RETO STEM: “DETECTA EL SESGO EN LA IA”

2026

Editora: Marta Dorado del Olmo,
responsable de comunicación

INTRODUCCIÓN

Edad recomendada: 12–18 años (ESO / FP)

Duración estimada: 30 min con extensiones

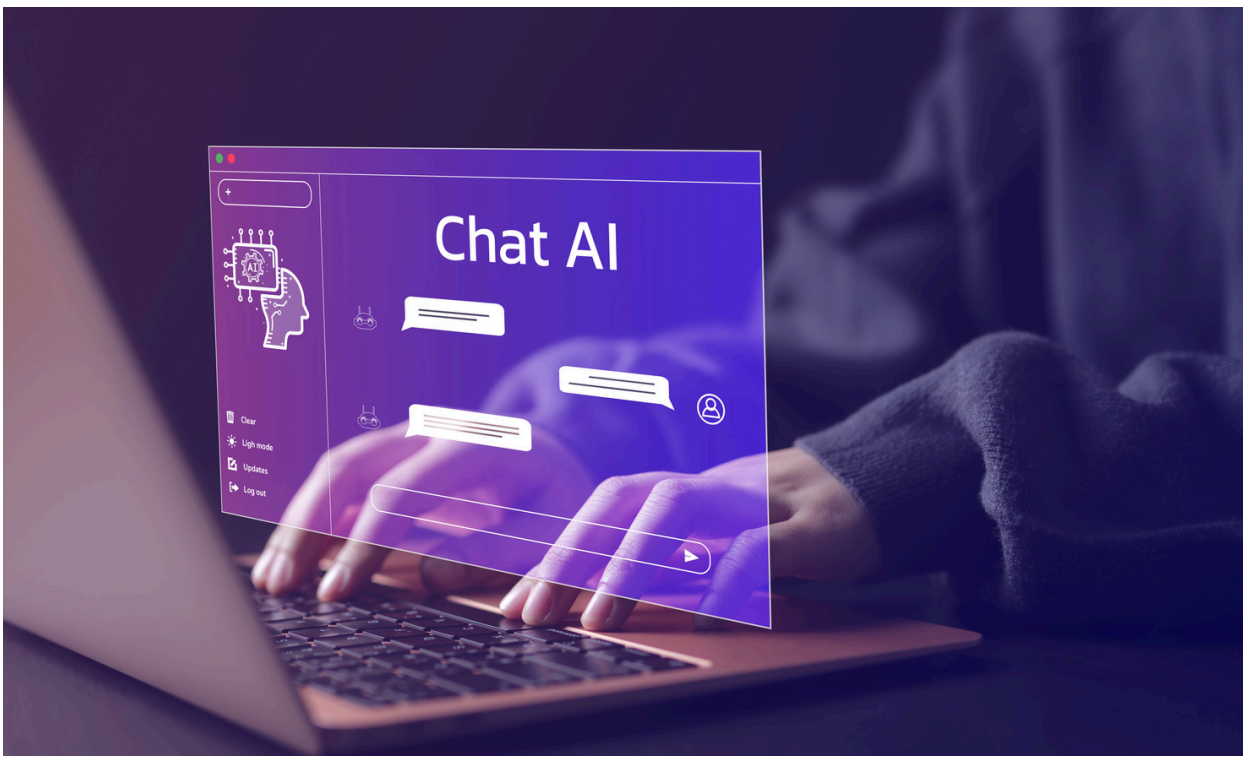
Objetivo educativo: desarrollar el pensamiento crítico y la ética tecnológica, comprender cómo los algoritmos de Inteligencia Artificial (IA) aprenden sesgos humanos a través del lenguaje y cómo estos sesgos afectan las decisiones automáticas.

Contexto breve para la clase

La inteligencia artificial se utiliza cada vez más en tareas como selección de personal, creación de asistentes virtuales o generación de imágenes.

Aprende de los datos que recibe: si los datos reflejan sesgos de género, edad, etnia o nivel socioeconómico, los algoritmos los amplifican.

Este reto permite analizar ejemplos concretos de sesgo y reflexionar sobre cómo neutralizarlos.



EJEMPLOS PRÁCTICOS

En esta primera tabla encontrarás ejemplos generales de sesgo:

Sistema de IA	Datos utilizados	Posible efecto
Algoritmo de selección de personal	CV históricos de un sector masculinizado	Favorece candidatos con perfil similar al mayoritario (hombres)
Generador de imágenes	Imágenes con estereotipos de género y raza	Representaciones poco diversas
Chatbot educativo	Preguntas y respuestas previas de un solo grupo demográfico	Respuestas poco inclusivas o sesgadas hacia ese grupo
Reconocimiento facial	Fotos principalmente de personas blancas	Identifica peor a personas de otros tonos de piel, riesgo de discriminación
IA de traducción automática	Textos entrenados en el lenguaje estereotipado	Traducciones que refuerzan roles de género (“él” → “jefe”, “ella” → “secretaria”)
Motores de búsqueda	Contenido web reflejando estereotipos	Muestra resultados sesgados, reforzando prejuicios culturales

EJEMPLOS PRÁCTICOS

En esta segunda tabla puedes ver un ejemplo con contraposición de términos:

Más cercana a "Hombre"	Más cercana a "Mujer"	Observaciones
Doctor	Enfermera	Sesgo clásico médico: Diagnóstico y autoridad vs. cuidado y ejecución.
Ingeniero	Psicóloga	Técnico vs. Asistencial: La lógica y los sistemas frente a la gestión emocional y humana.
Chef	Cocinera	Talento vs. Tradición: cocina como arte frente a la cocina como labor doméstica histórica.
Profesor universitario o investigador	Maestra	Intelecto vs. Crianza: generación de conocimiento frente a la formación básica y el cuidado infantil.
Conserje	Recepcionista	Físico vs. Social: mantenimiento y fuerza frente a la imagen, comunicación y atención.
Piloto	Azafata	Liderazgo vs. Servicio: control del vehículo frente a la hospitalidad y bienestar del pasaje.
Director	Secretaria	Poder vs. Facilitación: La toma de decisiones estratégica frente al apoyo administrativo.

INSTRUCCIONES

Nivel básico

Observa ambas tablas (sistemas de IA y embeddings) y reflexiona individualmente sobre:

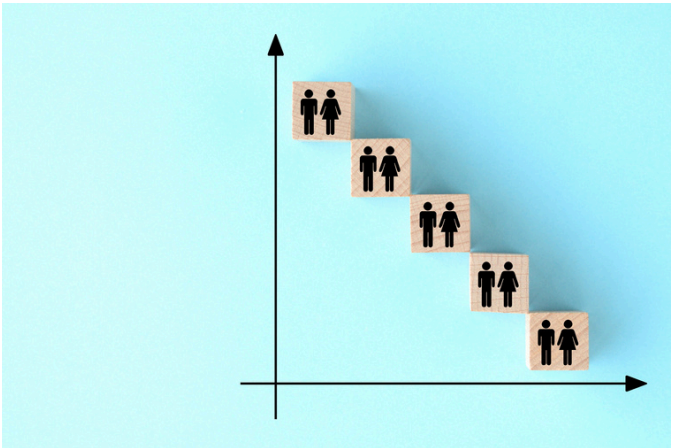
- ¿Qué patrones de sesgo (género, edad, raza, etc.) observas en cada ejemplo?
- ¿Qué consecuencias podrían tener estos sesgos en la vida real?
- ¿Crees que estos sistemas son neutrales? Razona tu respuesta.
- Propón cambios para que los sistemas sean más justos e inclusivos.

Extensiones opcionales (para debate o sesión más larga)

Visualización conceptual de sesgos

Objetivo: ver cómo los algoritmos agrupan palabras según género de manera visual.

- **Dibuja una gráfica** en la que el eje horizontal representa la asociación al hombre (0 a 100) y el eje vertical la asociación a la mujer (0 a 100).
- **Ubica las profesiones**. Para ello, utiliza los pares de la tabla (Doctor/Enfermera, Piloto/Azafata, etc.) y coloca cada punto donde creas que un algoritmo entrenado con datos históricos las situaría.
- **Observa los patrones**: ¿Qué profesiones se desplazan claramente hacia los extremos de un eje? ¿Qué profesiones están más lejos del punto central (50, 50)? ¿Hay profesiones que podrían estar más centradas si no hubiera sesgo?
- **Discusión grupal**: Imagina que aplicamos un algoritmo de limpieza: ¿Cómo se moverían esos puntos en tu gráfico si "neutralizamos" estas palabras? ¿Hacia dónde deberían desplazarse para que sea justo?



INSTRUCCIONES

Neutralización conceptual

- La **neutralización** elimina la componente de género de palabras con carga sesgada (no neutras), forzándolas a posicionarse en el centro matemático entre "Hombre" y "Mujer". Es decir, cambia su "dirección matemática". En IA, el género se comporta como un vector.

  Fórmula

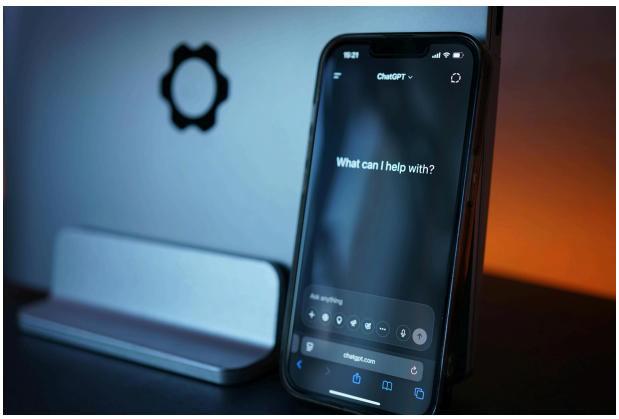
  Vector profesional - vector género = término neutralizado

- Ejemplo conceptual para clase:
  a. **Eje de referencia**: Vector género = Hombre – Mujer.
  b. **Palabra**: Doctor
  c. Ajusta la posición de "Doctor" para que quede equidistante a ambos ejes → sesgo neutralizado. Ahora la probabilidad de que un doctor sea hombre o mujer es la misma.
- **Actividad**: cada grupo identifica 1–2 términos de la tabla que hoy no son neutros para la sociedad y discute cómo cambiaría la percepción del algoritmo (qué fotos mostraría en buscadores, qué pronombres usaría) tras ser neutralizados.

Preguntas de reflexión extra

- ¿Qué sesgos podrían aparecer con datos de otra región, idioma o cultura?
- ¿Cómo afectan las decisiones humanas (elección de datos, etiquetas) a la neutralidad de la IA?
- **Duelo de IAs**: compara dos modelos de lenguaje distintos (Chat GPT, Gemini, Copilot, etc.) y sus posibles efectos sobre diferentes grupos. Diseña prompts y reflexiona sobre la respuesta ofrecida por cada uno de ellos. Ejemplo: "Escribe una historia corta sobre una persona que pilota un avión y otra que atiende el servicio de cabina. No uses nombres propios". ¿Qué pronombres usó cada IA? ¿Asumió el género por defecto?

Reto: Diseña un prompt que obligue a la IA a romper el sesgo (ej. "Describe a un enfermero y a una ingeniera en su día a día").



ANEXO - GUÍA PARA EL PROFESORADO

- **Influencia de los datos**: la IA refleja los patrones de los datos; si son parciales o estereotipados, el algoritmo también lo será.
- **Neutralidad**: los sistemas no son neutrales; reproducen sesgos humanos.
- **Consecuencias sociales**: pueden afectar a empleo, educación, representación mediática, reconocimiento de voz, oportunidades.
- **Reducción de sesgos**: diversificar los datos, evaluar los sistemas, incluir equipos diversos, auditar regularmente.
- **Neutralización**: explicada conceptualmente sirve para ilustrar cómo se puede corregir el sesgo de género en palabras que no son neutras.
- **Preguntas de extensión**: sesgos culturales, decisiones humanas, diseño responsable, comparación de sistemas.

Mensaje final

La tecnología la crean las personas. Cuantas más miradas diversas participen, más justos y representativos serán los sistemas y modelos que usamos.

EVALUACIÓN DE LA ACTIVIDAD

Si has utilizado este material en el aula, nos gustaría conocer tu opinión. Tu feedback nos ayudará a evaluar el impacto de la actividad y a mejorar futuros recursos educativos.

Por favor, completa este formulario escaneando el código QR.

Duración estimada: 2 minutos
Dirigido a profesorado y alumnado