1.3 Uso Responsable de la IA

Mensaje clave

- **La responsabilidad es una cuestión de diseño, no de control.** La detección de IA no funciona correctamente.
- **Y la detección de IA es injusta cuando falla.** Los detectores señalan erróneamente trabajos humanos como escritos por IA y pasan por alto trabajos escritos por IA. Las autoridades educativas de Francia y los Países Bajos desaconsejan su uso.
- **La invención con aparente seguridad y los sesgos heredados se derivan de cómo se construyen y evalúan estos sistemas.** La consecuencia debería ser: enseñarlos, en lugar de esperar una solución.
- **El principio de trabajo es: pensar primero, luego la IA.** Hacer que las personas se comprometan con su propia respuesta vence a explicar la máquina, e incorporar límites en el software elimina el daño que causa el acceso sin restricciones.
- **Los prompts por sí solos no bastan.** Un chatbot al que se le pide actuar como tutor tiende a volver a dar las respuestas directamente, lo que convierte los límites fiables en una cuestión de compra.
- **La amplificación es lo predeterminado; la equiparación debe construirse** a partir del acceso, el diseño y el acompañamiento juntos. Lo más escaso, pero también lo más importante de los tres, es un docente formado.
- **La inclusión es la ganancia más clara**, y donde el propio profesorado ve el argumento más convincente.

> Si solo recordáis una frase de esta página: los diques se están rompiendo. Enseñad a nadar.

El diario de lectura, una segunda mirada

Volvamos a nuestro estudiante de 3.º de la ESO y su diario de lectura sospechosamente pulido. El instinto es casi universal, y es el instinto de un buen docente: averiguar si usó IA. Construir un dique mejor.

Este capítulo trata de por qué ese instinto, por sí solo, no lleva a ninguna parte. Veremos a continuación qué funciona en su lugar.

Empecemos por la parte incómoda. **El profesorado no puede distinguir de forma fiable un texto de IA de un texto del alumnado.** En dos estudios de un equipo alemán, se dio a 89 docentes en formación y a 200 docentes con experiencia una mezcla de redacciones de estudiantes y de chatbot y se les pidió que las clasificaran. Ningún grupo lo logró de forma fiable, pero ambos estaban seguros de poder hacerlo. La experiencia ayudó solo un poco: el profesorado con experiencia siguió sin detectar la mayoría de los textos de IA. En una universidad británica, los investigadores enviaron respuestas generadas por chatbot a exámenes en línea reales y sin supervisión a través de los canales normales, sin avisar a quienes corregían.

«Podéis usarla, pero tenéis que ser capaces de explicar el resultado con vuestras propias palabras».

El detector que falla con el alumnado equivocado

Los investigadores probaron siete de los detectores de IA más usados con redacciones del examen TOEFL escritas por personas que aprendían inglés como lengua extranjera, recopiladas en un foro chino en línea. Junto a ellas, colocaron redacciones de estudiantes estadounidenses nativos de 8.º curso. El trabajo de los hablantes nativos se clasificó correctamente casi siempre. De las redacciones en lengua extranjera, **más de seis de cada diez fueron señaladas erróneamente como escritas por máquina**.

El mecanismo no es ningún misterio. Muchas de estas herramientas no detectan el trabajo de los grandes modelos de lenguaje. Lo que miden es la *previsibilidad*. Un texto que usa palabras comunes en el orden esperado puntúa como artificial. Quienes escriben en una segunda lengua, con un vocabulario más reducido y estructuras de frase más rígidas, pueden producir exactamente esa señal. Por otro lado, los detectores entrenados para una prueba concreta, así como los estudios en otros idiomas, muestran diferencias menores. Hasta ahora no hay en el mercado ningún detector lo bastante fiable como para acusar a nadie de engaño intencionado.

Los errores de la herramienta recaen sistemáticamente sobre el alumnado que ya carga con más dificultades. Las instituciones han hecho los números. La Universidad de Vanderbilt señaló que, aunque una herramienta así se equivocara solo el 1 % de las veces, eso seguiría significando unos **750 trabajos señalados erróneamente al año** de entre sus 75 000 entregas. En consecuencia, la desactivaron. Una guía elaborada por investigadores para el ministerio federal de educación de Alemania es tajante: la tecnología *«no es lo bastante fiable como para aportar una prueba jurídicamente segura»*. Una prueba europea independiente de catorce detectores llegó al mismo veredicto: *«ni precisos ni fiables»*.

Buena parte de Europa ya ha dejado de debatir esto. El ministerio de educación de Francia y su autoridad de protección de datos aconsejan a las escuelas no usar detectores, y la agencia neerlandesa de TIC Kennisnet los considera muy poco fiables. En la educación superior, algunas universidades, como la ZHAW de Zúrich, prohíben usarlos para calificar. En Inglaterra, los

organismos examinadores solo permiten el software de detección como un elemento más de un juicio que corresponde al docente, que es quien *«mejor conoce a su alumnado»*. También merece la pena llevarse a casa una cuestión de protección de datos: **pegar el trabajo de un estudiante en una herramienta en línea supone tratar sus datos personales.** Para ello se necesita una base jurídica y un contrato con el proveedor. → 1.4 Uso Conforme de la IA

Parece que el pánico por las trampas es más liviano que el ruido que lo rodea. Los investigadores encuestaron a estudiantes de tres institutos estadounidenses antes de que existiera ChatGPT y de nuevo después. Las trampas se mantuvieron, en líneas generales, donde estaban. Una encuesta de seguimiento a más de 4000 estudiantes de otros seis institutos estadounidenses, un año después, llegó al mismo resultado.

La verdadera pregunta nunca fue «¿cómo lo pillo?». Es: ¿qué sigue midiendo una tarea para casa una vez que su producto puede generarse a demanda?

«Al principio lo veía como hacer trampas. Ahora lo veo como una señal de que la propia tarea tiene que cambiar».

Dos cosas que no se van a arreglar sin más, así que hay que enseñarlas

La invención con aparente seguridad viene integrada de serie. Estos sistemas generan continuaciones plausibles de un texto, y plausible no es lo mismo que verdadero. Las revisiones de la investigación coinciden en que ninguna técnica ha eliminado todavía el problema, aunque puede reducirse y detectarse en parte. La razón es que las pruebas estándar usadas para evaluar los grandes modelos de lenguaje premian una suposición hecha con seguridad por encima de admitir incertidumbre, así que, mientras se midan así, persistirá una cierta tasa de falsedades fluidas y bien formadas.

Los sesgos del material de entrenamiento también vienen integrados. Y además no se reducen a medida que mejoran los modelos. Un estudio europeo sobre estereotipos de género en distintos idiomas encontró que los modelos más grandes, incluso tras el entrenamiento adicional pensado para hacerlos más seguros y justos, a veces estereotipan *más*, no menos. Si una propiedad no se puede eliminar mediante ingeniería, se convierte en materia de enseñanza. Y puede enseñarse: en un programa finlandés, el porcentaje de estudiantes capaces de explicar de dónde procedía el sesgo de un generador de imágenes pasó del 7 % al 44 %. La consecuencia para la enseñanza debe ser formar en el análisis del error en lugar de en evitar el error. Dicho de otro modo: los fallos de la máquina se convierten en la lección.

«Dejad que el alumno compruebe la solución del chatbot como si fuera el profesor. Buscando los errores es como por fin aprende el método».

Qué funciona: diseño, no control

La evidencia apunta a un principio sencillo y fácil de recordar: pensar primero, luego la IA.

- **Haced que la persona se comprometa primero.** En un experimento con unas 200 personas adultas, quienes tenían que tomar su propia decisión *antes* de ver la sugerencia de la máquina aceptaron sus sugerencias erróneas con menos frecuencia que quienes veían una breve explicación del razonamiento de la máquina. Explicar la IA no resolvió el problema. Lo que ayudó fue forzar el juicio propio. Forzar el juicio significa obligar a pensar. Como pensar cansa, a los participantes estos métodos fueron los que menos les gustaron, y ayudaron sobre todo a quienes ya disfrutaban del pensamiento esforzado. Es lo que todos conocéis como el efecto Mateo: a quien más tiene, más se le da.
- **Incorporad los límites al sistema.** El chatbot del Capítulo 1.2, que no daba las respuestas directamente y en su lugar hacía preguntas, es la misma idea llevada al software: el alumnado que lo usó no perdió nada en el examen, mientras que quienes usaron la versión sin restricciones salieron mal parados.
- **Haced visible el proceso en lugar de perseguir el producto.** Borradores, notas de trabajo, una breve conversación sobre el trabajo. Esto os muestra lo que un estudiante es capaz de hacer sin que nadie tenga que demostrar qué hizo.

Dinamarca está poniendo a prueba este principio: en las escuelas participantes, el alumnado puede usar IA mientras se prepara para el examen oral de inglés, y el examen escrito incluye una parte manuscrita sin ayudas.

Dos límites merecen conocerse antes de confiar en esto.

El primero: el consejo popular de que el profesorado simplemente indique a un chatbot que actúe como tutor socrático puede no sobrevivir a una conversación larga. Las primeras pruebas, procedentes de un estudio con estudiantes simulados que aún no ha sido revisado por pares (octubre de 2026), sugieren que, bajo una presión persistente, estos tutores vuelven a dar las respuestas directamente. Los límites fiables tienen que estar incorporados al sistema, lo que los convierte en una decisión de compra más que en una habilidad de prompting.

El segundo: hacer que una IA explique su razonamiento suena como la salvaguarda obvia, pero las explicaciones no evitan de forma fiable el exceso de confianza. En un estudio, dejaron sin cambios la aceptación de consejos *erróneos* por parte de las personas, y las explicaciones demasiado técnicas o demasiado simplificadas pueden aumentar la confianza mal depositada. La explicación es solo una condición para el aprendizaje.

Actitud: dos modos de fallo, no uno

Nos preocupa que el profesorado deposite demasiada confianza en la IA. Sin embargo, deberíamos preocuparnos igual por lo contrario. Incluso las personas con experiencia pueden dejarse llevar por el exceso de confianza. En un pequeño piloto en escuelas estadounidenses, cuatro de cada diez

docentes encuestados consideraron que la retroalimentación de una herramienta de calificación con IA no era útil, era mayoritariamente vaga o incluso incorrecta. La investigación sobre el sesgo de automatización muestra que las personas tienden a aceptar la sugerencia de una máquina incluso cuando tienen motivos para dudar de ella. **Detectar un error y corregirlo son pasos distintos.**

El polo opuesto es la ocultación. El profesorado esconde su propio uso de la IA de sus colegas y de su alumnado, prohibiéndoles hacer lo que ellos mismos hacen: el 77 % usa la IA en privado, mientras que el 57 % dice que el alumnado nunca debería usarla para generar ideas. Mientras tanto, un estudio con estudiantes académicamente destacados de 1.º a 3.º de la ESO encontró que entre un tercio y dos quintas partes ya afirmaban que la IA sabe más que sus profesores. Una profesión que oculta sus prácticas no puede servir de modelo de una práctica responsable, y la transparencia no es solo una cortesía. Es el mecanismo por el que funciona la enseñanza con el ejemplo.

¿La IA nivela el terreno de juego o lo inclina?

En tareas relacionadas con el trabajo, la IA iguala las diferencias. En un experimento, los participantes con niveles educativos más altos rindieron significativamente mejor que los de niveles más bajos. Con un asistente de IA, desaparecieron aproximadamente tres cuartas partes de esa brecha. Al retirar de nuevo el asistente, una parte considerable de la brecha volvió a aparecer. **Lo que la IA cerró fue la brecha en la producción, no la brecha en la capacidad.**

En tareas de aprendizaje, ocurre lo contrario. Un estudio no encontró ningún beneficio medio en absoluto, sino una distancia creciente entre el alumnado que ya sabía mucho y el que no. Otro observó a estudiantes que escribían en una lengua extranjera: quienes escribían peor recibieron más correcciones de la IA y consiguieron aplicar con éxito menos de ellas, y varios encontraron desalentadora la avalancha de retroalimentación. La retroalimentación sin la capacidad de priorizarla no ayuda; sepulta.

La brecha ya existe antes de que nadie toque una herramienta. En Alemania, el 80 % del alumnado de los hogares más favorecidos ve la IA como una oportunidad, frente al 55 % de los menos favorecidos. En Gran Bretaña, la división pasa por la sala de profesores más que por el armario de dispositivos: el 45 % del profesorado de escuelas privadas ha recibido formación formal en el uso de la IA, frente al 21 % en las públicas. Dentro del sector público, las escuelas más ricas vuelven a superar a las más pobres. Una encuesta de 2026 a más de 25 000 menores de 17 países europeos encuentra la misma división social en cuánto y de qué forma tan variada usan la IA. Una frase de un comentarista resume todo el problema mejor que cualquier cifra: *«los ricos tienen acceso tanto a la tecnología como a personas que les ayudan a usarla, mientras que los pobres solo tienen acceso a la tecnología»*.

Así que el cuello de botella es qué alumno tiene un docente que ha recibido formación. La IA compensa cuando tres cosas se sostienen juntas: acceso seguro, una herramienta diseñada para

aprender en lugar de para responder, y un docente que acompaña su uso. Quitad cualquiera de las tres, y en su lugar amplifica. Como las tres tienden a estar presentes en las mismas escuelas y ausentes en las mismas escuelas, la amplificación es lo predeterminado, y la equiparación es algo que hay que construir.

Revision #14
Created 2026-09-01 09:14:59 UTC by Alexander
Updated 2026-10-20 16:16:15 UTC by Alexander