

Evaluación de la información generada por la IA

Desarrollar habilidades para evaluar de forma crítica la precisión, la fiabilidad y el sesgo de los contenidos generados por la IA



Objetivos del aprendizaje En pocas palabras



- Identificar los tres pilares fundamentales de la evaluación de contenidos generados por IA: precisión, fiabilidad y sesgo (ARB).
- Aplicar marcos de verificación estructurados (como CRAAP o ROBOT) adaptados específicamente a los resultados de la IA generativa.
- Reconocer los modos específicos de fallo de la IA, incluyendo las alucinaciones, la fabricación de citas y el sesgo cultural.
- Desarrollar actividades prácticas en el aula que transformen los errores de la IA en ejercicios de pensamiento crítico.
- Implementar protocolos de transparencia que exijan a los estudiantes documentar y criticar su uso de las herramientas de IA.

La urgencia de una alfabetización crítica en materia de IA

Conceptos clave y definiciones

- **Cambio en el consumo de información:** los estudiantes utilizan la IA como principal motor de búsqueda y herramienta de redacción, a menudo sin pasar por los procesos tradicionales de verificación.
- **La trampa de la «plausibilidad»:** los modelos de lenguaje grande (LLM) están diseñados para sonar convincentes, no necesariamente para ser veraces. Dan prioridad a la probabilidad lingüística sobre la precisión factual.
- **Repercusión en las humanidades:** la IA suele tener dificultades con los matices, el contexto cultural y la profundidad interpretativa, componentes fundamentales de los estudios en humanidades.
- **Integridad académica 2.0:** La integridad ya no se limita a «no copiar»; consiste en verificar la veracidad de las afirmaciones de las herramientas que utilizamos.
- **Empoderamiento:** La evaluación de la enseñanza cambia el enfoque de «vigilar a los estudiantes» a «empoderar a los académicos» para que utilicen las herramientas de forma responsable.



La necesidad fundamental

De contar con competencias en evaluación de IA

EL RETO:

- Las herramientas de IA generan textos convincentes y fluidos que pueden ocultar imprecisiones y sesgos.
- Los estudiantes recurren cada vez más a la IA para investigar, redactar y analizar sin verificar la información.
- Los métodos tradicionales de detección de plagio resultan insuficientes en la era de la IA

LA OPORTUNIDAD:

- La evaluación crítica transforma la IA de una amenaza a una poderosa herramienta pedagógica.
- Los docentes de humanidades se encuentran en una posición única para enseñar a realizar evaluaciones matizadas.
- Desarrollar habilidades de evaluación prepara a los estudiantes para un mundo profesional integrado con la IA.

LA NECESIDAD:

- La integridad académica depende de información verificable y fiable.
- El uso ético de la IA requiere transparencia y compromiso crítico.





Los tres pilares: **precisión, fiabilidad, sesgo**

EXACTITUD

(Verificación de datos)

- ¿Se pueden contrastar las afirmaciones concretas con fuentes académicas reconocidas?
- ¿Presenta el contenido incoherencias lógicas?
- ¿Inventa la IA datos, fechas o acontecimientos (alucinaciones)?

FIABILIDAD

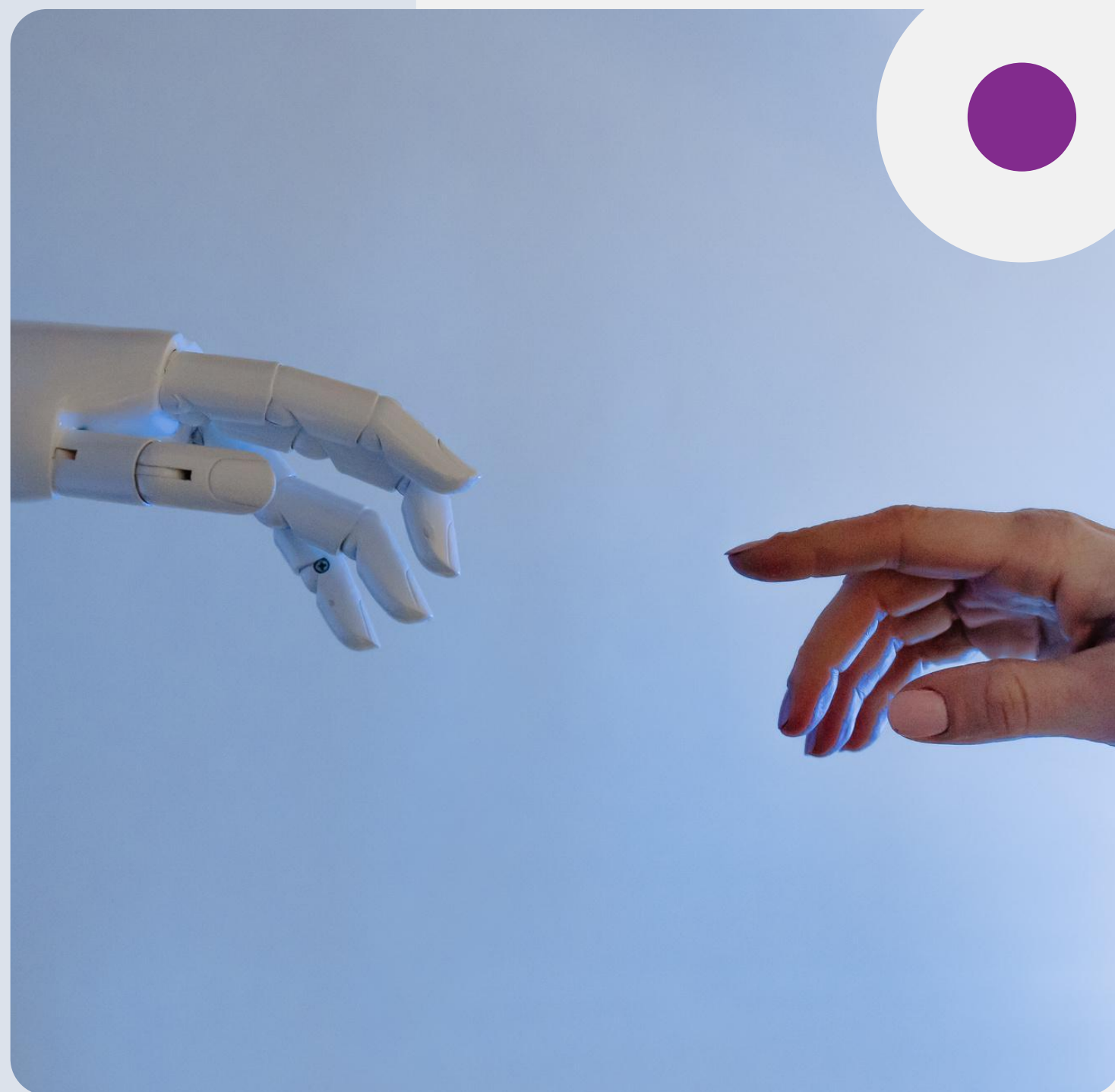
(Verificación de la fuente)

- ¿Puede la IA proporcionar citas reales y verificables? (A menudo no puede).
- ¿El resultado es coherente en diferentes solicitudes o modelos?
- ¿La IA proporciona fuentes transparentes?

SESGO

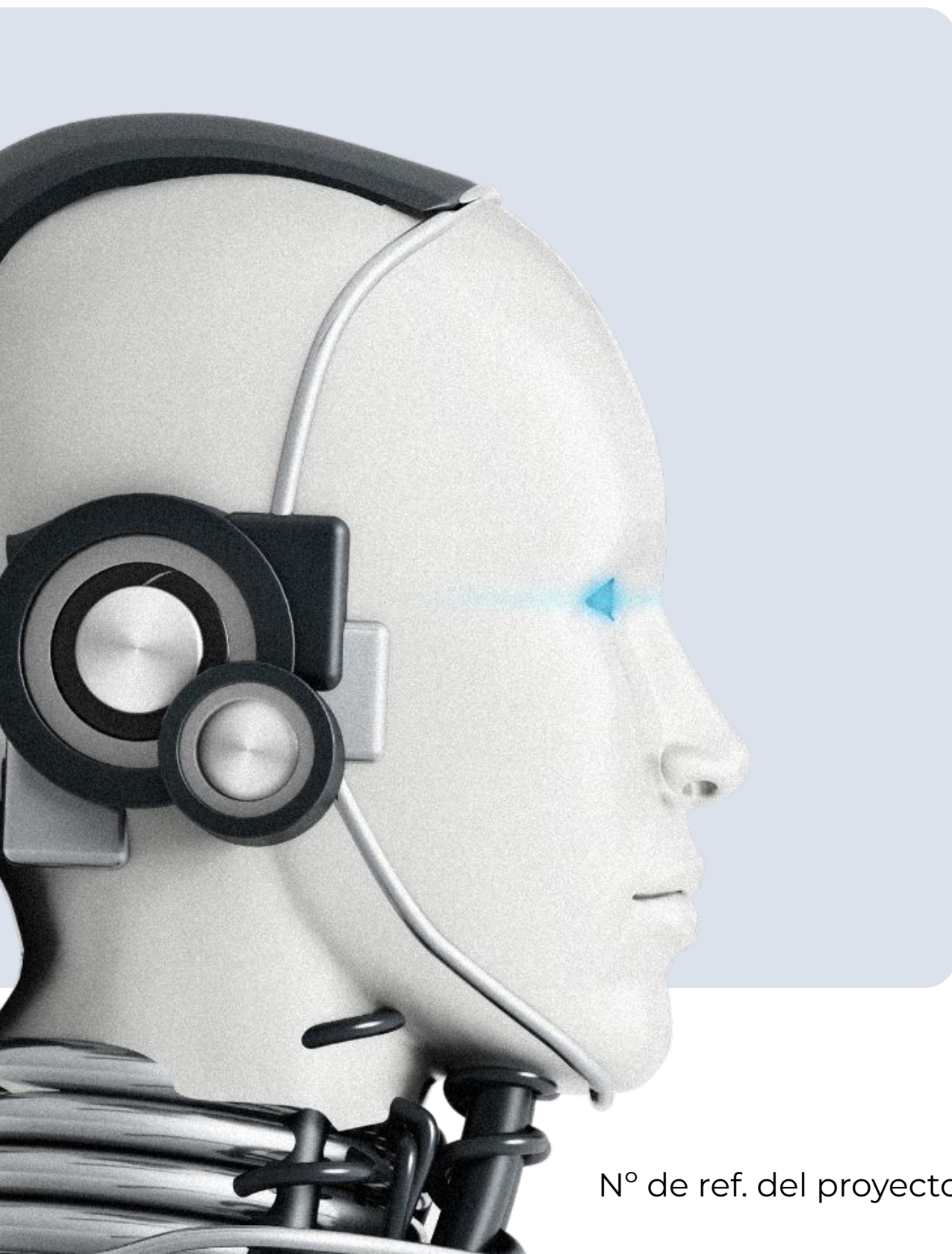
(La comprobación de la perspectiva)

- ¿Qué voz predomina? (Es habitual el sesgo occidental o anglocéntrico).
- ¿Qué voz falta? (Grupos marginados, estudios académicos no anglosajones).
- ¿El tono es objetivo o contiene estereotipos implícitos?



Análisis de las Limitaciones de la IA

- **Alucinaciones:** la generación segura de información falsa. La IA predice la siguiente palabra, pero no «conoce» los hechos.
- **Fabricación de citas:** la IA puede generar títulos de libros que parecen reales, escritos por autores reales, pero que en realidad no existen.
- **Lagunas de conocimiento:** es posible que los modelos no tengan constancia de acontecimientos ocurridos en los últimos 6-12 meses.
- **Colapso del contexto:** la IA a menudo simplifica debates complejos en resúmenes genéricos, perdiendo la «textura» del discurso de las humanidades.
- **Adulación:** la IA tiende a estar de acuerdo con la premisa del usuario, incluso si esta es errónea o incorrecta.



Actividad interactiva:

Descubre la alucinación

SITUACIÓN: Un estudiante entrega un trabajo sobre literatura polaca en el que cita:

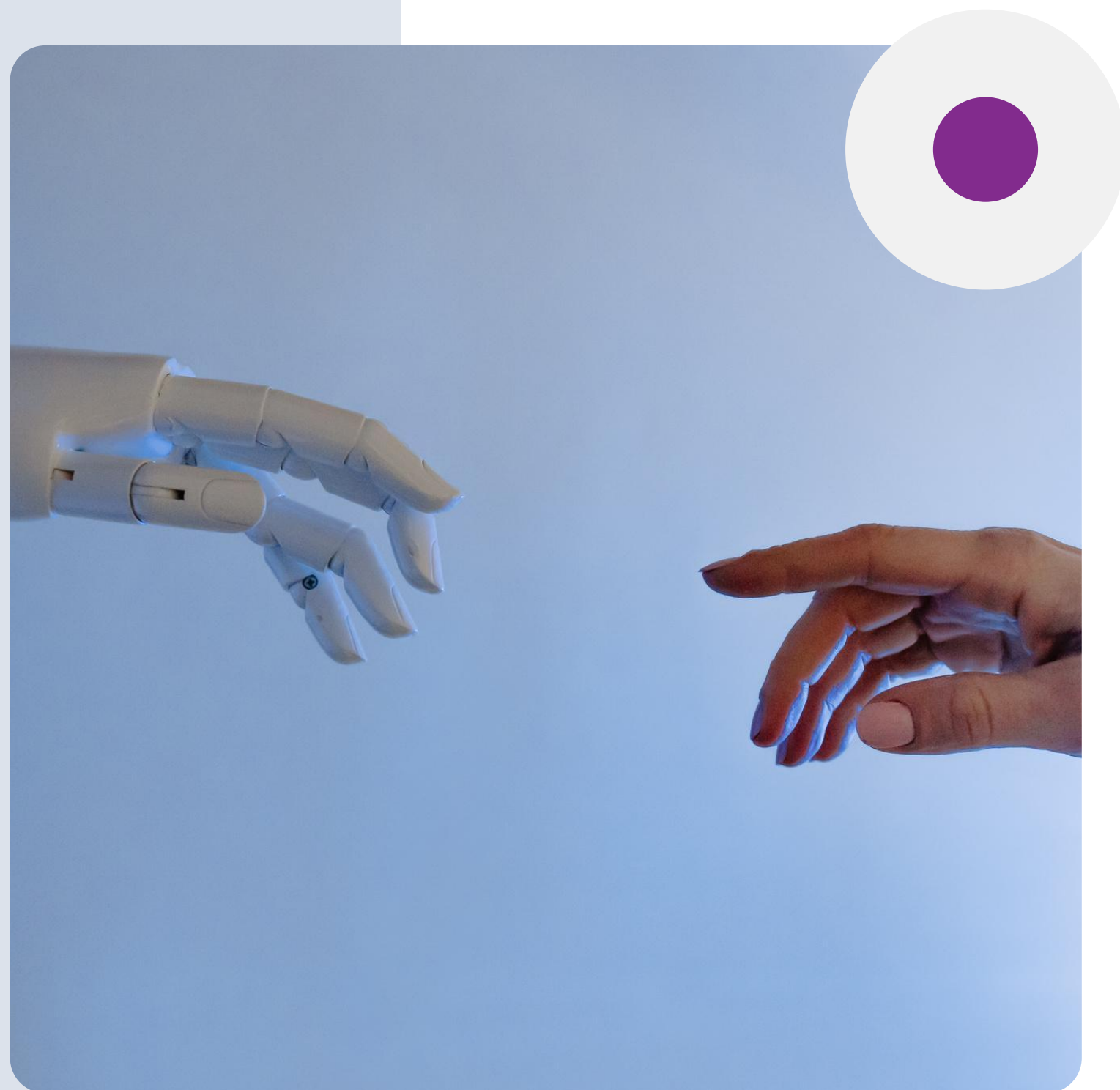
«Kowalski, J. (2019). La metáfora cuántica en las primeras obras de Tokarczuk.
Editorial de la Universidad de Varsovia».

ANÁLISIS:

- El título: Suena académico y verosímil.
- El autor: «Kowalski» es un nombre común, lo que dificulta su rápida verificación.
- La editorial: Una institución real.
- LA REALIDAD: Este libro no existe. La IA combinó palabras clave reales para inventar una fuente.

ACCIÓN:

- ¿Cómo verificarías esto? (Google Scholar, catálogo de la biblioteca, búsqueda por DOI).
- ¿Qué comentarios le darías al estudiante? (Céntrate en el proceso de verificación, no solo en el error).

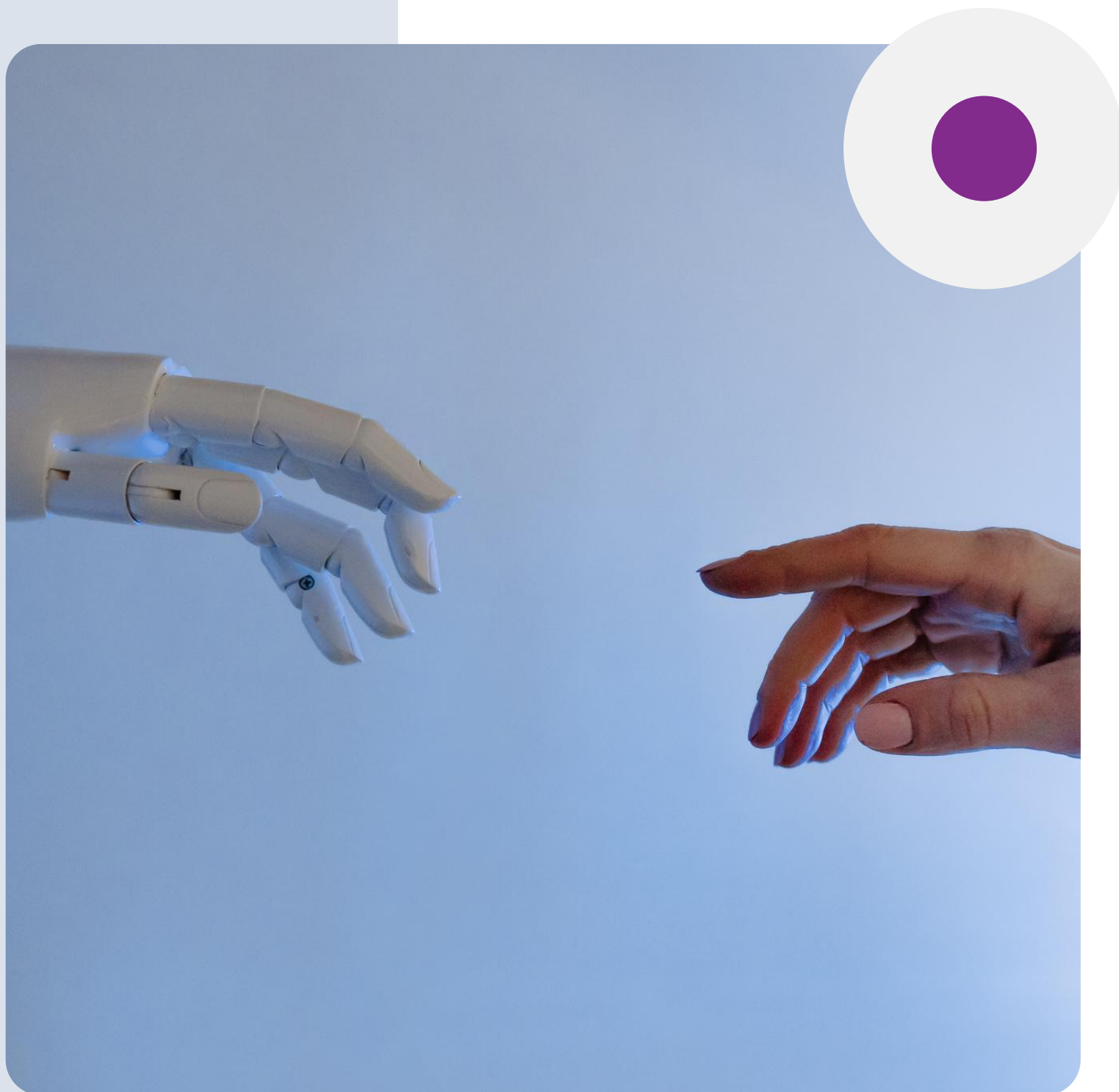


Marcos: adaptación De CRAAP a la IA

La prueba CRAAP tradicional frente a las necesidades de la era de la IA:

- **Actualidad:** los datos de la IA suelen estar desactualizados. Comprueba la fecha de «corte de conocimientos» del modelo.
- **Relevancia:** la IA responde a la pregunta, pero ¿responde a la pregunta de investigación? (Cumplimiento de la pregunta frente a profundidad intelectual).
- **Autoridad:** la IA NO tiene autoridad. Es un agregador. La comprobación de la «autoridad» debe pasar a verificar las fuentes *originales* que la IA afirma utilizar.
- **Exactitud:** Requiere verificación externa (triangulación con textos fiables).
- **Propósito:** El propósito de la IA es satisfacer al usuario, no buscar la verdad. Hay que ser consciente de su tendencia a complacer.

NUEVA MÉTRICA: Reproducibilidad. ¿Se puede obtener el mismo resultado de nuevo? Si no es así, no es una herramienta analítica fiable.



Estrategias de verificación Para el aula

1. **Lectura transversal:** No leas la página de arriba abajo; abre nuevas pestañas para verificar los datos inmediatamente.
2. **Triangulación:** Verifica cualquier afirmación de la IA con al menos dos fuentes distintas e independientes (por ejemplo, un texto primario y un artículo revisado por pares).
3. **La política de citas «cero confianza»:** Asume que toda cita generada por la IA es falsa hasta que se demuestre su veracidad mediante un DOI o un enlace a una biblioteca.
4. **Búsqueda inversa de imágenes:** en el caso de imágenes o gráficos generados por IA, comprueba si los datos existen en otros lugares.
5. **Auditoría de expertos:** anima a los estudiantes a utilizar la IA en un tema que *ya* conozcan bien para detectar los errores sutiles (el «método sándwich»: conocimiento humano -> borrador de IA -> crítica humana).

Actividad interactiva

Detectando sesgos

SITUACIÓN: Le pides a la IA «Los 10 filósofos más influyentes del siglo XX».

RESULTADO: La lista incluye a 9 hombres y 1 mujer. Todos son europeos o estadounidenses..

ANÁLISIS:

- Sesgo de representación: ¿Por qué se excluye a los pensadores no occidentales?
- Sesgo lingüístico: El modelo da prioridad a los datos de entrenamiento en inglés.
- Sesgo histórico: Reproduce la marginación histórica presente en el conjunto de datos.

SUGERENCIAS PARA SOLUCIONAR EL PROBLEMA:

- Indicación incorrecta: «Enumera filósofos influyentes».
- Buena indicación: «Enumera filósofos influyentes del siglo XX, asegurando un equilibrio de género e incluyendo perspectivas no occidentales de Asia, África y América Latina».



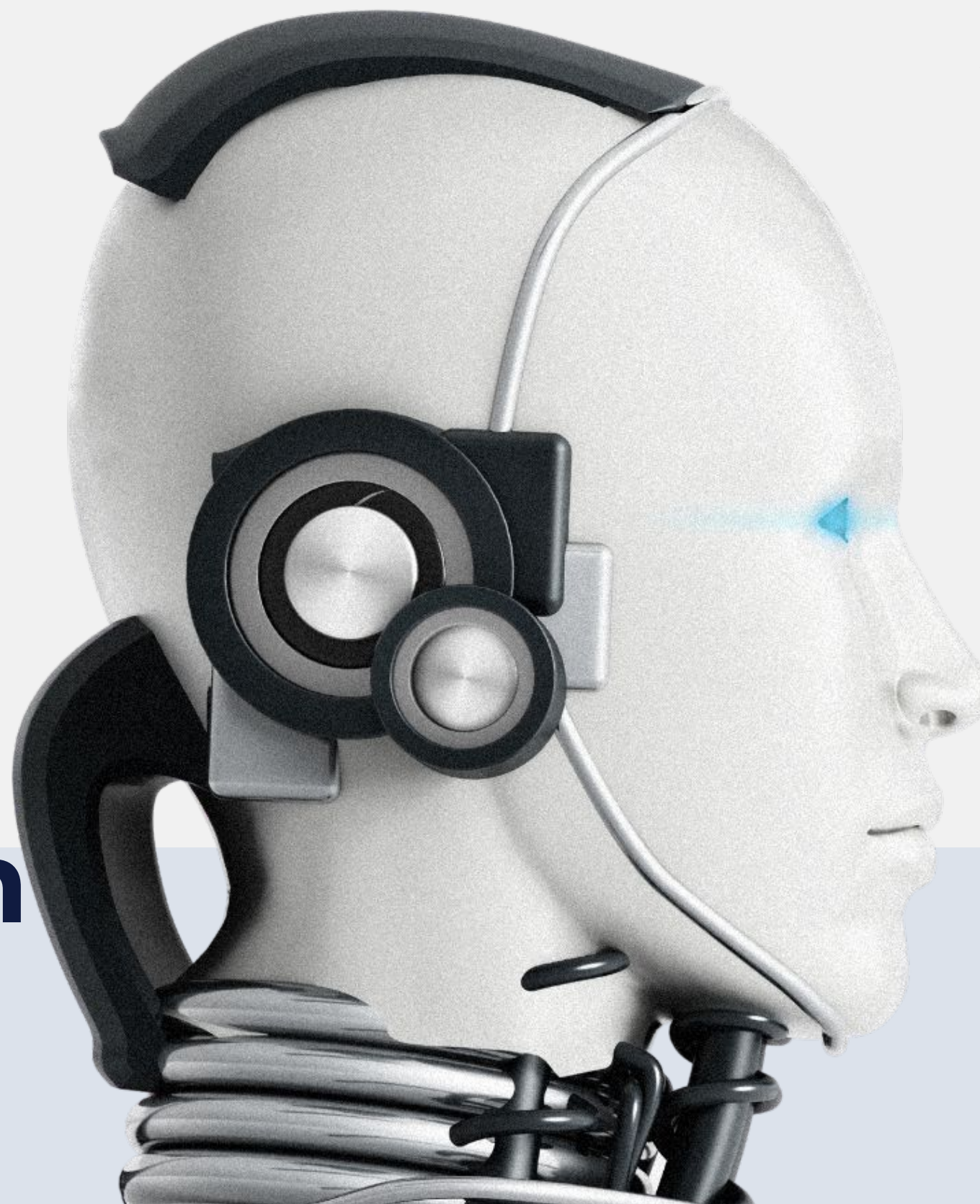
Estrategias pedagógicas

Enseñar el uso crítico de la IA

- **La tarea «Auditoría de IA»:** los alumnos generan un ensayo mediante IA sobre un tema y, a continuación, redactan una crítica en la que identifican todos los errores, sesgos y lagunas.
- **Documentación del proceso:** exigir a los alumnos que entreguen sus registros de chat. Calificar la *calidad de sus indicaciones* y su *verificación* del resultado.
- **Debates en clase:** humanos contra IA. Un equipo debate un tema; el otro equipo utiliza IA. Comparar la profundidad y los matices de los argumentos.
- **Rediseño de la evaluación:** Pasar de «resume este texto» (la IA lo hace bien) a «critica este caso práctico local específico utilizando esta teoría concreta» (la IA tiene dificultades en este aspecto).
- **Rúbricas transparentes:** Añadir una fila en la rúbrica para la «Evaluación crítica de las fuentes», premiando explícitamente la corrección de los errores de la IA.

Trust AI

Lista de verificación para la evaluación



Antes de utilizar contenido generado por IA, pregúntate:

- VERIFICABILIDAD: ¿Hay datos o fechas concretos que pueda comprobar?
- FUENTES: ¿He encontrado la fuente primaria de cada cita?
- PERSPECTIVA: ¿He solicitado explícitamente puntos de vista diversos?
- ACTUALIDAD: ¿Es este tema reciente? (Si es así, sé especialmente escéptico).
- LÓGICA: ¿Se sostiene el argumento o es solo retórica fluida?
- ÉTICA: ¿He revelado que he utilizado IA para este trabajo?
- TOQUE HUMANO: ¿Requiere esto mi interpretación personal o mis conocimientos locales?

Puntos clave Y próximos pasos

1. La IA es una herramienta, no un oráculo: predice textos, pero no descubre la verdad.
2. La verificación es imprescindible: debemos pasar de ser consumidores de información a auditores de información.
3. La transparencia genera confianza: los debates abiertos sobre las limitaciones de la IA reducen la ansiedad y las conductas indebidas.
4. Las habilidades de las humanidades son habilidades para la IA: el contexto, el espíritu crítico y el razonamiento ético son precisamente lo que se necesita para gestionar la IA.



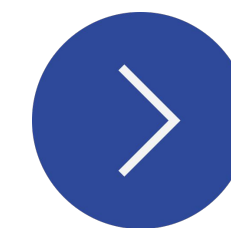
Participación interactiva

Comienza la actividad





Referencias



1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
2. Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2), e35179.
3. Caulfield, M. (2017). *Web Literacy for Student Fact-Checkers*. Retrieved from <https://webliteracy.pressbooks.com/>.
4. Borji, A. (2023). A Categorical Archive of ChatGPT Failures. *arXiv preprint arXiv:2302.03494*.
5. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779..
6. Teaching. *International Journal of Teaching and Learning in Higher Education*, 32(1), 39-48.



«El futuro de la educación no consiste en sustituir la mente humana, sino en ampliarla de forma responsable. Sigue cuestionando, sigue enseñando, sigue aprendiendo».

El consorcio



<https://trustai.unibit.bg/>

Trust AI project

Aviso legal: Financiado por la Unión Europea. No obstante, las opiniones y puntos de vista expresados son exclusivamente de los autores y no reflejan necesariamente los de la Unión Europea ni los de la Agencia Nacional (AN). Ni la Unión Europea ni la AN se hacen responsables de los mismos.

Nº de ref. del proyecto: 2024-1-BG01-KA220-HED-000249185



Co-funded by
the European Union

