

Cuando una respuesta convincente no basta: IA, sesgos y pensamiento crítico

Mariana Ramirez
Universidad de los Andes

Foro EMAD 2026

24 de octubre de 2026

- ▶ En esta charla
 - ▶ ¿Qué es la inteligencia artificial?
 - ▶ Conocer y reflexionar sobre
 - ▶ Limitaciones y sesgos
 - ▶ Riesgos
 - ▶ Aspectos éticos
 - ▶ Herramientas para el resto de las charlas

¿Qué es la inteligencia artificial?

- ▶ Campo o rama de la informática - tareas que normalmente asociamos con la inteligencia humana
 - ▶ Reconocer imágenes
 - ▶ Tomar decisiones
 - ▶ Generar imágenes
 - ▶ Seguir una conversación

- ▶ Modelo discriminativo - no generativo

- ▶ Clasificar o etiquetar



- ▶ Modelo Generativo

- ▶ Crear datos a partir de lo que han aprendido



¿Cómo da una respuesta?

- ▶ Modelos que generan texto
 - ▶ Entrenamiento - identificar patrones de cómo se combinan las palabras, qué palabras suelen seguir a otras, cómo se estructura una idea, etc
 - ▶ Predecir qué fragmento de texto es más probable que siga a otro

¿Cómo da una respuesta?

- ▶ Modelos que generan texto
 - ▶ No recuperan una verdad comprobada ni razonan desde experiencia humana
 - ▶ Generan la continuación que parece más adecuada según patrones aprendidos y el contexto de nuestra pregunta
 - ▶ La fluidez del lenguaje puede hacernos atribuir comprensión, intención o certeza
 - ▶ Puede sonar convincente y, aun así, estar equivocado

¿Cómo funciona?

- ▶ Menor conocimiento y mayor confianza en la IA, conlleva a más delegación y menor reflexión crítica (Klein-Avraham y Baram-Tsabari, 2026)
- ▶ Mayor alfabetización en IA, lleva a seguir menos los consejos de IA generativa, especialmente cuando son defectuosos
 - ▶ Aceptar la ayuda cuando tenemos razones para considerarla confiable y cuestionarla cuando la tarea exige verificación

Limitaciones y sesgos

Importancia de la reflexión al usarlo

Limitaciones y riesgos

- ▶ El objetivo básico es producir texto probable, no demostrar que cada afirmación es verdadera
 - ▶ Alucinaciones
 - ▶ Factuales - Datos falsos o inventados
 - ▶ Fieles - No relacionada o irrelevante
 - ▶ De interpretación - No entiende la pregunta o imagen
 - ▶ ¿Qué hacer?
 - ▶ Pedir fuentes y abrirlas
 - ▶ Usar nuestra experticia en el tema

Limitaciones y riesgos

- ▶ Los datos de entrenamiento contienen vacíos, errores, contradicciones y conocimiento desactualizado.
 - ▶ Sesgos
 - ▶ Histórico o social - Desigualdades y estereotipos existentes
 - ▶ Selección - No tiene datos suficientes
 - ▶ Retroalimentación - Asume como correcta una respuesta favorecida
 - ▶ ¿Qué hacer?
 - ▶ Medir que información le damos y preguntarnos si es necesaria

► Diseño del modelo

- ▶ Entrenado para ser útil, agradable y responder a preferencias humanas
 - ▶ Complacencia - confirmar o suavizar la contradicción
 - ▶ Siempre da una respuesta - presión por se útil
 - ▶ Personalización inapropiada
 - ▶ Pérdida de información en conversaciones largas
- ▶ ¿Qué hacer?
 - ▶ Pedir poca complacencia
 - ▶ Revisar si mi forma de interactuar esta sesgada “¿por qué esto es cierto?” a “¿esto es cierto?”

Algunos riesgos (Proceso cognitivo)

- ▶ Pérdida de habilidades de razonamiento crítico
 - ▶ Estudiantes que delegan a la IA tienen peores resultados (Bastani et al. 2025)
 - ▶ Excepto cuando se usa como tutor
 - ▶ Personas que usan mucho la IA tienen peores resultados en pruebas de razonamiento cognitivo (Gerlich, 2025)
 - ▶ Factores de edad y nivel de escolaridad
- ▶ Fomentar un escepticismo saludable motiva a los estudiantes a verificar los resultados por sí mismos

- ▶ Depende de nuestra experiencia con el tema
- ▶ Los estudiantes van a usarlo
- ▶ Democratiza el acceso al conocimiento
- ▶ Nuevas oportunidades para el docente
 - ▶ Más tiempo para acompañar y orientar
 - ▶ Herramienta para potenciar la enseñanza y aprendizaje
- ▶ Cuidado con ampliar las brechas

