escribe respuestas enteras.

Ese es todo el truco. No busca la respuesta en una base de datos; escribe la continuación más probable de tu texto, fragmento a fragmento. En el juego de construir ciudades es un asesor que habla con mucha seguridad, incluso cuando no ha mirado el mapa.

Término	Qué quiere decir
Prompt	La instrucción o el texto de entrada que le das al modelo. Más contexto y más precisión dan mejores respuestas
Token	Un fragmento de texto (una palabra o parte de una palabra) con el que trabaja el modelo. Los precios y los límites se cuentan en tokens
Alucinación	Una respuesta falsa o inventada, bien escrita y convincente

Término	Qué quiere decir
Fecha de corte del conocimiento	El modelo solo «sabe» lo que había en sus datos de entrenamiento; lo posterior no lo sabe, salvo que se lo des o que busque en la web

Qué hace bien	Sus límites
Escribir, resumir, traducir y reformular	Por sí solo no comprueba los hechos: puede inventarse datos, citas y fuentes
Explicar y generar código, pruebas y documentación	Puede repetir los **sesgos** de sus datos de entrenamiento
Sacar información de textos desordenados	La misma pregunta puede dar respuestas diferentes
Darte un primer borrador rápido	Lo que escribes puede salir de la empresa (privacidad); gasta mucha energía

> **Cuidado: «suena seguro, así que es correcto»**
> Un LLM escribe una respuesta correcta y una equivocada exactamente con el mismo estilo seguro. El problema no es que falle: todas las herramientas fallan. El problema es que **falla sin avisar**. Por eso cada resultado es un borrador hasta que una persona lo revisa.

En clase: cazadores de errores

Esta es la **única actividad de la unidad en la que se autoriza la IA generativa**, y solo en estas condiciones:

> **Qué está autorizado y qué no**
> **Autorizado:** usar el asistente de IA que indique el profesor, **en clase y solo durante esta actividad**, para hacer las preguntas de abajo y analizar las respuestas.
>
> **No autorizado:** escribir tus conclusiones con IA, cualquier otra actividad de clase y cualquier parte de la Entrega 4. No escribas nunca en el asistente datos personales (tuyos o de otra persona) ni datos internos de la empresa.

Por parejas. Cada respuesta se debe contrastar con una **fuente real**: una web oficial, la prensa, la documentación del lenguaje, o ejecutando el código.

Paso	Pregúntale al asistente	Compruébalo con
1	Tres preguntas sobre cosas que conoces muy bien: tu pueblo o tu barrio, tu instituto, un deporte o un juego que dominas	Tu propio conocimiento, confirmado con una web
2	Dos datos sobre la empresa de tu proyecto que **ya has citado** en tu dosier (año de fundación, número de hoteles o de trabajadores…)	La comparación con la fuente de tu dosier
3	Una función corta en el lenguaje que usas en Programación	Ejecutándola, con un caso normal y un caso límite
4	Las fuentes de una de sus respuestas	Abriendo los enlaces: ¿existen? ¿Dicen lo que afirma el modelo?

Para cada respuesta, marca **correcta**, **incorrecta** o **no se puede comprobar**, y pega el enlace que has usado para comprobarla. Después pon nombre al **tipo** de error: dato inventado, fuente inexistente, código que no funciona, respuesta desactualizada, sesgo.

Normas de uso responsable

De los errores que cazas salen cinco normas. La mayoría de las empresas que usan IA escriben algo parecido.

Norma	Qué quiere decir
Revísalo	Todo lo que produce un modelo es un borrador hasta que alguien lo revisa
Protege los datos	Nada de datos personales, ni código ni documentos confidenciales, en herramientas que la empresa no ha aprobado
Sé transparente	Di cuándo has usado IA y para qué
El responsable eres tú	Quien firma el trabajo o publica el código responde de él, no la herramienta
Respeta la autoría	Respeta la autoría y las licencias del contenido y del código

Tu grupo las convertirá en normas para **vuestra** empresa, con al menos una norma propia de su sector. En un hotel, por ejemplo: «el chatbot nunca confirma una reserva ni un precio sin el sistema de reservas».

No es solo buena educación. Desde el **2 de febrero de 2025**, el Reglamento de IA obliga a las empresas que usan IA a asegurarse de que su personal tiene suficientes conocimientos sobre ella, lo que se llama **alfabetización en IA**.

El Reglamento de IA: normas según el riesgo

> El **Reglamento de IA** es el **Reglamento (UE) 2024/1689**, la ley europea sobre inteligencia artificial. Entró en vigor el **1 de agosto de 2024** y se aplica por etapas.

Afecta sobre todo a los **proveedores** (quien desarrolla o vende un sistema de IA) y a los **responsables del despliegue** (quien usa uno de manera profesional) en la Unión Europea. Su idea es un **enfoque basado en el riesgo**: cuanto más daño puede hacer un sistema a las personas, más obligaciones tiene. La Comisión Europea lo resume en cuatro niveles.

Nivel	Qué quiere decir	Ejemplos
Inaceptable	Prácticas prohibidas	Puntuación social de la ciudadanía; manipulación que aprovecha vulnerabilidades; reconocimiento de emociones en el trabajo o en la escuela (salvo por motivos médicos o de seguridad)
Alto	Permitido con obligaciones estrictas: gestión de riesgos, calidad de los datos, documentación, registro, supervisión humana	Ordenar currículums para contratar a personas, evaluar a estudiantes, decidir sobre préstamos, infraestructuras críticas
Limitado (transparencia)	Obligaciones de transparencia	Los chatbots deben decir que estás hablando con una máquina; el contenido generado o manipulado (*deepfakes*) se debe etiquetar
Mínimo	Sin obligaciones específicas	Filtros de spam, IA en videojuegos

Los grandes modelos de lenguaje también tienen obligaciones propias, como **modelos de IA de uso general**: documentación, transparencia y respeto de los derechos de autor, con más obligaciones para los más grandes.

Qué se aplica y desde cuándo

Fecha	Qué empieza a aplicarse
1 ago 2024	Entrada en vigor
2 feb 2025	Prácticas prohibidas y alfabetización en IA
2 ago 2025	Gobernanza y obligaciones de los modelos de IA de uso general
2 ago 2026	Aplicación general, incluidas las obligaciones de transparencia
2 dic 2027	Sistemas de alto riesgo como la contratación, el crédito o la educación (aplazado por el *Ómnibus de IA*)
2 ago 2028	Sistemas de alto riesgo integrados en productos regulados

> **El calendario ya ha cambiado una vez**
> Las fechas del alto riesgo se fijaron primero para 2026 y 2027. Una modificación, el **Ómnibus de IA**, entró en vigor el **27 de julio de 2026** y las retrasó. Cuando escribas sobre el Reglamento de IA, comprueba las fechas vigentes en la página de la Comisión Europea y cítala: las leyes cambian. Y no lo confundas con el **RGPD** (Reglamento (UE) 2016/679), la ley de protección de datos que verás en la Unidad 5.

En clase: ¿qué nivel de riesgo?

En vuestros grupos de proyecto, clasificad cada caso y justificadlo en una frase. Después clasificad las dos ideas de IA que anotasteis para vuestra empresa en la lección anterior.

Caso	Vuestro nivel
Un chatbot que responde las preguntas de los clientes en la web de un hotel	
Un programa que ordena currículums y descarta candidatos	
El filtro de spam de una empresa de transporte	
Un sistema de una administración pública que puntúa a la ciudadanía según su comportamiento	
Una IA que corrige exámenes y decide quién entra en un curso	
Un vídeo promocional con la voz de una persona generada con IA	
Cámaras que detectan el estado de ánimo de los trabajadores de una oficina	
Un modelo que decide si un banco concede un préstamo	

Entrega en Classroom (por parejas, «S16 · Cazadores de errores y niveles de riesgo», antes de las 23:59 del mismo día): vuestra tabla de errores con el enlace de cada comprobación, las cinco normas adaptadas a vuestra empresa y la clasificación de los ocho casos y de vuestras dos ideas. Recuerda: la autorización de la IA terminó con la caza de errores.

Para tu proyecto

Esta lección alimenta la **parte de grupo** de la Entrega 4 · Plan de IA:

Trabajo de hoy	Se convierte en	Criterio
Vuestras cinco normas, adaptadas a la empresa	Parte de grupo: normas de uso responsable	c
El nivel de riesgo de vuestras propias ideas	Parte de grupo: el nivel de riesgo de cada propuesta según el Reglamento de IA	c
Qué puede hacer y qué no la IA generativa	Sección C: presente y futuro de la IA en el sector	c, f

Ideas clave

- La **IA generativa** crea contenido; un **LLM** escribe prediciendo el siguiente fragmento de texto, una y otra vez.
- Escribe igual de bien las respuestas correctas y las equivocadas: las **alucinaciones** llegan sin avisar. Contrasta siempre con fuentes.
- Uso responsable: **revísalo**, **protege los datos**, **sé transparente**, **el responsable eres tú**, **respeta la autoría**.
- El **Reglamento de IA** (Reglamento (UE) 2024/1689) sigue un **enfoque basado en el riesgo**: inaceptable, alto, limitado (transparencia), mínimo.
- Las prohibiciones y la alfabetización en IA se aplican desde el 2 de febrero de 2025; la transparencia, desde el 2 de agosto de 2026; las obligaciones del alto riesgo se aplazaron a diciembre de 2027 y agosto de 2028.
- En este módulo, la IA solo está permitida cuando el profesor la autoriza para una actividad concreta.

Comprobación rápida

1 ¿Qué hace básicamente un gran modelo de lenguaje cuando responde?

A Busca la respuesta en una base de datos de hechos verificados
B Predice, fragmento a fragmento, la continuación más probable del texto
C Busca en internet y copia la mejor página
D Ejecuta reglas escritas por expertos para cada pregunta

2 Un chatbot te da una cita de un informe, con el título y el año. Buscas el informe y no existe. ¿Cómo se llama esto?

A Un token
B Un prompt
C Una alucinación
D Una fecha de corte del conocimiento

3 ¿Qué dos acciones siguen las normas de uso responsable?

A Pegar una lista de clientes con sus teléfonos en un chatbot público para escribir correos más deprisa
B Decirle a tu equipo que un primer borrador se escribió con una herramienta de IA aprobada por la empresa
C Probar una función generada por un asistente antes de añadirla al proyecto
D Publicar una respuesta de la IA tal cual, porque la responsable es la herramienta

4 Según el Reglamento de IA, ¿qué nivel tiene un programa que ordena currículums y descarta candidatos?

- **A** Inaceptable: está prohibido
- **B** Riesgo alto: permitido con obligaciones estrictas como la supervisión humana
- **C** Limitado: solo tiene que decir que es una IA
- **D** Mínimo: sin obligaciones

5 Un hotel pone un chatbot en su web. ¿Qué es lo mínimo que exige el Reglamento de IA?

- **A** Nada: los chatbots son de riesgo mínimo
- **B** Una licencia de la Comisión Europea
- **C** Informar a los usuarios de que están hablando con una IA
- **D** Apagarlo por la noche

6 ¿Cuál de estas es una práctica prohibida según el Reglamento de IA?

- **A** Un filtro de spam
- **B** Un videojuego con enemigos controlados por IA
- **C** Cámaras que reconocen las emociones de los trabajadores de una oficina
- **D** Un recomendador de una tienda en línea

7 Tu grupo propone una IA que predice qué máquinas de una panadería se averiarán la semana que viene, a partir de los datos de sus sensores. ¿Qué nivel de riesgo encaja mejor, y qué debéis hacer igualmente?

- **A** Inaceptable; debéis abandonar la idea
- **B** Probablemente mínimo: no decide nada sobre personas, pero una persona debería seguir revisando sus avisos
- **C** Alto, porque usa sensores
- **D** Limitado, así que debe anunciar a las máquinas que es una IA

Ya sabes cómo escribe un modelo generativo, por qué lo revisas y qué le pide Europa. Ahora cambia de lado: eres la empresa y quieres saber si la IA sale a cuenta. Siguiente: Automatización y rentabilidad.

Soluciones

1. B · Un **LLM** es un predictor de texto. Por eso puede escribir una respuesta convincente que es falsa: nada de la predicción comprueba los hechos.

2. C · Una respuesta inventada pero convincente es una **alucinación**. Las fuentes inventadas son una de sus formas más habituales; por eso siempre abres los enlaces.

3. B, C · **Sé transparente** y **revísalo**. Los datos personales nunca van a herramientas no aprobadas, y la persona que publica es la responsable, no la herramienta.

4. B · Las decisiones de contratación afectan a la vida de las personas, así que el Reglamento de IA las clasifica como **riesgo alto**: gestión de riesgos, calidad de los datos, documentación y **supervisión humana**.

5. C · Los chatbots están en el nivel **limitado (transparencia)**: las personas deben saber que hablan con una máquina. Esas obligaciones de transparencia se aplican desde el **2 de agosto de 2026**.

6. C · El reconocimiento de emociones **en el trabajo o en la escuela** es una práctica de riesgo **inaceptable** (salvo por motivos médicos o de seguridad), prohibida desde el 2 de febrero de 2025.

7. B · Una predicción de mantenimiento no decide sobre los derechos de las personas, así que encaja en el riesgo **mínimo**. El uso responsable sigue valiendo: alguien lo **revisa** antes de actuar, y los datos deben ser buenos.