CUADERNOS DE DOCENCIA UNIVERSITARIA 51

¿Cómo hacer un uso responsable de la inteligencia artificial en educación?

Aspectos técnicos y éticos de la utilización de la inteligencia artificial en el aula

Jordi Vitrià Marca

Octaedro Editorial

IDP Institut de Desenvolupament Professional
UNIVERSITAT DE BARCELONA

Cuadernos de docencia universitaria 51

Título: *¿Cómo hacer un uso responsable de la inteligencia artificial en educación? Aspectos técnicos y éticos de la utilización de la inteligencia artificial en el aula*

Primera edición: julio de 2025

Recepción del original: 4/10/2024
Aceptación: 10/1/2025

Ediciones OCTAEDRO
Bailèn, 5, pral. - 08010 Barcelona
Tel.: 93 246 40 02
www.octaedro.com - octaedro@octaedro.com

Universitat de Barcelona
Institut de Desenvolupament Professional
Passeig de la Vall d'Hebron, 171, 08035 Barcelona
Tel.: 934035175

ISBN: 978-84-1079-134-3

Diseño y producción: Servicios Gráficos Octaedro

SOBRE EL AUTOR

Jordi Vitrià

Catedrático de la Universidad de Barcelona, donde imparte un curso de iniciación a los usos de los algoritmos basados en datos y cursos avanzados de IA y ética en la ciencia de datos. Actualmente es miembro del Departamento de Matemáticas e Informática de la UB. También es director del Máster de Principios Fundamentales de la Ciencia de Datos, codirector del Posgrado de Ciencia de Datos e Inteligencia Artificial de la UB y director de la Cátedra de IA y Medios.

ÍNDICE

INTRODUCCIÓN

En agosto de 2020, cientos de estudiantes se congregaron frente al Ministerio de Educación británico en Londres, coreando: «Fuck the algorithm!» ('¡Que le den al algoritmo!') (Hao, 2020). A los pocos días, sus protestas llevaron a las autoridades a dar marcha atrás y descartar las «notas» que un algoritmo había generado para los exámenes que los estudiantes no pudieron hacer debido a la pandemia de la COVID-19.

Los exámenes en cuestión eran los *A-level*, un conjunto de pruebas que los alumnos realizan alrededor de los 18 años. Estos exámenes son los últimos antes de ingresar en la universidad y tienen un gran peso en el proceso de admisión en la educación superior. Las universidades hacen ofertas basadas en estas calificaciones y, generalmente, un estudiante debe alcanzar determinados resultados para asegurarse una plaza.

El Gobierno británico decidió que, dada la situación creada por la pandemia de la COVID-19, no era aconsejable hacer estos exámenes de forma presencial. En su lugar, propuso que las notas se determinaran mediante un algoritmo[1] a partir de los siguientes datos:

1. La *distribución de calificaciones de los centros* a los que pertenecían los alumnos durante los tres años anteriores.
2. Un *ranking de cada alumno* dentro de su propio centro y para cada asignatura concreta, basado en la opinión del profesor de la asignatura.
3. Las *calificaciones previas de cada alumno* en cada asignatura de su centro.

El resultado de este algoritmo fue que casi el 40 % de los estudiantes recibieron notas más bajas de lo habitual, mientras que solo un 2 % obtuvieron notas más altas, lo cual provocó protestas públicas y acciones legales. Ante esta situación, el Gobierno británico se retractó y decidió que los estudiantes recibirían calificaciones basadas en la estimación

1. En nuestro caso, el algoritmo en cuestión era un algoritmo que estaba diseñado con técnicas de inteligencia artificial.

de sus profesores sobre cuál habría sido su calificación si los exámenes se hubieran desarrollado según lo previsto.

El algoritmo en cuestión había analizado la distribución histórica de las notas de un centro escolar y asignaba la nota de un alumno en función de su *ranking*. Por ejemplo, si un alumno se encontraba en la mitad de la lista de su centro, su nota era, aproximadamente, igual a la obtenida en años anteriores por otra persona en la misma posición del *ranking*. El objetivo de los desarrolladores del algoritmo al incorporar la distribución de las calificaciones fue corregir una desviación histórica observada en algunos centros educativos, que tendían a reportar calificaciones significativamente más altas en comparación con los resultados obtenidos en los *A-level*.

¿Qué es un algoritmo?

Un algoritmo es un conjunto claro y detallado de pasos que se siguen para resolver un problema o responder a una pregunta. Funciona como una receta que indica qué hacer en cada momento, sin necesidad de creatividad o habilidades especiales para seguirlo.

Un ejemplo sencillo de un algoritmo es el que usamos para sumar dos números altos a mano:

1. Escribe los dos números uno debajo del otro, alineando sus dígitos por columnas (unidades, decenas, centenas, etc.).
2. Comienza sumando los dígitos de la columna más a la derecha (unidades).
3. Si la suma es superior a 9, escribe el dígito de las unidades del resultado y lleva el 1 (el acarreo) a la siguiente columna de la izquierda.
4. Repite el proceso con cada columna hacia la izquierda, sumando los dígitos y cualquier acarreo.
5. Una vez que hayas sumado todas las columnas, el número final es el resultado de la suma.

Este algoritmo garantiza que, si sigues los pasos correctamente, siempre obtendrás el resultado correcto, sin necesidad de hacer ningún razonamiento adicional.

Los algoritmos son nuestra principal forma de comunicarnos con los ordenadores, porque estos solo pueden procesar instrucciones precisas, lógicas y estructuradas que puedan traducirse en operaciones concretas sobre su electrónica.

La inteligencia artificial (IA) amplía el concepto tradicional de los *algoritmos* al introducir la capacidad de generarlos automáticamente a partir de datos. Esto significa que, en lugar de que un humano diseñe explícitamente cada paso de un algoritmo para resolver un problema, la IA puede analizar grandes volúmenes de datos, identificar patrones y aprender las reglas necesarias para alcanzar un objetivo específico.

El análisis posterior de un grupo de expertos detectó dos problemas en las notas generadas por el algoritmo: la *escasa precisión* del algoritmo y la *presencia de sesgos injustos*.

El primer problema era un problema de «calidad» del algoritmo. Cuando se testeó con alumnos de años anteriores, el algoritmo predijo en muchos casos notas bastante diferentes a las que esos estudiantes acabaron obteniendo. Ante tal resultado, parecía muy arriesgado utilizar los resultados del algoritmo de manera determinante sin considerar su alto grado de error.

El segundo problema consistía en un error sistemático en las predicciones del algoritmo que beneficiaba a los alumnos que asistían a centros con menos de 15 alumnos por aula. El resultado del análisis mostró que la proporción de notas altas y muy altas que obtuvieron los alumnos de nivel socioeconómico elevado gracias al algoritmo se incrementaron respecto al año anterior más del doble que en el caso de los alumnos que asistían a escuelas de entornos de nivel socioeconómico bajo. Dicho de otra manera, el algoritmo no solo reproducía los sesgos estructurales de la sociedad británica, sino que los amplificaba.

La lección de esta historia presenta varios puntos de interés. En primer lugar, el uso de algoritmos en ámbitos importantes de la vida de las personas, como la educación, el mercado laboral o la justicia, debe cumplir con un *nivel de exigencia muy elevado* en aspectos como la precisión y la ausencia de sesgos no deseados. Por ejemplo, en el campo de los recursos humanos, algunas empresas han implementado sistemas de inteligencia artificial (IA) para filtrar currículos o evaluar candidatos durante los procesos de selección. En casos notables, como el de un sistema utilizado por Amazon, se detectó que los algoritmos reforzaban estereotipos de género, desfavoreciendo a las mujeres que se presentaban para ocupar a puestos tecnológicos (Dastin, 2018). De modo similar, en el ámbito de la seguridad pública, herramientas como el *software* de predicción del crimen han sido objeto de críticas por amplificar desigualdades sociales y raciales. En ciudades de EE. UU. Unidos como Chicago o Los Ángeles, algoritmos diseñados para identificar zonas de alto riesgo han dado como resultado un mayor control policial en comunidades racializadas, perpetuando, así, desigualdades históricas (Richardson, 2019).

En segundo lugar, la detección de problemas como los comentados solo es posible si se exige un cierto nivel de *transparencia* que permita a expertos y ciudadanos analizar y cuestionar los resultados generados por estos sistemas.

Finalmente, la solución de problemas relacionados con algoritmos no depende únicamente de expertos, sino también de una ciudadanía *crítica e informada* que exija responsabilidad en su diseño y uso. Tal como sucedió con los estudiantes en el Reino Unido, otras movilizaciones, como las protestas contra el uso de reconocimiento facial por parte de la policía en San Francisco en 2019, han demostrado que la presión pública puede forzar a las instituciones a replantearse el uso de estas tecnologías.

Experiencias como las descritas demuestran que, aunque los algoritmos tienen un gran potencial, su implementación sin supervisión rigurosa puede acentuar las desigualdades y perpetuar problemas estructurales. Por ello, su adopción ha de ir acompañada de un marco ético, una supervisión transparente y una participación activa de la sociedad para garantizar su impacto positivo.

Ante la introducción de la IA en muchos aspectos fundamentales de nuestras vidas, y para evitar problemas como el expuesto, el Parlamento Europeo aprobó recientemente la Ley de la Inteligencia Artificial (EUR Lex, 2024), que pretende regular el uso de esta tecnología en diferentes sectores, incluido el educativo. En este sector, la Ley busca generar una mayor confianza y seguridad en la adopción de tecnologías de IA, ya que las herramientas empleadas deberán ajustarse a estrictos requisitos de transparencia, protección de datos y derechos fundamentales que, en principio, mitigarán sus efectos adversos, entre los cuales están los que hemos visto en el ejemplo anterior.

Este libro no tiene por objetivo explicar el impacto de la nueva ley en el sector, ni tan siquiera dar una serie de recomendaciones genéricas sobre los usos más pertinentes de la IA en educación. La IA puede ser una herramienta poderosa para gestionar, apoyar y mejorar el aprendizaje de los estudiantes, las tareas de los docentes y la gestión del sistema educativo, pero, como hemos visto, puede ocasionar daño o perpetuar y agravar los problemas y las desigualdades existentes en la educación si

no se aplican adecuadamente. El objetivo de este libro es metodológico: sentar unas bases simples y comprensibles que permitan a todos los interesados en temas educativos evaluar de forma crítica posibles situaciones concretas con las que se van a encontrar en su práctica docente. Para sentar estas bases, es necesario conocer algunos conceptos básicos de estas tecnologías, pero también la naturaleza de ciertos desafíos éticos y las consecuencias sociales que deben abordarse y que van mucho más allá de la privacidad y la seguridad de los datos.

El capítulo 1 define la IA y sus usos potenciales en la educación, centrándose en su capacidad para mejorar la toma de decisiones basada en datos, predecir y modelar el rendimiento de los estudiantes, además de asistir en la creación de materiales de aprendizaje personalizados. También examina los tipos de IA, incluyendo la IA simbólica, la IA predictiva, el aprendizaje profundo y la IA generativa, destacando sus características y aplicaciones. Asimismo, aborda los retos y los problemas que pueden surgir de la implementación de la IA en la educación, como las desigualdades en el acceso, la discriminación y las fisuras en cuanto a privacidad.

El capítulo 2 explica el ciclo de vida de un sistema de IA predictiva, desde la recolección de datos hasta el desarrollo de modelos y su implementación en aplicaciones concretas. Ilustra cómo la IA predictiva puede usarse para predecir el éxito de los estudiantes, con un ejemplo hipotético de una universidad que utiliza datos históricos para prever el rendimiento futuro de los alumnos. Además, muestra cómo se crean y entrenan los modelos de IA predictiva utilizando algoritmos como la regresión, los árboles de decisión y las redes neuronales.

El capítulo 3 se adentra en el funcionamiento de la IA generativa, en lo relativo a cómo se entrena con modelos de lenguaje y cómo genera nuevos contenidos a partir de datos existentes. Discute la paradoja de cómo los modelos generativos pueden ser competentes en tareas no previstas utilizando la analogía de la biblioteca de Babel imaginada por J. L. Borges para ilustrar su capacidad de producir «texto plausible», pero no necesariamente «texto correcto». Asimismo, explora las aplicaciones de la IA generativa en la educación, entre las cuales la creación de contenidos, el desarrollo de entornos de aprendizaje personalizados, la asistencia en la escritura y creación de contenidos por parte de los

estudiantes, la evaluación y la retroalimentación. También toca los aspectos sociotécnicos de la IA generativa, como la concentración de poder y conocimiento en manos de grandes empresas tecnológicas, junto con las implicaciones éticas y sociales de su uso.

El capítulo 4 propone una metodología para evaluar el uso responsable de la IA en la educación, compuesta por cuatro pasos: definir el problema educativo y el valor que aporta la IA, comprender el marco legislativo y social, identificar los riesgos asociados y evaluar la viabilidad de la propuesta. Describe el modelo Canvas EduIA como una herramienta para evaluar la viabilidad de las aplicaciones de IA en la educación, destacando la importancia de considerar factores como la equidad, la autonomía de alumnos y profesores, la privacidad y la transparencia. También resalta la importancia de la toma de decisiones grupal en la evaluación de la IA, a fin de mitigar los sesgos individuales y promover una perspectiva más completa y equilibrada.

Finalmente, las conclusiones resumen las herramientas y los usos más destacados de la IA en la educación, como la tutoría inteligente, el análisis de aprendizaje, la evaluación automatizada, el aprendizaje adaptativo y los asistentes virtuales. Definen lo que significa hacer un uso responsable de la IA en la educación, poniendo énfasis en las consideraciones éticas, la confianza, la reputación y el cumplimiento legal y regulatorio. También se proporcionan directrices para asegurar el uso responsable de la IA en la educación, subrayando la necesidad de un proceso deliberativo, la participación de los actores clave y la adaptación a las particularidades de cada contexto educativo.

1. INTELIGENCIA ARTIFICIAL Y EDUCACIÓN

La progresiva digitalización de nuestro entorno personal, familiar, de estudio y de trabajo ha facilitado el desarrollo de un conjunto de tecnologías, basadas en la ciencia de datos y la inteligencia artificial (IA), que permiten automatizar algunas tareas que hasta ahora estaban relacionadas exclusivamente con la inteligencia humana y que abren la puerta a un nuevo modelo para la toma de decisiones y la resolución de problemas. El mundo de la educación no ha sido una excepción y, durante estos últimos años, hemos visto múltiples propuestas para transformar la forma en que se enseña y se aprende, que van desde la personalización del aprendizaje y el desarrollo de tutores virtuales o asistentes de enseñanza hasta la gestión y administración educativa.

El uso de las tecnologías de IA en la actualidad se divide en dos áreas principales (Kennedy *et al*., 2023). La primera es la generación de conocimiento a partir de datos, que permite extraer información valiosa e identificar patrones significativos en grandes volúmenes de datos, lo que facilita la toma de decisiones fundamentadas. Esto tiene aplicaciones en análisis predictivo, identificación de tendencias, clasificación y segmentación de datos, así como en la automatización de descubrimientos científicos. Este enfoque es especialmente relevante en sectores como la medicina, con diagnósticos asistidos por IA; las finanzas, con la detección de fraudes; y la logística, con la optimización de rutas, entre otros. La segunda área es la facilitación de la comunicación con los ordenadores mediante interfaces naturales, que posibilita que las personas interactúen de forma más intuitiva y fluida con la tecnología. Esto se logra a través del reconocimiento de voz y lenguaje natural, recurriendo a asistentes virtuales del tipo Alexa, Siri o Google Assistant; el reconocimiento de imágenes y gestos, utilizado en sistemas de seguridad, videojuegos e interfaces accesibles; el procesamiento y generación de texto, como en *chatbots*, traductores automáticos y herramientas de redacción; y la interacción multimodal, que combina varias formas de comunicación, como voz y gestos, para controlar dispositivos. Estas dos áreas se complementan, ya que mejoran nuestra capacidad de análisis y manejo de datos al tiempo que eliminan barreras entre humanos y máquinas, lo cual redunda en una tecnología más inclusiva y eficiente.

El valor fundamental de la IA en el ámbito educativo radica en la diversidad de sus aplicaciones, que se pueden clasificar en tres sectores principales. En primer lugar, la toma de decisiones basada en datos permite analizar grandes volúmenes de información para identificar patrones y tendencias útiles que mejoran la toma de decisiones en áreas como las políticas educativas, la administración escolar, la selección de materiales y métodos de enseñanza, la evaluación del aprendizaje y las adaptaciones curriculares para satisfacer diferentes necesidades educativas. Aunque siempre debería haber un ser humano detrás de cada decisión, el uso adecuado de los datos puede ayudar a minimizar los errores que suelen ocurrir cuando las decisiones se toman confiando demasiado en la intuición o basándose únicamente en prácticas previas que han dado buenos resultados.

En segundo lugar, la predicción y el modelado, a través de modelos estadísticos y algoritmos de aprendizaje automático, facilitan el análisis del historial académico de los estudiantes, su participación en el aula y otros factores para prever su rendimiento futuro. Esto permite a los educadores intervenir de manera temprana con los estudiantes que presentan riesgo, proporcionando el apoyo adicional necesario. Además, pueden ajustar el contenido, las evaluaciones y los desafíos de acuerdo con las necesidades de aprendizaje individuales de cada alumno, optimizando, así, su trayectoria educativa.

La IA, por último, asiste a la educación mediante la generación de materiales de aprendizaje personalizados que se adaptan a las necesidades, niveles de habilidad e intereses de cada persona. Por ejemplo, la IA puede crear textos, problemas y ejemplos en tiempo real para estudiantes que avanzan a diferentes ritmos. También permite a los educadores diseñar preguntas de examen y ejercicios de práctica con una amplia variedad y niveles de dificultad adaptados a las habilidades de los estudiantes, mejorando su preparación para los exámenes. Asimismo, la IA puede ayudar a superar las barreras de comunicación y lenguaje en entornos educativos multilingües gracias a la generación de traducciones en tiempo real y materiales de aprendizaje en varios idiomas.

La implementación de estas tecnologías en la educación, pese a que tiene un inmenso potencial, también puede presentar desafíos y problemas: desigualdades en el acceso, riesgos de discriminación, problemas de pri-

vacidad, etc. Abordar estos problemas exige una planificación cuidadosa, inversión en infraestructura y formación, una comprensión clara de las limitaciones de la tecnología, y políticas sólidas para minimizar los peligros asociados a esta tecnología. La colaboración entre desarrolladores de tecnología, educadores, psicólogos, filósofos, alumnado, familias y legisladores es crucial si se quieren maximizar los beneficios de la IA en la educación, a la vez que se minimizan sus riesgos y se superan los desafíos.

¿Qué es la inteligencia artificial?

No hay una definición aceptada universalmente de la palabra *inteligencia*, pero podemos elaborar una que nos sirva de referencia:

> La inteligencia es una facultad natural que permite a ciertos organismos vivos alcanzar objetivos complicados en situaciones complejas de manera racional, especialmente cuando se enfrentan a circunstancias nuevas y deben adaptarse.

La *inteligencia* no es una característica exclusiva de los humanos, hasta el punto de que podríamos decir que cualquier organismo vivo mínimamente complejo posee un cierto nivel de inteligencia. Teniendo esto en cuenta, podemos afirmar que la facultad de la inteligencia se manifiesta de muchas maneras posibles, siendo la inteligencia humana una de las manifestaciones más complejas y amplias de esta facultad.

Como hemos dicho, dada su complejidad y variedad, el concepto de *inteligencia* no goza de una definición precisa y sus límites no están claramente establecidos. Sin embargo, existe un cierto consenso en torno a la idea de que, en el caso de los humanos, se puede descomponer en una serie de competencias.

Una *competencia* es una habilidad específica, o un conjunto de habilidades, que permite a un individuo o sistema resolver determinado tipo de problemas o adaptarse a nuevas situaciones mediante el procesamiento de información. En el contexto de la inteligencia humana, las competencias pueden ser múltiples y variadas, y abarcar áreas cognitivas, emocionales, sociales y prácticas.

Las *competencias cognitivas* son fundamentales para el funcionamiento humano, ya que permiten procesar información, comprenderla, almacenarla en la memoria y aplicarla en la resolución de problemas y la toma de decisiones. Entre las más importantes se encuentran la percepción, la capacidad de aprendizaje, el razonamiento, la planificación y la toma de decisiones, cada una de ellas con aplicaciones prácticas en la vida diaria.

La *percepción* es la habilidad para captar e interpretar información a través de los sentidos y hace posible una interacción efectiva con el entorno. Por ejemplo, cuando conducimos un coche, la percepción nos ayuda a identificar señales de tráfico, detectar otros vehículos y evaluar la velocidad y la distancia para evitar accidentes.

La *capacidad de aprendizaje* se refiere a la facultad de adquirir y retener conocimientos o habilidades mediante la experiencia o la enseñanza. Un ejemplo cotidiano sería aprender un nuevo idioma: las personas procesan información nueva, como vocabulario y gramática, y la aplican en situaciones comunicativas.

El *razonamiento* implica analizar información y llegar a conclusiones lógicas. Esto puede observarse cuando un individuo resuelve un problema matemático o decide cómo sea de reparar un electrodoméstico averiado, analizando las posibles causas y escogiendo la mejor solución.

La *planificación* permite organizar actividades o recursos para alcanzar un objetivo específico. Por ejemplo, planificar unas vacaciones implica establecer un presupuesto, elegir destinos, reservar alojamiento y organizar itinerarios para aprovechar al máximo el tiempo disponible.

Finalmente, la *toma de decisiones* se refiere a escoger la mejor opción entre diversas alternativas, considerando la información disponible y la incertidumbre asociada al efecto de la decisión. Un ejemplo sería decidir qué carrera estudiar, evaluando factores como los intereses personales, las perspectivas laborales y los costes educativos, a fin de adoptar una decisión informada.

Estas competencias trabajan de manera interconectada, facilitando la adaptación y el éxito en un entorno complejo y en constante cambio.

El término *artificial intelligence* se utilizó por primera vez en una escuela de verano, la *Dartmouth Summer Research Project on Artificial Intelligence*, realizada en 1956 en el Dartmouth College, una universidad de élite de los EE. UU. La conferencia reunió a veinte de las mentes más brillantes en informática y ciencias cognitivas de la época para discutir la siguiente conjetura (McCarthy *et al*., 1955):

> [...] que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tanta precisión que se puede hacer una máquina para simularlo. Se intentará encontrar cómo hacer que las máquinas usen el lenguaje, formen abstracciones y conceptos, resuelvan distintos tipos de problemas que ahora están reservados para los humanos y se mejoren a sí mismas. Creemos que se puede lograr un avance significativo en una o más de estas cuestiones si un grupo cuidadosamente seleccionado de científicos trabaja en él durante un verano.

El objetivo planteado para un verano está todavía pendiente de solución y su consecución es el objetivo de una numerosa comunidad científica que actualmente incluye miles de científicos de todo el mundo.

Durante las siguientes décadas, la IA se desarrolló a trompicones, con períodos de rápido progreso intercalados con los llamados «inviernos» de la IA, temporadas durante las cuales la financiación de la investigación en IA fue escasa (Crevier, 1993). Hoy en día, y muy especialmente desde 2023, nos hallamos en un «verano» de la IA, que algunos creen ya definitivo para el paso de la IA del mundo de la investigación al mundo de las aplicaciones. Esto se ha producido como resultado de la confluencia de tres elementos críticos: el crecimiento exponencial de los datos existentes gracias a Internet y la Web, la generación de nuevos algoritmos de aprendizaje automático y el desarrollo de procesadores con una elevada potencia de cálculo en paralelo.

Desde sus inicios, definir el concepto de IA ha resultado un tema arduo, como es natural cuando hablamos de un fenómeno que no acabamos de entender. En este libro asumimos una posición pragmática y la definiremos tal y como lo hace la Ley Europea de Inteligencia Artificial (EUR Lex, 2024):

> [...] un sistema basado en máquinas que está diseñado para funcionar con diversos niveles de autonomía y que puede mostrar capacidad de adaptación tras su despliegue, y que, para objetivos explícitos o implícitos, infiere, a partir de la entrada que recibe, cómo generar salidas tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales.

La definición que nos da la Ley Europea de Inteligencia Artificial resalta algunos puntos clave que podemos desglosar en un lenguaje más accesible.

Primero, la IA *puede crear sistemas que operan de manera (parcialmente) autónoma*, lo que significa que pueden operar por sí mismos sin necesidad de que alguien le indique constantemente qué hacer. Por ejemplo, un asistente virtual como Alexa o Siri puede responder a tus preguntas y tomar decisiones sin que tengas que explicarle cómo hacerlo cada vez.

Además, tiene *capacidad de adaptación*, esto es, puede aprender de su experiencia o de la información que recibe durante su funcionamiento. Esto le permite mejorar su desempeño con el tiempo. Un ejemplo cotidiano sería un sistema de recomendación, como el de Netflix, que aprende qué series o películas te gustan según lo que has visto y te sugiere nuevas opciones.

La IA también *puede tomar decisiones* o *hacer predicciones* basándose en la información que se le da. Por ejemplo, un sistema de navegación GPS puede utilizar datos en tiempo real sobre el tráfico y las preferencias del usuario para sugerirle la mejor ruta.

Finalmente, hace falta considerar que estas decisiones o recomendaciones pueden tener un impacto tanto en el mundo físico como en el virtual, afectando a otras personas o a la sociedad en su conjunto.

En la actualidad existe un consenso entre la mayoría de los investigadores del área sobre el hecho de que estamos todavía lejos no solo de construir artefactos con inteligencia general, sino, incluso, de comprender todos los aspectos de la inteligencia. Pero esto no significa que no haya ciertas áreas de investigación en la IA que hayan logrado algunos resultados parciales bastante interesantes. Las áreas más importantes son:

- El *aprendizaje automático* a partir de datos, conocido en inglés como *machine learning*, es una rama de la IA que se centra en la construcción de algoritmos que permiten a las máquinas aprender y hacer predicciones o tomar decisiones basadas en datos. Los sistemas de aprendizaje automático construyen modelos matemáticos que se ajustan o entrenan a partir de los datos de entrada para que puedan,

sin estar explícitamente programados para cada caso, tomar decisiones razonables cuando se exponen a nuevos datos.
Ejemplo: Imagina que tienes una máquina que intenta adivinar si una fruta es una manzana o una naranja basándose en algunos datos. Para ello, usas el aprendizaje automático (o *machine learning*). Primero, proporcionas a la máquina datos de ejemplo sobre diferentes frutas, como:
- *Color:* ¿Es roja o naranja?
- *Tamaño:* ¿Es pequeña como una manzana o más grande como una naranja?
- *Textura:* ¿Es lisa o rugosa?

A partir de estos datos, la máquina aprende a identificar patrones. Por ejemplo:
- Las manzanas suelen ser rojas, pequeñas y lisas.
- Las naranjas suelen ser naranjas, grandes y rugosas.

Ahora, cuando le muestras una fruta que no ha visto antes, la máquina analiza sus características (color, tamaño y textura) y usa lo que aprendió de los ejemplos previos para decirte si se trata de una manzana o de una naranja.

Este es un ejemplo sencillo de cómo funciona el aprendizaje automático: la máquina aprende a tomar decisiones basándose en los datos que le proporcionas, sin necesidad de que tú le digas explícitamente todas las reglas.

- La *percepción visual del entorno* es una fuente de información que nos permite explorar nuestro entorno y hacernos una idea del estado del mundo. Este campo combina métodos del aprendizaje automático, el procesamiento de imágenes y la visión artificial para permitir a las máquinas «ver» y procesar imágenes y vídeos de manera similar a como lo hacen los humanos, incluso a escalas y velocidades que, en ocasiones, superan las capacidades humanas.
  Ejemplo: Un ejemplo adicional de percepción visual del entorno es el uso de aplicaciones para el diagnóstico médico basado en imágenes. Estas aplicaciones emplean algoritmos de visión artificial y aprendizaje automático para analizar imágenes médicas como radiografías, tomografías o resonancias magnéticas, así como para detectar posibles problemas de salud. Por ejemplo, un sistema de IA puede exami-

nar imágenes de la retina, el sistema puede identificar condiciones como retinopatía diabética o degeneración macular. Estos sistemas no solo ayudan a los especialistas a tomar decisiones más rápidas y precisas, sino que también pueden procesar grandes volúmenes de datos en poco tiempo, facilitando el cribaje de enfermedades en poblaciones amplias y mejorando la accesibilidad al diagnóstico en áreas con pocos recursos médicos.

- El *reconocimiento y producción del habla* a partir de los sonidos es un elemento clave como forma de comunicación con los humanos y abre la puerta a un estilo de colaboración mucho más natural entre humanos y computadoras.
  Ejemplo: Un ejemplo simple es el uso de asistentes de voz como Alexa, Siri o Google Assistant. Estos sistemas pueden reconocer comandos de voz del estilo: «¿Qué tiempo hará mañana?» o: «Pon una alarma a las 7 de la mañana», procesar el lenguaje hablado y responder con información o acciones específicas, como proporcionar el pronóstico del tiempo o configurar la alarma. Esto permite una interacción más natural, ya que puedes «hablar» con la tecnología en lugar de escribir o pulsar botones.
- El área de la *percepción, el análisis, la generación y la traducción del lenguaje natural* se ocupa de la interacción entre las computadoras y los humanos mediante el lenguaje humano. El objetivo es hacer posible que los ordenadores comprendan, interpreten y produzcan lenguaje humano de una manera útil.
  Ejemplo: Un ejemplo simple es el uso de *chatbots* en servicios de atención al cliente. Así, cuando escribes: «¿Cuál es el estado de mi pedido?» en el chat de una tienda en línea, el *chatbot* entiende tu pregunta en lenguaje natural, busca la información en su base de datos y responde con algo del estilo: «Tu pedido llegará mañana por la tarde». Esto facilita una interacción rápida y eficiente, sin necesidad de intervención humana directa.
- El *razonamiento y la planificación* son áreas de investigación clave en el campo de la IA, enfocadas en dotar a las máquinas de la capacidad de tomar decisiones lógicas y elaborar planes de acción para alcanzar objetivos específicos. Estas áreas abordan algunos de los desafíos más complejos y fascinantes en la creación de sistemas inteligentes que pueden actuar de manera autónoma.

Ejemplo: Un ejemplo simple es el funcionamiento de un robot aspirador. El robot analiza el diseño de una habitación, identifica obstáculos como muebles y decide el mejor camino para limpiar todo el espacio de manera eficiente. Además, si su batería está baja, planifica regresar a su base de carga antes de continuar la tarea. Este proceso combina razonamiento y planificación para alcanzar su objetivo: limpiar la casa de forma autónoma.

Estos resultados se han ido produciendo a lo largo de los últimos setenta años de manera progresiva, con muchos altibajos que han hecho que la IA se vaya incorporando de forma gradual a nuestras vidas.

¿Cómo se construye un sistema de inteligencia artificial?

Las primeras aplicaciones de IA, desarrolladas en los años setenta, se basaron en el *enfoque simbólico,* también conocido como «basado en reglas» (Boden, 2014). Este enfoque consistía en definir un conjunto de reglas claras y precisas que indican a la máquina, de forma explícita, cómo actuar en diferentes situaciones. Para usar estas reglas, se empleaba un motor de inferencia, que es como el «cerebro» del sistema: un programa informático que analizaba los datos de entrada, buscaba las reglas relevantes y las aplicaba para producir un resultado o tomar una decisión. Es decir, el motor de inferencia conectaba las reglas con los datos para resolver problemas o responder preguntas.

Las reglas se implementan a través de la programación y surgen de la experiencia humana en la materia. Una vez que se han definido las reglas, la IA se encarga de aplicarlas a los datos de entrada. Por ejemplo, un sistema basado en reglas puede ser utilizado para clasificar correos electrónicos como *spam* o no *spam* basándose en la presencia de ciertas palabras clave o patrones en el correo.

¿Cómo funciona un sistema basado en reglas para detectar *spam*?

Un sistema basado en reglas para detectar *spam* funciona aplicando un conjunto predefinido de reglas o criterios para identificar correos electrónicos no deseados. El proceso podría ser el que sigue:

1. Los expertos en seguridad y correos electrónicos definen una serie de reglas basadas en características comunes del *spam*. Estas reglas pueden ser simples o complejas, dependiendo del sistema. Por ejemplo:
   a) Correos con ciertas palabras clave como «gratis», «ganar» u «oferta limitada».
   b) Presencia de enlaces a dominios sospechosos.
   c) Uso excesivo de mayúsculas y signos de exclamación.
2. Cada correo electrónico entrante se analiza para ver si cumple alguna de las reglas definidas. Esto incluye revisar el cuerpo del mensaje, los encabezados, los enlaces y los archivos adjuntos.
3. A cada regla se le asigna un peso o puntuación. Si un correo cumple una regla, se le suman los puntos correspondientes a su puntuación total.
4. Se establece un umbral de puntuación. Si la puntuación total del correo supera dicho umbral, el correo se marca como *spam*.

Durante décadas se desarrollaron este tipo de «sistemas expertos» de IA, que podían estar basados en centenares de reglas, para una vasta gama de aplicaciones, como diagnósticos médicos, aplicaciones financieras y procesos de fabricación (Jackson, 1990). Independientemente del número de reglas, es fácil para un humano entender la lógica de las decisiones de estos sistemas.

Los sistemas basados en reglas son todavía útiles y utilizados en ciertos ámbitos, pero, a medida que el número de reglas crece y que sus interacciones se multiplican, su desarrollo y su mantenimiento se complican extraordinariamente. Otra de sus limitaciones es que los expertos tienen que expresar de forma explícita las reglas, un aspecto que no es factible en muchos ámbitos de la inteligencia como el diagnóstico a partir de imágenes médicas.

Muchos de los avances actuales de la IA se sustentan en un modelo alternativo a las reglas: las técnicas de aprendizaje automático o *IA predictiva* (Alpaydin, 2021). Estas técnicas se basan en el análisis de datos para extraer de ellos de forma automática las reglas de decisión. El papel del experto en este caso no es enunciar las reglas, sino que se reduce a etiquetar los datos de manera que el sistema de IA sea capaz de generar las reglas a partir de esta información implícita.

¿Cómo funciona un sistema basado en IA predictiva para detectar *spam*?

Un sistema basado en IA predictiva para detectar *spam* utiliza técnicas de análisis de datos y modelos predictivos para identificar correos electrónicos no deseados. El proceso podría ser este:

- Se recopila un gran número de correos electrónicos etiquetados como *spam* y no *spam*. Estos datos servirán para entrenar el modelo.
- Se extraen características relevantes de los correos electrónicos como palabras clave, frecuencia de palabras, longitud del correo, presencia de enlaces, direcciones de remitente, etc.
- Se seleccionan uno o varios algoritmos de aprendizaje automático. Algunos algoritmos comunes para la detección de *spam* incluyen *Naive Bayes*, regresión logística, árboles de decisión, máquinas de soporte vectorial (SVM) o redes neuronales.
- Se divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de pruebas.
- El modelo se «entrena» utilizando el conjunto de entrenamiento, ajustando los parámetros del algoritmo seleccionado para minimizar el error en la clasificación de correos electrónicos. El resultado es un modelo matemático que permite analizar un correo y determinar su clase a partir de las características relevantes.
- Se evalúa el modelo entrenado utilizando el conjunto de pruebas para medir su precisión.

En el caso de un sistema de aprendizaje automático aplicado a la educación, los datos podrían estar formados por la información sobre el rendimiento académico de los estudiantes, como calificaciones en exámenes, trabajos entregados, asistencia a clases y participación en actividades extracurriculares. Las etiquetas, en este contexto, podrían ser los resultados de los estudiantes, como sería si aprueban o no una asignatura, o la nota media de sus estudios.

La tarea del sistema de IA sería, entonces, procesar estos datos para inferir de forma automática la relación entre los datos de entrada (rendimiento y actividades de los estudiantes) y la etiqueta de salida (resultados académicos).

Cabe señalar que, en este caso, la relación entre los datos de entrada y los datos de salida no estará especificada en forma de reglas de fácil lectura para un humano. En su lugar, se expresa mediante complejas relaciones matemáticas implementadas en algoritmos, como las redes neuronales. Esto hace que sea imposible, para un ser humano, entender la lógica de las decisiones del sistema simplemente inspeccionando el algoritmo resultante.

Los modelos basados en la IA predictiva constituyen la base de la segunda oleada de aplicaciones de IA que arrancó en los años noventa. Estos modelos se han usado en ámbitos tan dispares como las finanzas, la banca, la salud, la medicina, el comercio electrónico o el transporte y la logística.

En 2012 apareció un nuevo tipo de modelos, los modelos de *aprendizaje profundo*, que reforzaron esta tendencia y ampliaron la gama de aplicaciones (Zhang *et al*., 2023). El aprendizaje profundo es un subcampo de la IA predictiva que se vale de redes neuronales para analizar otros tipos de datos (imágenes, vídeos, textos) con una elevada precisión.

Las ventajas del aprendizaje profundo incluyen su capacidad para procesar y aprender a partir de grandes cantidades de datos no estructurados y su habilidad para realizar tareas de clasificación y predicción con mucha precisión. No obstante, también presenta desafíos, tales como la necesidad de grandes conjuntos de datos para entrenar a los modelos de manera efectiva, su alto coste computacional y la dificultad para interpretar los modelos, en lo que se conoce como el problema de la «caja negra» (Castelvecchi, 2016).

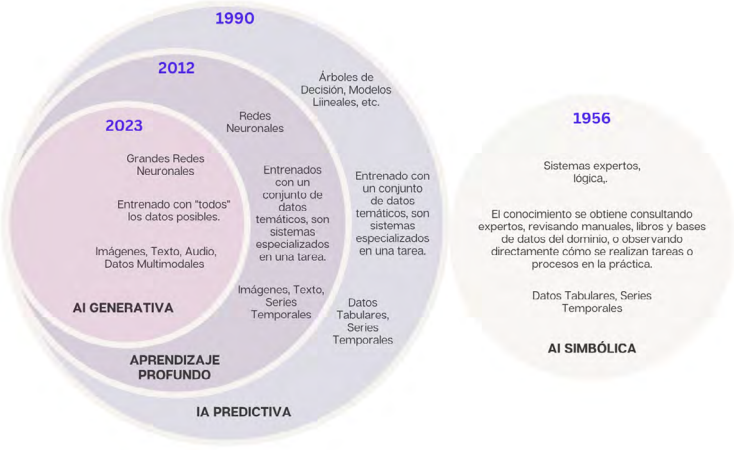

Figura 1. Las herramientas de IA que podemos usar en la actualidad se han desarrollado mayoritariamente dentro de cuatro marcos conceptuales: el simbólico, el predictivo, el aprendizaje profundo y el generativo. Esta figura muestra los cuatro marcos, sus herramientas principales, cómo crean sus modelos y, finalmente, a qué tipo de datos pueden aplicarse.

Las redes neuronales profundas mejoraron significativamente el procesamiento del lenguaje natural respecto a las tecnologías basadas en los modelos predictivos o las basadas en el modelo simbólico, permitiendo la representación del significado del texto de una forma más precisa y útil para muchas aplicaciones. Esto hizo posible la existencia de los asistentes virtuales nacidos en la primera década del siglo XXI, capaces de interpretar algunos aspectos del lenguaje humano como identificar una necesidad y realizar una acción para satisfacerla, o responder con un guion predefinido a una determinada pregunta.

Pero en 2017 se produjo un avance fundamental en el campo del aprendizaje profundo que transformó radicalmente el procesamiento del lenguaje natural: la introducción de un nuevo tipo de red neuronal profunda basada en mecanismos de atención (Vaswani *et al*., 2017). Este diseño hizo posible la construcción de los grandes modelos de lenguaje –*large language models* (LLM)– (Mann *et al*., 2020), sentando, así, los pilares de lo que hoy conocemos como *inteligencia artificial generativa*.

¿Cómo aprende un sistema de IA generativa?

Un sistema de IA generativa aprende a partir de enormes volúmenes de datos utilizando técnicas avanzadas de aprendizaje automático, particularmente, redes neuronales profundas. Aquí se muestra cómo funciona el proceso de aprendizaje con un ejemplo específico: un modelo de generación de texto como GPT (*generative pre-trained transformer*).

1. Se recopilan ingentes cantidades de datos de texto de diversas fuentes, como libros, artículos, sitios web, etc. Estos datos han de ser variados y representativos de la lengua (y de los registros de esta lengua) que se pretende modelar.
2. El texto se convierte en secuencias de elementos numéricos (que corresponden a palabras o subpalabras) que la red neuronal puede procesar.
3. Se utiliza una arquitectura de red neuronal profunda, como un *transformer*, para entrenar el modelo. Un *transformer* tiene múltiples capas de atención que le permiten entender las relaciones contextuales entre las palabras en una secuencia.
4. El modelo se entrena en dos fases: preentrenamiento y ajuste fino (*fine-tuning*).
5. El modelo se entrena en una tarea de lenguaje general, como predecir la siguiente palabra en una oración. Por ejemplo, dado el texto: «El gato está sobre el», el modelo aprende a predecir «tejado» a partir del contexto. Esto se hace utilizando grandes volúmenes de texto sin etiquetas, donde el modelo aprende las estructuras del lenguaje y las relaciones entre palabras.
6. Tras el preentrenamiento, el modelo se ajusta utilizando un conjunto de datos específico para la tarea deseada, como generación de texto coherente o respuesta a preguntas.

7. Una vez entrenado, el modelo puede generar texto nuevo. Se le proporciona una entrada inicial (*prompt*) y el modelo genera texto basado en lo que ha aprendido:

   > **Entrada** (*prompt*): «*Érase una vez, en un pequeño pueblo,*»

   > **Salida** (respuesta del modelo): «*Érase una vez en un pequeño pueblo, rodeado de verdes colinas y bosques densos. Los habitantes vivían en armonía, y cada primavera celebraban una gran fiesta en honor a la naturaleza. Un día, un extraño visitante llegó al pueblo, trayendo consigo historias de tierras lejanas y aventuras extraordinarias...*»

El lanzamiento del LLM más famoso, ChatGPT (OpenAI, 2022), abrió la puerta de esta tecnología a cualquier persona con acceso a Internet, que puede interactuar con una interfaz web o una aplicación en el móvil de forma sencilla. Esto puso de relieve los impresionantes avances de los LLM, ya que anteriormente los LLM más potentes solo estaban disponibles para investigadores con muchos recursos y aquellos con conocimientos técnicos muy profundos.

Posteriormente, estos modelos se han generalizado y son capaces de aceptar entradas multimodales (texto, imágenes, vídeo), y también de generar salidas multimodales, constituyendo el elemento principal de la llamada IA generativa.

El papel de los expertos en el desarrollo de la IA generativa se reduce aún más que en el caso de la IA predictiva, puesto que son capaces de aprender patrones complejos partiendo de los datos sin necesidad de intervención o etiquetado manual por parte de expertos. Los modelos se entrenan para predecir o reconstruir partes de los datos de entrada a partir de otras partes de los mismos datos. Este enfoque aprovecha la estructura inherente de los datos para aprender representaciones útiles y generales. Por ejemplo, un modelo podría aprender a predecir la próxima palabra en una oración o completar una imagen parcialmente oculta utilizando solo el contexto proporcionado por los propios datos. Este método reduce la necesidad de elevados cantidades de datos etiquetados, que pueden ser costosos y laboriosos de producir.

La IA generativa permite a las computadoras crear contenido por sí mismas. Imaginemos que, en lugar de simplemente responder a preguntas concretas, para lo cual ha sido «entrenada» con una base de datos específica, una computadora puede interpretar imágenes, resumir textos, componer música o generar ideas sobre cualquier temática. Y

no solo eso, sino, incluso, llegar a comunicarse con los usuarios en lenguaje natural.

Tabla 1. Principales características de los distintos tipos de IA

Tipo de IA	Necesidad de datos	Tipos de datos	Rol de un experto
IA simbólica	Muy pocos. Solo sirven para validar el sistema, puesto que la información que queremos capturar está en la cabeza del experto.	Datos tabulares.	Definir de forma explícita las reglas que pueden aplicarse a un concepto.
IA predictiva	Centenares o miles de datos, para descubrir en ellos patrones que representen los conceptos que queremos aprender.	Datos tabulares. Series temporales. Imágenes o audios muy simples.	Etiquetar cada uno de los datos con el concepto que le corresponde.
Aprendizaje profundo	De decenas de miles a millones de datos, para descubrir en ellos patrones que representen los conceptos que queremos aprender.	Imágenes, audio, vídeos, textos.	Etiquetar cada uno de los datos con el concepto que le corresponde.
IA generativa	Muchos millones de datos.	Imágenes, audio, vídeos, textos.	Ninguno.

Hoy en día existen aplicaciones prácticas de los cuatro tipos de IA que hemos visto (tabla 1), pero la mayoría corresponden a la IA predictiva o al aprendizaje profundo, siendo la IA generativa, que ha generado potencialmente más expectativas, todavía minoritaria.

Usos de la inteligencia artificial en educación

De forma general, los usos de la IA en el entorno educativo pueden caracterizarse en función del receptor de los potenciales beneficios del uso de esta tecnología.

Figura 2. Usos de la IA en educación

La *IA centrada en el estudiante* tiene como objetivo mejorar la experiencia de aprendizaje individual y personalizar la educación según las necesidades específicas de cada alumno. Esto incluye el uso de tutores inteligentes que proporcionan tutorías personalizadas y adaptativas, ayudando a los estudiantes a aprender a su propio ritmo. También se emplean plataformas de aprendizaje adaptativo que ajustan el contenido educativo en función del progreso y el rendimiento del alumno, presentando nuevos desafíos solo cuando está preparado para avanzar. Asimismo, los asistentes virtuales basados en IA pueden responder preguntas, ofrecer explicaciones adicionales y guiar al alumnado a través de materiales de estudio complejos, facilitando su comprensión y aprendizaje.

Por lo que toca a la *IA centrada en el profesor*, el objetivo es apoyar a los educadores mejorando la eficiencia y la efectividad de sus métodos de enseñanza. Esto se logra mediante herramientas que analizan el rendimiento de los estudiantes durante sus actividades formativas, identificando áreas en las cuales necesitan más apoyo, cosa que permite a los profesores ajustar sus estrategias pedagógicas. Por otra parte, aplicaciones de automatización de tareas administrativas liberan a los docentes de actividades repetitivas como calificar ejercicios o gestionar la asistencia, gracias a lo cual se pueden centrar en la enseñanza. También se utilizan plataformas de desarrollo profesional que ofrecen recursos y formación personalizada adaptados a las necesidades e intereses de cada educador.

Por último, la *IA centrada en la institución* puede optimizar la gestión y las operaciones de las instituciones educativas mediante diversas aplicaciones. Los *sistemas de gestión del aprendizaje* (LMS) integran IA para mejorar la administración de cursos, la comunicación entre alumnado y docentes, y el seguimiento del rendimiento estudiantil. Además, herramientas basadas en IA analizan grandes volúmenes de datos para predecir tendencias futuras en matriculación, identificar necesidades de recursos o proponer mejoras en el desarrollo curricular. También se emplean soluciones de IA para incrementar la seguridad en el campus por medio de sistemas avanzados de vigilancia y para facilitar el acceso inclusivo a recursos educativos destinados a estudiantes con discapacidades.

Cada uno de los modelos de IA que hemos visto puede tener un uso en estas áreas, dependiendo del caso concreto.

Riesgos

El análisis de riesgos es un paso esencial para garantizar un uso responsable y ético de la IA en la educación. Este proceso implica identificar, evaluar y mitigar los posibles impactos negativos que pueden surgir al implementar sistemas de IA en contextos educativos. En este ámbito, donde la interacción con estudiantes, docentes y comunidades es altamente sensible, es fundamental basarse en principios claros que guíen el desarrollo, la implementación y la evaluación de estas tecnologías.

La Unesco, en su *Recomendación sobre la Ética de la Inteligencia Artificial* (Unesco, 2022), propone un conjunto de principios que ofrecen un marco ético para abordar los riesgos asociados al uso de la IA, también en el ámbito educativo. Estos principios son clave para analizar y mitigar los riesgos vinculados con la equidad, la privacidad, la transparencia y la sostenibilidad, entre otros. Estos principios se resumen en los siguientes puntos:

- *Proporcionalidad e inocuidad:* los sistemas de IA tienen que ser proporcionales a los objetivos educativos y no causar ningún daño a los

estudiantes. Es preciso evaluar los riesgos potenciales de la IA en la educación y tomar medidas para mitigarlos.

- *Equidad y no discriminación:* los sistemas de IA deben promover la justicia social y la igualdad de oportunidades para todo el alumnado, sin discriminación por raza, género, origen o capacidad. Los Estados miembros han de fomentar el acceso inclusivo a la IA, adaptando los contenidos y los servicios al contexto local y respetando el multilingüismo y la diversidad cultural.
- *Sostenibilidad:* la IA ha de contribuir a una educación sostenible que tenga en cuenta las necesidades de las generaciones presentes y futuras.
- *Derecho a la privacidad y protección de datos:* la privacidad de los estudiantes debe ser respetada, protegida y promovida. Es esencial proteger los datos personales de los estudiantes y utilizarlos de manera responsable y ética.
- *Supervisión y decisión humana:* los docentes tienen que mantener el control sobre el proceso educativo. La IA debe ser una herramienta de apoyo, no un sustituto de los docentes.
- *Transparencia y explicabilidad:* los sistemas de IA han de ser transparentes y explicables, de modo que estudiantes y docentes comprendan por qué han tomado una determinada decisión. Esto incluye informar a los usuarios cuándo un producto o servicio ha sido proporcionado mediante IA.
- *Responsabilidad y rendición de cuentas:* se tienen que establecer mecanismos claros de responsabilidad y rendición de cuentas para el uso de la IA en la educación. Esto implica investigar y reparar los daños causados por sistemas de IA, además de establecer mecanismos de aplicación y medidas correctivas para garantizar los derechos humanos y el estado de derecho.

Más allá de estos principios, la Unesco destaca la importancia de la sensibilización y la educación sobre la IA. Esto incluye promover programas de sensibilización sobre los avances de la IA y sus repercusiones en los derechos humanos. También es fundamental fomentar la investigación sobre la IA, incluyendo la investigación sobre la ética de la IA.

Pasar de los principios éticos al uso práctico constituye una tarea compleja que comporta hacer frente a diversos desafíos tanto técnicos

como sociales y educativos. Por más que los principios como la equidad, la transparencia y la sostenibilidad ofrezca una guía de peso, su implementación en contextos reales presenta numerosas dificultades.

En primer lugar, los principios éticos acostumbran a ser amplios y abstractos, lo que deja margen para diferentes interpretaciones dependiendo del contexto. Por ejemplo, garantizar la equidad en el uso de la IA puede implicar acciones muy distintas en una escuela rural sin acceso a Internet y en un entorno urbano con recursos tecnológicos avanzados. Esta falta de claridad sobre cómo interpretar y aplicar estos principios puede dificultar su traducción en normas y prácticas concretas.

Otro reto que resaltar es la falta de recursos y capacidades en muchos sistemas educativos. La implementación de la IA requiere una infraestructura tecnológica adecuada, como dispositivos y conectividad, que no siempre está disponible en todas las regiones. Además, los docentes y administradores educativos a menudo no cuentan con la formación necesaria para utilizar herramientas de IA de modo efectivo. Incluso cuando existen los recursos técnicos, el coste elevado de desarrollo y mantenimiento de estas tecnologías puede representar una barrera, en especial en sistemas educativos con presupuestos limitados.

La complejidad técnica de los sistemas de IA también implica un obstáculo. Muchos modelos de IA, y más aún los más avanzados, como los basados en aprendizaje profundo, son difíciles de entender hasta para los expertos. Esto complica la transparencia y la explicabilidad, dos principios esenciales para generar confianza en estas tecnologías. Asimismo, los sistemas de IA dependen de los datos con los que se entrenan. Si estos datos contienen sesgos históricos o desigualdades, la IA puede perpetuar o incluso amplificar estos problemas, contradiciendo principios como la equidad.

Por otro lado, la ausencia de una regulación adecuada agrava la situación. En muchos países, las normativas sobre el uso de la IA en la educación son inexistentes o están poco desarrolladas, lo cual ocasiona incertidumbre acerca de cómo proteger la privacidad de los estudiantes o cómo garantizar la rendición de cuentas. Asimismo, las diferencias entre marcos regulatorios de distintos países hacen que cueste más implementar principios éticos de manera uniforme a escala global.

A esto se le suman los conflictos de intereses. Un gran número de herramientas de IA son desarrolladas por grandes empresas tecnológicas cuyo objetivo principal es el beneficio económico. Esto puede entrar en conflicto con principios como la sostenibilidad o la equidad, sobre todo si las tecnologías se diseñan pensando más en mercados lucrativos que en contextos educativos diversos. Por otra parte, los responsables de los sistemas educativos pueden priorizar resultados rápidos, como sería el aumento de las puntuaciones en los exámenes, sobre una implementación reflexiva y ética de estas herramientas.

Finalmente, la diversidad de los contextos educativos añade otro nivel de complejidad. Las escuelas y los estudiantes tienen necesidades y condiciones muy distintas, lo que significa que no hay una solución única que funcione para todos. Adaptar los sistemas de IA a estos contextos diversos requiere tiempo, esfuerzo y recursos adicionales, cosa que puede retrasar o dificultar su implementación.

2. ¿CÓMO FUNCIONA LA INTELIGENCIA ARTIFICIAL PREDICTIVA?

La comprensión del ciclo de vida de un sistema de inteligencia artificial predictiva es uno de los elementos básicos para tener una visión realista y crítica de la aplicación de la IA en la educación.

El ciclo consta de cuatro pasos: la recolección de un conjunto de datos relevante para resolver un problema, la creación de un modelo que codifica el conocimiento implícito en los datos, la creación de una aplicación informática que implementa el modelo y, por último, el uso de esta aplicación en un entorno concreto.

De los datos a los modelos

Para entender estos pasos, puede ser útil pensar en un ejemplo hipotético.[2] Supongamos que una universidad quiere usar datos históricos de estudiantes para construir modelos para predecir *el éxito futuro de los estudiantes que empiezan un grado*. Estos modelos pueden ser útiles para apoyar cambios en políticas y prácticas de la institución. Por ejemplo, este modelo permitiría identificar al alumnado que puede estar en riesgo de rendimiento académico bajo o de abandono, a fin de intervenir con recursos de apoyo como tutorías o asesoramiento académico desde el principio de su recorrido académico.

Nuestro primer conjunto de datos

Los *datos* de cada alumno que inicialmente pueden verse como necesarios para el objetivo definido pueden incluir elementos muy variados: variables sociodemográficas (género, raza, país de origen, número de hermanos, barrio y ciudad de residencia, etc.), notas de bachillerato, notas de selectividad, listado de actividades extraescolares que ha realizado, deportes que practica, etc. En el caso de antiguos alumnos, también disponemos de la nota media de todas las convocatorias de

2. Este ejemplo está inspirado en el caso propuesto por Google como ejemplo de sesgo algorítmico en: https://pair.withgoogle.com/explorables/hidden-bias

examen a las que se han presentado durante la realización de su grado, que puede representar el «éxito» del estudiante en la universidad. En términos de un sistema de IA, el primer conjunto de datos constituye la *entrada* del sistema (el conjunto de datos que queremos interrogar, que normalmente se denotan con la letra *X*) y el segundo, la nota media de todas las convocatorias de examen a las que se han presentado en el transcurso de su grado, son los datos de *salida* o etiquetas (los datos que constituyen la respuesta a la entrada, que se suelen denotar con la letra *Y*).

Nuestro primer modelo

El siguiente paso es la selección del tipo de *modelo* matemático que usaremos para representar el conocimiento implícito que creemos que contienen los datos. Los modelos creados mediante técnicas de IA suelen expresarse de forma matemática.

En nuestro ejemplo, podríamos empezar definiendo de forma manual, sin recurrir a ningún tipo de análisis de los datos, un modelo de éxito de un estudiante, que llamaremos *Y*:

$$Y = \textit{Nota media del bachillerato}$$

La interpretación de expresión matemática es muy simple: el éxito de un estudiante que empieza un grado, representado por el valor numérico *Y*, es igual a la nota media de las asignaturas que cursó durante el bachillerato, representado por el valor numérico *X*.

Debido a la limitación del dato de entrada *X* usado, la nota media que sacó en el bachillerato, no podemos esperar que este modelo sea muy preciso. Para evaluar esta precisión, podemos usar una base de datos de antiguos alumnos con las características antes mencionadas. Para ello, basta con comparar el valor que el modelo predice para cada alumno de la base de datos y la nota media real de su grado.

En la figura 3 podemos ver el resultado para un conjunto de 100 estudiantes. Los puntos azules corresponden a los valores (*x,y*) de los alumnos disponibles en el conjunto de datos y la línea roja corresponde a la predicción que hubiera realizado el modelo *Y* para estos datos.

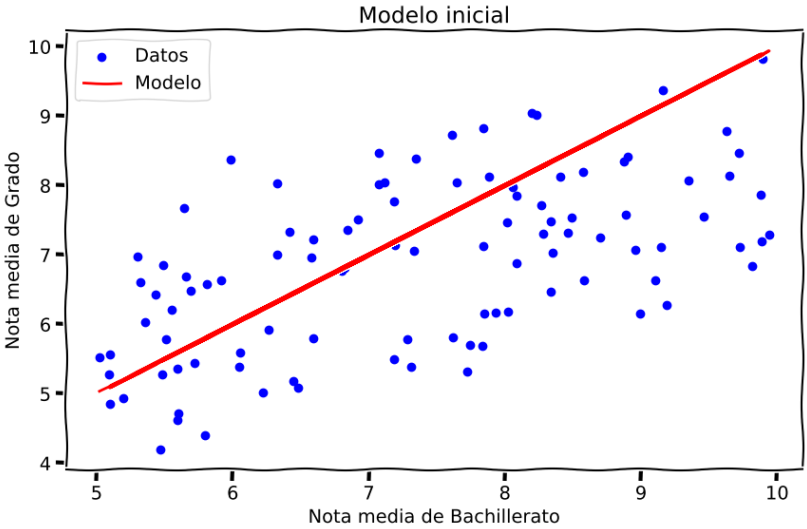

Figura 3. Ejemplo del resultado de aplicar un modelo simple (sin ningún tipo de técnica de IA)

Supongamos el caso de un estudiante que obtuvo una nota media de 7,7 en bachillerato y una calificación final de 5,2 en el grado universitario. Si nuestro modelo se hubiera utilizado para predecir la nota del grado, habría estimado un 7,7. El error de predicción en este caso habría sido de +2,5 puntos, lo que representa una clara sobreestimación del rendimiento real del estudiante.

En general, podemos observar que el modelo sobreestima de forma notable la nota de la mayoría de los estudiantes (lo cual es normal, puesto que el desempeño medio de un estudiante en la universidad es, en general, más bajo que en el bachillerato) y la subestima en algunos casos. Es más, solo es capaz de predecir la nota más o menos bien en muy pocos casos (los puntos azules que están cerca de la línea roja).

Si queremos construir un modelo más preciso, con menos error de predicción, podemos usar herramientas de IA, que nos permiten descubrir de forma automática cómo podemos diseñar el modelo *Y* a partir de la combinación de las características disponibles de cada estudiante.

¿Cómo mejorar el modelo?

Siguiendo con nuestro ejemplo, el siguiente paso hacia un modelo predictivo más preciso podría ser estimar un modelo un poco más complejo,[3] en términos matemáticos, a partir de los datos.

Es decir, si denotamos la nota media del bachillerato como x_1, pasar del modelo que teníamos,

$$Y = x_1$$

a un modelo de este tipo:

$$Y = ax_1 + b,$$

donde a y b representan dos parámetros numéricos desconocidos. En este caso, la tarea del aprendizaje predictivo es determinar los valores a y b tales que minimizan el error de predicción del modelo.

Si aplicamos uno de estos algoritmos de IA disponibles para esta tarea a un conjunto de datos históricos (que llamaremos *conjunto de aprendizaje*), obtenemos el modelo de la figura 4. El proceso que nos permite ajustar los valores a y b se llama *proceso de entrenamiento*.

¿Cómo funciona un proceso de entrenamiento?

El proceso de entrenamiento empieza por dividir los datos de aprendizaje en dos conjuntos: conjunto de entrenamiento y conjunto de prueba. El conjunto de entrenamiento se utilizará para ajustar el modelo, mientras que el conjunto de prueba se usará para evaluar su rendimiento.

El segundo paso es escoger un algoritmo de aprendizaje adecuado para el problema. Esto puede incluir modelos de regresión, árboles de decisión, redes neuronales, entre otros (Alpaydin, 2021).

A continuación, el modelo se entrena utilizando el conjunto de entrenamiento. El algoritmo ajusta sus parámetros internos para minimizar el error de predicción mediante técnicas como el descenso de gradiente.

Por último, el modelo final se evalúa con datos que no ha visto antes, para asegurar que no ha sobreajustado y que generaliza correctamente a datos nuevos.

3. A fin de no complicar la comprensión del rol de los modelos en IA, en la explicación vamos a usar modelos matemáticos muy simples, si bien, en la realidad, estos modelos serían mucho más complejos.

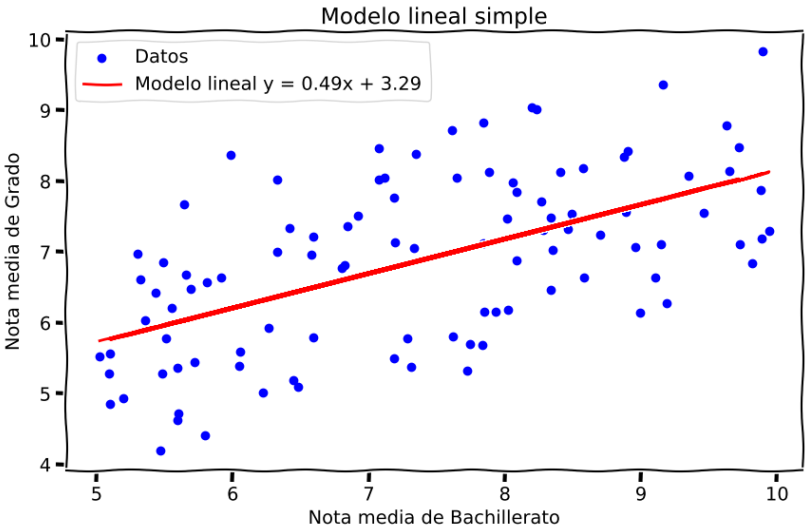

Figura 4. Modelo lineal simple obtenido con técnicas de IA predictiva

El método de aprendizaje predictivo ha determinado, a partir del análisis de los datos históricos disponibles, que si a = 0,49 y b = 3,29, obtenemos un modelo un poco mejor, representado por la línea roja. Este modelo es mejor que el anterior, puesto que no sobreestima ni subestima de forma sistemática, pero el error global que comete es igualmente muy grande (la mayoría de los puntos azules está lejos de la línea roja).

Para disminuir todavía más el error, podemos usar el resto de las características disponibles[4] (como el número de actividades extraescolares) y no limitarnos a la nota de bachillerato. Si tenemos n características distintas, podemos representar cada estudiante i por este conjunto de características: $(x_1^i, x_2^i ..., x_n^i)$.

En nuestro ejemplo, si usamos todas las características disponibles de cada estudiante para construir un modelo más complejo,

$$Y = a_1 x_1 + ... + a_n x_n + b$$

podríamos obtener un modelo con unos errores como los que se muestran en la figura 5.

4. Ello nos obliga a «complicar» el modelo matemático que usamos.

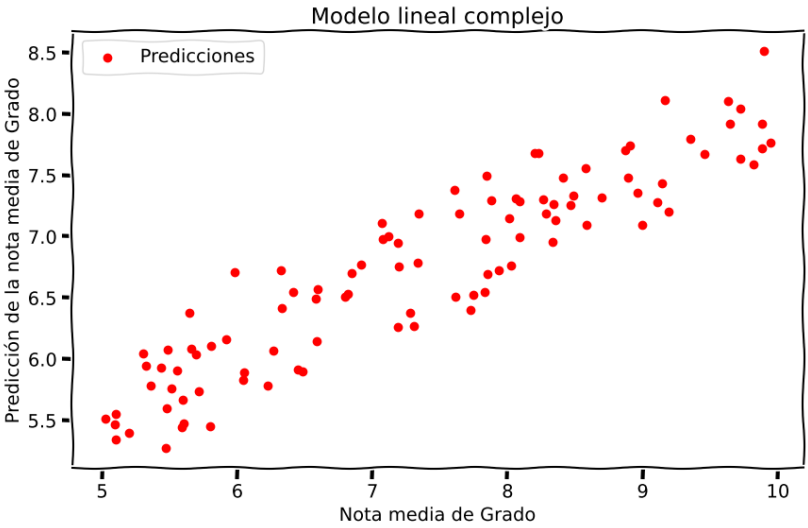

Figura 5. Error para cada uno de los datos de un modelo más complejo

En esta gráfica podemos apreciar la relación entre la nota de grado real y la que ha predicho el modelo complejo.[5] El nuevo modelo es mucho mejor que los anteriores desde todos los puntos de vista: no sobreestima ni subestima, y el error global es pequeño.

Por último, si en vez de un modelo lineal usamos el aprendizaje automático para generar un modelo mucho más flexible,[6] con mejor capacidad para modelizar las relaciones entre las características, podríamos llegar a modelos con un error de predicción muy pequeño, como el de la figura 6.

La obtención de modelos predictivos precisos que se nutren de grandes volúmenes de datos depende, en gran medida, de la pertinencia de las características seleccionadas para abordar el problema específico. La calidad de los datos juega un papel esencial en la eficacia de estos modelos. Incluso los algoritmos de aprendizaje más avanzados pueden

5. En este gráfico, un modelo predictivo perfecto generaría una línea roja diagonal.

6. No es objetivo de este libro hacer un repaso de los algoritmos de IA que pueden usarse para este fin, y que van desde los árboles de decisión hasta las las redes neuronales, puesto que los problemas que queremos discutir sobre el uso de modelos predictivos en educación son independientes de la forma concreta del modelo.

producir resultados inexactos o sesgados si los datos no son de alta calidad o si las características no están adecuadamente seleccionadas.

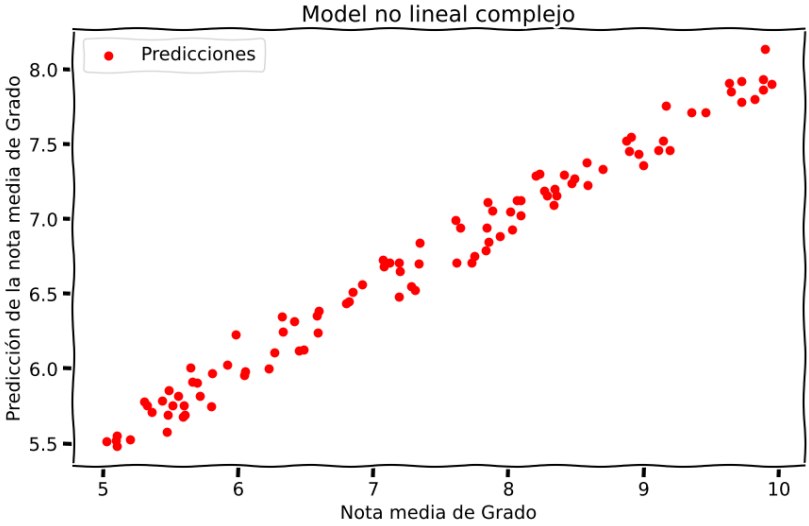

Figura 6. Error para cada uno de los datos de un modelo no lineal complejo

Del modelo al uso

Una vez disponemos de un modelo específicamente desarrollado para esta tarea, el camino hasta sus posibles usos es ya directo.

Al inicio del siguiente curso, la universidad puede construir una aplicación informática tal que, dados los datos de entrada *X* de los nuevos alumnos, haga la predicción de *Y*. Y puede poner esta aplicación a disposición de los miembros de sus facultades.

Hasta este punto, el modelo que hemos construido no tiene todavía un uso definido: será cada una de las facultades la que lo determinará. Por ejemplo, una facultad (uso 1) puede aplicar este modelo a un estudiante recién ingresado y predecir la nota media que va a sacar al cabo de cuatro años para determinar su idoneidad para recibir una beca de apoyo al estudio. Otras pueden aplicarlo a toda la cohorte de entrada (uso 2) y determinar qué recursos docentes se van a necesitar para mantener un nivel de conocimientos determinado a la salida de los estudios.

Como hemos visto en este ejemplo, no hay una correspondencia unívoca entre la tarea que se ha usado para entrenar el modelo predictivo y sus usos finales. Un mismo modelo puede ser utilizado para múltiples usos, unos más justificados y otros menos. Es en este punto donde la necesidad de un análisis del uso de los modelos predictivos se hace evidente.

3. ¿CÓMO FUNCIONA LA INTELIGENCIA ARTIFICIAL GENERATIVA?

Así como en el caso de la IA predictiva es fácil para cualquier persona construirse una idea intuitiva de su funcionalidad (dado un dato de entrada *X*, predecir el valor de salida *Y* más probable en función de los datos con los que ha sido entrenado), el caso de la IA generativa es algo más complejo, por la aparente falta de correspondencia entre la tarea con la que se entrena y los usos que se le adjudican en el mundo real. De hecho, lo curioso es que la IA generativa basada en texto no es sino un caso particular de IA predictiva, aunque sus usos son muy distintos.

La paradoja de la IA generativa radica en la relación entre su funcionamiento interno y los usos que se le asignan en el mundo real. S bien puede parecer que la IA generativa realiza tareas completamente distintas de las de la IA predictiva, en esencia, ambas se basan en principios similares de predicción.

Como ya hemos visto, la IA predictiva es relativamente fácil de entender: dado un conjunto de datos de entrada *X* (por ejemplo, el historial de un paciente), el sistema predice el valor de salida *Y* (como la probabilidad de desarrollar una enfermedad). Este enfoque refleja una relación directa entre el entrenamiento del modelo (aprender patrones en los datos) y su aplicación (predecir valores futuros basados en esos patrones).

En el caso de la IA generativa, como los modelos de generación de texto, las tareas que lleva a cabo parecen menos directamente conectadas con su proceso de entrenamiento. Por ejemplo, un modelo como GPT se entrena para predecir la palabra que sigue en una secuencia de texto dada. Sin embargo, en el mundo real, se utiliza para generar ensayos, escribir poesía, programar o responder preguntas complejas. Esto puede parecer desconectado, porque el modelo no fue explícitamente entrenado para «escribir ensayos» o «resolver problemas de programación».

La paradoja se resuelve al comprender, como apuntábamos, que la IA generativa basada en texto no es más que un caso particular de IA pre-

dictiva. Durante su entrenamiento, el modelo aprende a predecir la palabra o el fragmento de texto más probable en función del contexto proporcionado (el texto previo). Esta tarea predictiva básica, repetida a gran escala, permite al modelo construir secuencias completas que parecen creadas intencionalmente para tareas específicas.

Por ejemplo:

- Al escribir poesía, el modelo simplemente predice palabras que sean coherentes con el estilo lírico y el ritmo de las palabras previas.
- Al generar código, el modelo predice fragmentos de programas válidos y contextualmente relevantes en función del lenguaje y de las instrucciones dadas.

El aparente salto entre el entrenamiento (predecir la siguiente palabra) y los usos reales (escribir ensayos o resolver problemas) se debe a que la tarea de predecir palabras encapsula una riqueza implícita de conocimiento sobre el lenguaje, el contexto y el significado.

La relación entre los datos de entrada y la salida en un modelo de IA generativa no se especifica en forma de reglas claras y legibles para los humanos. En lugar de eso, el modelo aprende estas relaciones complejas a través de patrones en los datos durante el entrenamiento, utilizando técnicas avanzadas basadas en redes neuronales profundas (*transformers*).

Esto permite que la IA generativa produzca texto, imágenes, música y otros tipos de datos que parecen creados por personas, pero cuya lógica subyacente de generación puede costar interpretar.

Comprender a fondo la conexión entre la tarea simple para la cual han sido entrenados estos modelos y su alta competencia en otras tareas complejas no es el propósito de este libro. Con todo, es necesario desarrollar una idea intuitiva acerca de los modelos generativos que nos permita evaluar su utilidad para resolver determinadas tareas y, sobre todo, para prever sus riesgos. A tal fin, vamos a usar una analogía propuesta por Léon Bottou y Bernhard Schölkopf basada en un relato corto de J. L. Borges titulado *El jardín de los senderos que se bifurcan* (Bottou *et al*., 2023).

En el relato se describe una biblioteca que no solo contiene todos los textos producidos por los humanos, sino que, mucho más allá de lo que ya se ha escrito en la realidad, abarca también todos los textos que un humano podría leer y al menos comprender superficialmente. Esta colección de textos «plausibles» podría contener libros, diálogos, artículos, oraciones, páginas web, programas de computadora... en cualquier idioma, en cualquier forma.

Ahora imaginemos un sistema de IA que es capaz de memorizar todos estos textos y de generar, a partir de una secuencia de palabras concreta, la lista de palabras que la siguen en alguno de libros de la biblioteca, ordenadas por su frecuencia.

Si aplicamos este proceso de forma recursiva, la secuencia de palabras generada por la máquina a partir de una secuencia de entrada se encuentra en algún lugar de nuestra colección infinita de todos los textos plausibles y, por lo tanto, forma parte de uno de los libros de nuestra biblioteca.

Respuestas plausibles y respuestas correctas

Los usos y los riesgos de la IA generativa se desprende de una simple observación basada en la analogía entre los grandes modelos de lenguaje y la biblioteca de Borges: los textos que un individuo podría leer y, como mínimo, comprender de manera superficial no solo están formados por relatos relacionados con hechos «reales», sino que cualquier idea fantasiosa que un humano pueda entender, o simplemente cualquier idea que sea falsa pero comprensible, formará parte de la biblioteca. La descripción del sistema solar basada en la mecánica de Newton formará parte de la biblioteca, pero también las teorías terraplanistas. A partir de aquí es fácil entender que, ante el desconocimiento de lo que es verdad y de lo que no lo es, el programa de IA que genera el texto solo es capaz de generar «texto plausible», pero no «texto correcto». Por consiguiente, la salida de un modelo de IA generativa siempre tiene una cierta probabilidad de ser incorrecta.

Los grandes modelos de lenguaje han sido entrenados con datos textuales que no se han revisado desde el punto de vista de su realidad o factualidad. Desde esta perspectiva, podemos considerar que son datos que podrían formar parte de la biblioteca de Borges. Por ello, los datos producidos por los grandes modelos de lenguaje forman parte de la biblioteca de Borges y no es posible determinar *a priori* si van a ser reales o falsos. Desde este punto de vista, podemos decir que los datos producidos por una IA generativa cuando es interrogada por un usuario con una frase en lenguaje natural son:

- Útiles para resolver alguna tarea, si son correctos y tenemos forma de verificar su corrección.
- Inútiles, si son incorrectos y lo sabemos.
- Peligrosos, si son incorrectos y no lo sabemos.

Si pensamos en la IA generativa de esta manera y aplicamos un principio de precaución, la decisión sobre la adecuación de esta tecnología a un determinado uso, maximizando la utilidad y minimizando el riesgo, se vuelve clara y sencilla. La IA generativa es una magnífica herramienta cuando es utilizada por un usuario experto que puede emplearla como una ayuda eficaz en su trabajo. Los expertos disponen del conocimiento necesario para evaluar y contextualizar las respuestas generadas, cosa que les permite identificar posibles errores o inconsistencias, así como exprimir el potencial de la tecnología para mejorar su productividad y su creatividad.

Pero la situación cambia drásticamente cuando la IA generativa la utilizan usuarios que carecen de la capacidad para discernir la veracidad de una respuesta. En estos casos, la herramienta se vuelve peligrosa, ya que puede generar información errónea o engañosa que el usuario podría tomar como cierta si no tiene la habilidad de cuestionarla o de verificarla de un modo adecuado. Esto puede llevar a la toma de decisiones incorrectas, a la producción y difusión de información falsa, y a consecuencias negativas en diversos ámbitos, desde la educación y la investigación hasta los medios de comunicación y las redes sociales. Por lo tanto, es crucial implementar medidas que aseguren el uso responsable de la IA generativa, fomentando la capacitación de los usuarios y estableciendo mecanismos de control y verificación de la información.

Por último, es interesante darse cuenta de que la IA generativa propone un cambio en el rol de los datos, tal como se muestra en la figura 7. Los datos y sus etiquetas (producidas por humanos y relacionadas con una tarea concreta) deben estar disponibles para poder entrenar un modelo de IA predictiva y, a partir de ese momento, el modelo producido es útil para contestar preguntas relacionadas con esa tarea, y ninguna más.

En el caso de la IA generativa, la mera existencia de datos genéricos es suficiente para entrenar un modelo, y es en el momento que hacemos una pregunta cuando especificamos la tarea que queremos resolver y qué datos hacen falta para responder. En figura 8 podemos ver cómo ChatGPT contesta una pregunta para la cual no ha sido explícitamente entrenado si especificamos los datos de entrada (llamados *prompts*) y la tarea que deseamos resolver.

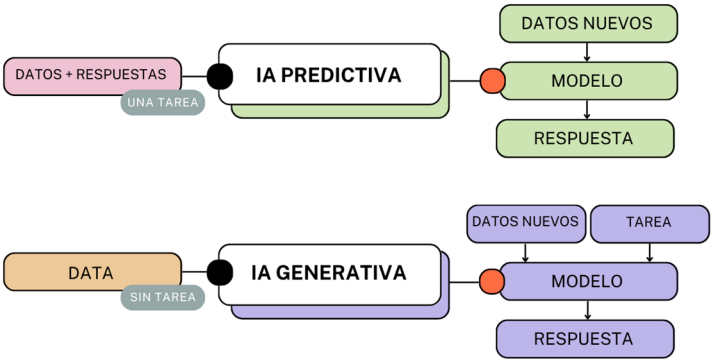

Figura 7. Relación entre datos, etiquetas y modelos en la IA predictiva y la IA generativa

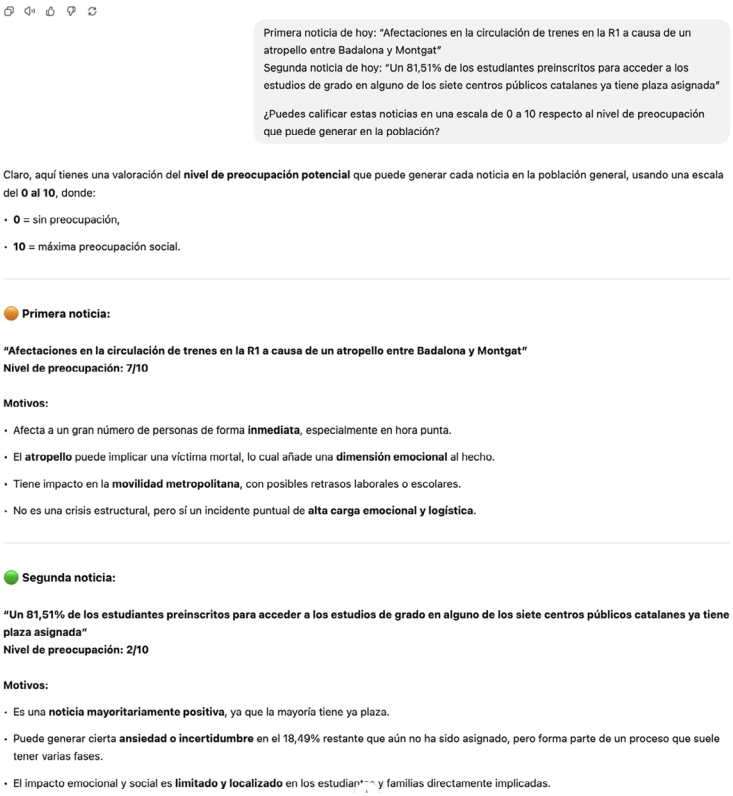

Figura 8. ChatGPT contesta una pregunta (predecir el nivel de preocupación que suscita una noticia entre los ciudadanos) para la cual no ha sido explícitamente entrenado

Ingeniería de *prompts*

Un *prompt* es una instrucción o conjunto de instrucciones proporcionadas a un sistema de IA, como un modelo de generación de texto o de imágenes, para guiar la salida o respuesta del sistema. En el contexto de la IA generativa, un *prompt* puede ser una frase, una pregunta, una descripción o cualquier tipo de entrada textual o visual que se utiliza para solicitar una respuesta específica del modelo.

A pesar de que pueda parecer que, para conseguir la respuesta deseada de un sistema de IA generativa, baste con escribir una pregunta o algún tipo de *prompt*, en la práctica no siempre es fácil que el usuario obtenga exactamente el resultado que espera. En algunos casos esto puede representar el trabajo de días o semanas. Esto ha conducido a un aumento en la demanda de trabajos relacionados con la «ingeniería de *prompts*», la cual se refiere a los procesos y técnicas para crear entradas que generen un resultado de IA generativa que se acerque lo máximo posible a la intención original del usuario.

Sobre los usos de la inteligencia artificial generativa

La IA generativa goza de un amplio rango de aplicaciones en la educación, al permitir crear contenidos, personalizar experiencias de aprendizaje y apoyar tanto a estudiantes como a profesores (Ribera *et al.*, 2024). Seguidamente, se describen algunos de los usos más destacados de la IA generativa en el ámbito educativo:

1. Creación de contenidos educativos:
   - *Generación de materiales de estudio:* dado un contexto especificado en un *prompt*, puede ser útil para ayudar al profesor en la creación de textos, imágenes, resúmenes, explicaciones y materiales didácticos, siempre y cuando estos materiales sean validados por el docente.
   - *Ayuda en la elaboración de ejercicios y problemas:* de la misma forma, puede ayudar en la proposición de problemas, cuestionarios y ejercicios prácticos adaptados a diferentes niveles de dificultad.
2. Desarrollo de entornos de aprendizaje personalizados:
   - *Tutorías virtuales:* se pueden crear tutores virtuales que pueden interactuar con los estudiantes, responder preguntas y proporcionar explicaciones adicionales en tiempo real. El uso de la IA generativa por parte del alumno ha de estar muy bien estudiado y monitorizado, puesto que este no es capaz de discernir entre respuestas correctas y respuestas plausibles pero incorrectas.
   - *Simulaciones y entornos de aprendizaje inmersivos:* se pueden generar simulaciones realistas y entornos virtuales para prácticas en campos como la ciencia, la medicina y la ingeniería.
3. Apoyo en la escritura y creación de contenidos por parte de los estudiantes:
   - *Asistencia en la escritura:* la IA generativa nos permite ir un paso más allá del corrector ortográfico para la asistencia en la escritura y crear herramientas que ayudan a los estudiantes a redactar ensayos, corregir errores gramaticales, mejorar el estilo de escritura y proponer ideas para sus trabajos. Este uso de la IA generativa siempre tiene que estar sujeto a una serie de normas relacionadas con la integridad académica y a su efectivo cumplimiento.

- *Creación de materiales para proyectos multimedia:* es posible generar imágenes, vídeos y presentaciones a partir de descripciones textuales proporcionadas por los estudiantes.

4. Evaluación y *feedback*:
   - *Generación de comentarios personalizados:* permite el análisis del trabajo de los estudiantes para proponer al profesor comentarios detallados y personalizados.
   - *Creación de exámenes y evaluaciones:* ayuda a diseñar exámenes y evaluaciones que se ajustan al nivel de conocimiento del estudiante.
5. Apoyo a la inclusión y accesibilidad:
   - *Conversión de texto a voz y viceversa:* se genera contenido en diferentes formatos para estudiantes con discapacidades visuales o auditivas.
   - *Traducción automática:* se traducen en tiempo real materiales educativos para estudiantes que utilizan otros idiomas.
6. Creación de recursos para profesores:
   - *Diseño de planes de clase:* se pueden generar de manera automática planes de clase y materiales de apoyo basados en los objetivos educativos y el currículo.
   - *Elaboración de recursos didácticos:* se pueden crear diapositivas, hojas de trabajo y guías de estudio que los docentes pueden utilizar en sus clases.
7. Motivación y compromiso de los estudiantes:
   - *Generación de historias y narrativas*: es posible la creación de historias y contenidos narrativos que pueden hacer que el aprendizaje resulte más atractivo y memorizable.
   - *Juegos educativos y gamificación:* se pueden desarrollar juegos y actividades interactivos que incentiven la participación activa del alumnado.

El caso de la evaluación basada en IA

La evaluación del alumnado mediante IA es una de las aplicaciones más prometedoras en educación. Aun así, no existe un consenso general sobre la concepción pedagógica de la evaluación con IA. Este problema se ha exacerbado en los últimos tiempos, con la aparición de la IA generativa, que ha resultado ser un elemento disruptor de muchas actividades de evaluación usadas en la actualidad.

El valor de la IA en los procesos evaluativos radica en el apoyo a los docentes y a los estudiantes, al integrar mejor la evaluación en el proceso de aprendizaje y ofreciendo una retroalimentación personalizada, más justa e inclusiva. Pero este modelo de evaluación debería estar soportado por un modelo pedagógico diferente, priorizando la personalización por encima de los modelos basados en calificaciones. Tampoco podemos despreciar la reducción de la burocracia y del tiempo dedicado a la corrección por parte del profesorado ni el hecho de que nos abre la posibilidad de hacer un seguimiento casi continuo de los estudiantes que puede permitir al profesorado conocer sus estilos de aprendizaje, identificar las dificultades y gestionar los apoyos necesarios.

Entre los riesgos y dificultades, destaca la posibilidad de que esta tecnología genere falta de confianza en los resultados por parte de los evaluados o sus familias, y también la de que se produzca un empobrecimiento si la evaluación pierde el factor humano que siempre tiene en consideración el contexto y la situación del alumno.

Aspectos sociotécnicos de la inteligencia artificial generativa

La IA generativa requiere enormes cantidades de datos y una gran potencia computacional, y eso sin contar innovaciones continuas en los métodos de entrenamiento. Estas capacidades, en su mayoría, solo están disponibles para las grandes empresas tecnológicas internacionales y unas pocas economías. Esto implica que la capacidad de crear y controlar la IA generativa está fuera del alcance de la mayoría de las empresas y países, lo cual da lugar una serie de dudas sobre algunos aspectos sociopolíticos relacionados con su uso, que van desde de la concentración de poder y conocimiento en manos de unas pocas entidades hasta el ensanchamiento de la brecha digital entre Estados (Narayanan, 2024).

Las empresas tecnológicas poseen no solo la capacidad de desarrollar sistemas de IA avanzada, sino también los recursos necesarios para hacerlo a una escala que la mayor parte de la comunidad investigadora no puede igualar. Esto incluye el acceso a potentes infraestructuras computacionales, enormes volúmenes de datos y talento especializado. Sin embargo, esta concentración de poder plantea unos cuantos desafíos significativos. Por un lado, estas empresas tienen una influencia directa

en la dirección que sigue el desarrollo de la IA y en las aplicaciones que se priorizan, lo cual puede conducir a un monopolio de la innovación. En este contexto, las decisiones sobre qué áreas explorar y qué tecnologías implementar tienden a estar condicionadas por objetivos comerciales o intereses específicos, dejando de lado las necesidades de comunidades menos representadas o las regiones con recursos limitados.

Por otro lado, esta dinámica puede perpetuar desigualdades existentes, pues los beneficios derivados de la IA generativa –como la creación de herramientas avanzadas, mejoras en la productividad o soluciones personalizadas– se diseñan desde perspectivas limitadas. Esto, aparte de restringir su utilidad en contextos diversos, amplía la brecha entre quienes tienen acceso a estas tecnologías y quienes no.

Otro aspecto crítico guarda relación con la privacidad y la ética. El desarrollo de sistemas de IA generativa acostumbra a requerir la recopilación y el procesamiento de grandes cantidades de datos, muchos de los cuales son personales y sensibles. La concentración de esta información en manos de unas pocas empresas pone sobre la mesa riesgos evidentes: desde posibles abusos de poder en forma de vigilancia masiva hasta la manipulación de información con fines políticos o comerciales.

En este escenario, es indispensable establecer marcos regulatorios que fomenten un equilibrio entre innovación tecnológica y responsabilidad social, promoviendo el acceso equitativo a la tecnología y asegurando que su desarrollo respete los principios éticos fundamentales. En caso contrario, existe el riesgo de que la IA, en lugar de democratizar las oportunidades, se convierta en un catalizador de nuevas desigualdades y tensiones globales.

Estas reflexiones también tienen que acompañar el uso de la IA generativa en el campo de la educación. Los docentes y los estudiantes deben ser capaces de tener una visión crítica respecto a los valores, las normas culturales y las costumbres sociales que están integrados en los modelos de entrenamiento de la IA generativa. Es crucial que se cuestionen y analicen a fondo estos asuntos, ya que pueden influir en los resultados generados por la IA y reforzar sesgos y prejuicios. Por ejemplo, si los datos utilizados para entrenar estos modelos provienen predominantemente de una cultura específica, es probable que los resultados hagan

prevalecer valores y perspectivas que no son universales, lo cual podría llevar a la exclusión o a la representación incorrecta de otras culturas y contextos.

Otro punto que tener en mente es que, tal y como se ha mencionado, los modelos de IA generativa se desarrollan a partir de ingentes cantidades de datos, a menudo obtenidos de Internet y generalmente sin el permiso de los propietarios de dichos contenidos. Como resultado, estos sistemas pueden haber violado derechos de propiedad intelectual, y por este motivo existen ya varias demandas internacionales en curso relacionadas con este tema. Los docentes y los estudiantes han de ser conscientes de que las imágenes o códigos generados con IA generativa pueden infringir derechos de propiedad intelectual de terceros y de que las imágenes, sonidos o códigos que crean y comparten en Internet pueden ser utilizados por las compañías que desarrollan esta tecnología sin su permiso.

Un ejemplo claro de esta problemática es el caso de herramientas como Stable Diffusion o DALL-E, que son modelos de generación de imágenes entrenados con enormes conjuntos de datos recopilados de Internet. Estos conjuntos incluyen imágenes disponibles públicamente, muchas de las cuales están protegidas por derechos de autor, como obras de artistas o fotografías comerciales. No obstante, durante el entrenamiento, estas imágenes son utilizadas sin el consentimiento explícito de los propietarios de los derechos. Otro caso concreto es el de GitHub Copilot, una herramienta de generación de código basada en IA entrenada con repositorios públicos de código de plataformas como GitHub. Aunque la herramienta es útil para programadores, en algunos casos puede generar fragmentos de código casi idénticos a los originales con los cuales fue entrenada, una circunstancia que hace inevitable preocuparse por eventuales violaciones de licencias de código abierto.

4. ¿CÓMO EVALUAMOS EL USO DE UN SISTEMA DE INTELIGENCIA ARTIFICIAL?

En capítulos anteriores hemos visto algunas de las capacidades de los sistemas de inteligencia artificial (IA), pero también algunas de sus limitaciones y riesgos (escasa precisión, falta de veracidad en sus respuestas, etc.). Hacer un uso responsable de la IA en educación, velando por que los beneficios de la IA se maximicen mientras se minimizan los riesgos y los posibles daños, es, pues, esencial para asegurar que estas tecnologías mejoran la enseñanza y el aprendizaje, respetando siempre los derechos y el bienestar de todas las personas involucradas.

El uso responsable de la IA en la educación no se garantiza simplemente siguiendo una serie de recomendaciones genéricas preestablecidas. Es un proceso que debe involucrar a diversos actores y llevarse a cabo en cada aula, escuela o institución educativa cuando se introducen nuevas aplicaciones. En otras palabras, se trata más de un proceso para hallar respuestas a una serie de preguntas específicas que de un mero proceso consistente en completar una lista de verificación predefinida. Con una metodología adecuada, la implementación responsable de la IA en la educación puede mejorar los resultados de aprendizaje, fortalecer el rol de los docentes y promover la equidad educativa. Sin dicha metodología, los riesgos pueden ser significativos.

En este capítulo se propone una metodología que puede adaptarse a múltiples casos. Cada uno de los pasos de esta metodología persigue responder a una pregunta, tal y como podemos ver en la figura 9.

Figura 9. Proceso de validación para asegurar el uso responsable de la IA en educación

En este libro se propone un proceso simple que comprende cuatro pasos: la definición del problema educativo que queremos tratar, el análisis del marco legislativo y social en el que esta aplicación se produce, el análisis de los riesgos asociados y, finalmente, la determinación de la viabilidad de la propuesta en un contexto dado.

Paso 1: Definir el problema y el valor que aporta la inteligencia artificial

El primer paso del proceso es definir de manera clara y transparente lo que se desea lograr con la IA: cuáles son los objetivos que se quieren alcanzar, mediante qué proceso y cuáles son los resultados esperados.

Por ejemplo, si el *objetivo* es mejorar la eficiencia administrativa de las tareas del profesor mediante una herramienta de apoyo al diseño de actividades durante el *proceso* de creación de materiales formativos, el *resultado* esperado podría definirse como una reducción efectiva del tiempo en tareas administrativas, manteniendo o mejorando la calidad del resultado de la tarea. De cara a fases posteriores de este proceso, es interesante expresar el resultado esperado en unos términos que se relacionen con el valor del uso de la IA, ya sea por su capacidad para personalizar el aprendizaje, por optimizar procesos, por mejorar la toma de decisiones o por fomentar la inclusión o promover la innovación.

También es importante tener en cuenta que, dado el estado actual de la IA en la educación y las muchas incógnitas que surgen de su aplicación, es útil conceptualizar el uso de una herramienta de IA en un proceso educativo como una intervención con un impacto esperado, que requiere validación bajo condiciones experimentales adecuadas. Visualizar el impacto esperado y pensar en detalle cómo se va a medir dicho impacto de la forma más objetiva posible es central, si queremos contar con evidencias que justifiquen el uso futuro de la herramienta.

Pero definir claramente el objetivo y el resultado esperado no es suficiente. En pasos posteriores hará falta comprobar su viabilidad en un contexto dado. Por ello, es interesante, en este paso, mapear los procesos involucrados, las personas que tienen algún papel en el proceso e identificar los puntos en los que la IA puede agregar valor.

En nuestro ejemplo haría falta identificar las subtareas que el profesor realiza e identificar aquellas en las que un sistema de IA predictiva o generativa pueden aportar valor: ¿en la planificación de lecciones?, ¿en la evaluación?, ¿en la comunicación con los estudiantes?

También se han de comunicar estos objetivos y resultados a todas las partes interesadas, asegurando que haya una comprensión compartida y un compromiso con la visión del proyecto. Esto ayudará a alinear los esfuerzos y a facilitar la colaboración entre las diferentes personas y equipos implicados.

Por último, se ha de establecer un plan de revisión y ajuste continuo. La tecnología y las necesidades organizativas cambian con el tiempo, por lo que es crucial revisar periódicamente el desempeño de la IA y efectuar los ajustes necesarios para cerciorarnos de que se sigue cumpliendo con los objetivos establecidos y y de que se aporta valor.

Paso 2: Comprender el marco legislativo y social

El primer elemento que se ha de considerar después de definir un uso de la IA en educación es analizar el marco legislativo y los límites que este nos impone. Algunos de los usos de la IA en educación han sido definidos por el Parlamento Europeo como usos de alto riesgo (EUR Lex, 2024):

> **La implementación de sistemas de IA en la educación es importante para ayudar a modernizar completamente los sistemas educativos, aumentar la calidad de la educación, tanto en línea como fuera de línea, y acelerar la educación digital, poniéndola, así, también a disposición de un público más amplio.** Deben clasificarse como de alto riesgo los sistemas de IA que se utilizan en la educación o la formación profesional, y en especial aquellos que determinan el acceso o **influyen sustancialmente en las decisiones sobre admisión** o distribuyen a las personas entre distintas instituciones educativas y de formación profesional o aquellos que evalúan a las personas a partir de pruebas realizadas en el marco de su educación o como condición necesaria para acceder a ella o **para evaluar el nivel apropiado de educación de una persona e influir sustancialmente en el nivel de educación y**

> **formación que las personas recibirán o al que podrán acceder, y supervisar y detectar comportamientos prohibidos de los estudiantes durante las pruebas,** ya que pueden decidir la trayectoria formativa y profesional de una persona y, en consecuencia, afectar a su capacidad para asegurar su subsistencia. Cuando no se diseñan y utilizan correctamente, estos sistemas pueden **invadir especialmente** y violar el derecho a la educación y la formación, y el derecho a no sufrir discriminación, además de perpetuar patrones históricos de discriminación, **por ejemplo, contra las mujeres, los grupos de una edad determinada, las personas con discapacidad o las personas de cierto origen racial o étnico o con una determinada orientación sexual.** (*AI Act, 2024*)

La implementación de sistemas de IA en la educación se considera capital para ayudar a modernizar los sistemas educativos, aumentar la calidad de la educación y acelerar la educación digital, poniéndola, de este modo, también a disposición de un público más amplio. Pero, cuando no se diseñan ni se utilizan correctamente, estos sistemas pueden invadir gravemente y violar el derecho a la educación y la formación, y el derecho a no sufrir discriminación, además de perpetuar patrones históricos de discriminación, por ejemplo, contra las mujeres, los grupos de una edad determinada, las personas con discapacidad o las personas de cierto origen racial o étnico o con una orientación sexual concreta.

En la actualidad, la Ley se centra específicamente en el uso de la IA para tomar decisiones en la educación que impactan en la participación del alumnado en los programas. No aborda de forma directa el uso de la IA en el aprendizaje. Sin embargo, estos aspectos no tienen unas fronteras bien definidas y no puede descartarse que en un futuro próximo se amplíen o modifiquen los criterios.

Asimismo, hemos de atenernos a las obligaciones que nos impone el marco legislativo. En el caso de que el uso previsto de un sistema de IA entre en la categoría de alto riesgo, las organizaciones responsables tendrán que disponer de un sistema de gestión de la calidad específico que asegure los siguientes puntos:

1. La existencia de un sistema de gobernanza de los datos.
2. Un nivel elevado de transparencia del sistema de IA.
3. Unos niveles elevados de robustez y precisión en los modelos de IA usados.

4. Unos niveles adecuados de ciberseguridad que protejan los sistemas de IA.
5. La existencia de un sistema de gestión de riesgos.
6. La existencia de una documentación técnica que refleje el desarrollo y mantenimiento del sistema de IA.
7. La existencia de mantenimiento de registros y trazas del funcionamiento del sistema de IA que permita su análisis.
8. Un nivel adecuado de supervisión humana en la toma de decisiones.

La presencia de este sistema de gestión de la calidad es solo el primer elemento, puesto que, antes de desplegar el sistema, la organización responsable deberá demostrar la conformidad de su aplicación ante las autoridades pertinentes. Esto quiere decir que:

- Una organización o entidad acreditada certifique que el sistema de IA cumple con los principios éticos y los estándares técnicos establecidos para asegurar la confiabilidad de la IA.
- El sistema de IA ha sido registrado en una base de datos centralizada de la Unión Europea, lo cual facilita la supervisión y el cumplimiento de las normativas.
- El sistema ha recibido la marca CE, que asegura que se cumplen las regulaciones europeas aplicables.
- Se ha hecho una evaluación sobre cómo el sistema de IA podría afectar los derechos fundamentales de las personas, asegurando que no se violen derechos como la privacidad, la no discriminación y otros derechos humanos esenciales.

Cabe destacar que el sistema de gestión de la calidad no solo afecta al desarrollo del sistema, sino que aplica a todo su ciclo de vida.

En el caso de que el uso previsto de un sistema de IA no entre en la categoría de alto riesgo, las obligaciones son menores, pero no despreciables. La organización responsable no tendrá que demostrar la conformidad de su aplicación ante las autoridades pertinentes antes de su despliegue, pero deberá ser capaz de pasar una auditoría en cualquier momento del ciclo de vida de la aplicación. Esta auditoría se centrará muy especialmente en la vertiente de la transparencia.

El marco legislativo no es el único punto en que debemos fijarnos cuando analizamos el marco de aplicación de la IA. Hay una segunda cuestión, centrada en los aspectos sociales y que es previa al marco legislativo, que se ha de responder: ¿es la IA el método más adecuado, desde un punto de vista social, para alcanzar los objetivos educativos propuestos?[7]

En el ámbito educativo la IA tiene que utilizarse para apoyar y enriquecer la experiencia de aprendizaje de los alumnos, promover su bienestar y el de los docentes, y mejorar los procesos administrativos necesarios para tal fin. Pero, más allá de estas consideraciones genéricas, cada caso concreto debe estar bien justificado. Volviendo al caso de los exámenes *A-level* del Reino Unido: ¿era el uso de un sistema predictivo como el descrito el método más adecuado para asignar los estudiantes a las universidades sin traicionar los principios que deben regir el sistema educativo?

Para responder esta pregunta, podemos hacer uso de un concepto esencial en cualquier ámbito relacionado con el bien común: la *legitimidad*. La legitimidad es el criterio que garantiza que una institución, persona o sistema tiene derecho a ejercer el poder que le ha sido otorgado de acuerdo con las leyes vigentes o los principios morales socialmente aceptados. La reacción de los estudiantes en el Reino Unido fue un caso claro de discusión de la legitimidad del uso de la IA por parte del Gobierno para un determinado tipo de decisión.

En general, podemos considerar que el uso de procesos de toma de decisiones automatizados en un escenario específico es legítimo si cumple varios criterios en dos planos distintos (Barocas, 2023):

1. Los resultados de las decisiones son:
   - *Precisos*: el sistema ofrece resultados que son correctos en la mayoría de los casos o muy cercanos al resultado ideal.
   - *Fiables*: el sistema ofrece resultados estables y coherentes en diferentes escenarios posibles.

7. Esta pregunta adquiere su dimensión real si pensamos en el problema de los exámenes *A-level* en el Reino Unido, puesto que fue un problema de este tipo el que obligó a dar marcha atrás al Gobierno británico.

- *Eficaces*: el sistema ofrece resultados que generan los efectos esperados en el mundo real.
- *Compatibles con los valores* de los individuos afectados por las decisiones.

2. El proceso de toma de decisiones está:
   - *Bien ejecutado, de forma coherente y correcta*: cuando la toma de decisiones es arbitraria en este sentido del término, los individuos sienten que están sujetos a distintos esquemas de toma de decisiones y reciben decisiones diferentes simplemente porque pasan por el proceso de toma de decisiones en distintos momentos. Este principio se basa en la creencia de que las personas tienen derecho a decisiones similares salvo que existan razones para ser tratados de forma diferente.
   - *Bien justificada*: sobre la base de unos principios compartidos y racionales. En caso contrario, diremos que la toma de decisiones es arbitraria. Las sociedades modernas no admiten la toma de decisiones arbitrarias sobre temas importantes para la vida de las personas, puesto que ello pone de manifiesto una falta de respeto hacia las personas que las reciben.[8]

Los años anteriores a la COVID-19, la decisión sobre el acceso a la educación superior en el Reino Unido se implementaba mediante un proceso burocrático, que incluía un examen y su evaluación manual, que estaba legitimado desde el punto de vista de la ciudadanía. Pero el uso del algoritmo de IA no fue percibido como legítimo por la misma ciudadanía. El origen de la falta de legitimidad del algoritmo era la *falta de precisión* (un nivel de error muy elevado en las predicciones), agravado por un sesgo negativo respecto a los estudiantes de nivel socioeconómico bajo, que fue percibido como injustificado y, por ende, *arbitrario*.

8. Cuando la toma de decisiones está relacionada con cuestiones que tienen pocas consecuencias en la vida de las personas, se pueden aceptar procesos de toma de decisiones arbitrarias. Por ejemplo, decidir mediante un sorteo cuál va a ser el orden de intervención de los alumnos en un examen oral grupal es un proceso de decisión arbitrario que puede ser aceptado socialmente.

Paso 3: Identificar los riesgos

La necesidad de un paso centrado en el análisis de las implicaciones éticas de la IA en la educación y de los riesgos asociados se vuelve evidente si se hace una simple búsqueda bibliográfica (Holmes, 2022). Cuando se exploran las noticias o la literatura actual, se observa una creciente preocupación por cómo la IA está transformando el ámbito educativo. Estas transformaciones presentan una serie de desafíos éticos que no pueden ser ignorados. Por ejemplo, si bien es cierto que el uso de algoritmos para personalizar el aprendizaje puede mejorar la experiencia educativa del alumnado, también plantea cuestiones sobre la privacidad de los datos, el sesgo algorítmico y la equidad en el acceso a estas tecnologías.

Asimismo, la integración de la IA en la educación puede alterar la relación entre docentes y estudiantes, así como los métodos pedagógicos tradicionales. Las decisiones automatizadas que afectan la evaluación y el progreso de los alumnos requieren una supervisión ética rigurosa a fin de evitar injusticias y asegurar que los procesos sean transparentes y equitativos. También es esencial considerar cómo la dependencia de la IA podría influir en el desarrollo de habilidades críticas y el pensamiento independiente del alumnado.

Sin ánimo de ser exhaustivos, los riesgos inducidos por un determinado uso de la IA en educación están relacionados con las siguientes preguntas:

Privacidad y seguridad de los datos

- ¿Cómo se protege la información personal de los estudiantes y se evita el acceso no permitido?
- ¿Se manejan de manera segura y ética los datos personales recogidos? ¿Cómo se asegura que la información personal de los estudiantes se usa solamente para los usos en los que es imprescindible?

Sesgo algorítmico

- ¿Qué factores relevantes, como género, raza o cultura, se deben tener en cuenta cuando pensamos en posibles problemas de discriminación?

- ¿Se han evaluado los algoritmos que queremos usar desde el punto de vista de la discriminación algorítmica?
- ¿Qué estrategias preventivas y de mitigación se pueden poner en marcha?

Equidad en el acceso

- ¿Existen desigualdades en el acceso a las herramientas que usaremos?
- ¿Cómo se aborda la brecha digital entre diferentes grupos socioeconómicos?
- ¿Cómo se asegura que las personas con discapacidades también se beneficien?
- ¿Son las tecnologías de IA inclusivas y accesibles para todos los estudiantes?

Transparencia en decisiones automatizadas

- Si el algoritmo sirve como sistema de ayuda en la toma de decisiones, ¿se comprende con claridad qué criterios utiliza para justificar las decisiones que propone?
- ¿Podemos asegurar que todos los procesos de decisión están sujetos a la supervisión humana?

Impacto en la relación docente-estudiante

- ¿Cómo afectan a los objetivos educativos los cambios introducidos por la IA en la dinámica de interacción y comunicación en el aula?
- ¿Cómo se redefine el papel del docente en un entorno en el que la IA juega un papel importante?

Métodos pedagógicos tradicionales

- ¿Cómo se adaptan los métodos de enseñanza existentes a esta nueva tecnología?
- ¿Se preservan y potencian las habilidades críticas y el pensamiento independiente?
- ¿Existe una potencial disminución de las habilidades sociales?

- ¿Cómo se fomenta un aprendizaje equilibrado entre tecnología y habilidades humanas?

Evaluación y progreso de los estudiantes

- ¿Son fiables y equitativas las evaluaciones automatizadas?
- ¿Cuál es el impacto de la evaluación automatizada en la motivación y desarrollo de los estudiantes?

Responsabilidad y rendición de cuentas

- ¿Cómo se educa al alumno sobre los aspectos de plagio y deshonestidad académica?
- ¿Quién se responsabiliza de las decisiones propuestas por la IA al alumno?
- ¿Se han establecido y comunicado a todos los profesores los marcos legales y éticos para la implementación de IA en la organización?

Capacitación y formación de usuarios

- ¿Es adecuada la formación para que los educadores utilicen la IA de manera efectiva y ética?
- ¿Se están desarrollando competencias digitales en estudiantes y docentes?

Algunas de estas implicaciones éticas pueden resolverse con soluciones tecnológicas (por ejemplo, proporcionando un entorno seguro que prevenga la fuga de datos personales o los ciberataques que manipulen los sistemas). Sin embargo, otras requieren un análisis ético por parte del docente y de la organización.

Paso 4: Evaluar la viabilidad

Una vez que se ha identificado el valor del uso de la IA, su seguridad jurídica y social, su alineación con las personas involucradas y también sus riesgos potenciales, es preciso evaluar la viabilidad de la propuesta tanto desde un punto de vista práctico como ético, y tomar una decisión final.

La toma de decisiones es el proceso cognitivo que nos permite seleccionar un curso de acción particular entre múltiples alternativas. Este proceso implica evaluar los pros y los contras de diferentes opciones, considerar los posibles resultados y, acto seguido, elegir una opción en particular.

Los seres humanos estamos inherentemente sujetos a una amplia gama de sesgos cognitivos que pueden nublar nuestro juicio y, por este motivo, se han desarrollado una serie de herramientas y de enfoques estructurados que pueden ayudar a generar una perspectiva más objetiva a través de la cual evaluar una decisión, minimizando los posibles sesgos. El uso de estas herramientas estandarizadas puede conducir a decisiones más consistentes, mejorando la calidad de los resultados con el tiempo.

La mejor herramienta o metodología para la toma de decisiones es aquella con la que estamos familiarizados y que sabemos usar: análisis DAFO, matrices de decisión, etc., pero, para ilustrar este paso, vamos a usar un modelo de Canvas específicamente diseñado para esta tarea: el EduIA Canvas. Los modelos de Canvas son herramientas visuales que se utilizan para diseñar, describir, analizar y gestionar modelos de negocio[9] y otros tipos de decisiones estratégicas. El término Canvas (que en inglés significa 'lienzo') alude a un cuadro o plantilla que se utiliza para estructurar y organizar ideas y conceptos de manera clara y concisa.

El uso de un modelo de Canvas sigue estos pasos:

1. *Selección del tipo de Canvas*: elegir el modelo de Canvas que mejor se ajuste a tus necesidades (por ejemplo, Business Model Canvas para un negocio completo, Lean Canvas para *startups*, etc.). En nuestro caso, definiremos un nuevo modelo de Canvas adaptado a nuestra problemática.
2. *Trabajo en equipo:* involucrar a las partes interesadas relevantes, como miembros del equipo docente, responsables de la organización o familiares de los alumnos.

9. Uno de los modelos de Canvas más conocidos es el Business Model Canvas, creado por Alexander Osterwalder, que ayuda a las empresas a delinear todos los aspectos clave de su modelo de negocio.

3. *Materiales necesarios:* preparar un lienzo grande o una pizarra, *post-it* y marcadores. También puedes usar herramientas digitales como Miro, Lucidchart o Canva para crear Canvas virtuales.
4. *Cumplimentación de los bloques:* el modelo de Canvas se compone de varios bloques o secciones que tienen que ser rellenados con información clave durante un proceso deliberativo del equipo.
5. *Revisión y validación*: has de revisar cada bloque con tu equipo y ajustar la información según sea necesario.

El modelo de Canvas EduIA

Generalmente, un Canvas se divide en bloques o secciones que representan diferentes aspectos clave de la problemática que se ha de tratar.

El modelo de Canvas que proponemos empieza con un bloque simple que nos permite clasificar el proyecto docente en una serie de clases predefinida según su objetivo genérico: mejorar los resultados de aprendizaje de los alumnos, fortalecer el rol de los docentes, promover el acceso o la equidad en la educación, o reducir la carga burocrática para concentrarse en actividades de valor añadido.

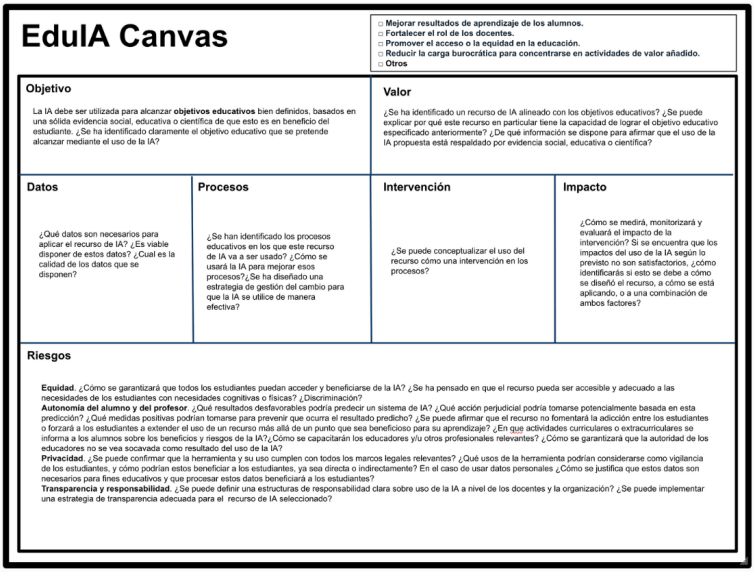

Figura 10. El modelo EduIA Canvas

Nota: El canvas se puede obtener escribiendo al autor: jordi.vitria@ub.edu

Seguidamente, tenemos el bloque formado por el objetivo de nuestro proyecto y el valor esperado.

La parte del objetivo tiene como función responder a las siguientes preguntas: dado que la IA debe ser utilizada para alcanzar objetivos educativos bien definidos, basados en una sólida evidencia social, educativa o científica de que esto es en beneficio del estudiante, ¿se ha identificado claramente el objetivo educativo que se pretende alcanzar mediante el uso de la IA?

En la parte del valor, las preguntas son: ¿se ha identificado un recurso de IA que sea útil para los objetivos educativos? ¿Se puede explicar por qué este recurso en particular tiene la capacidad de lograr el objetivo educativo especificado anteriormente? ¿De qué información se dispone para afirmar que el uso de la IA propuesta está respaldado por evidencia social, educativa o científica?

El bloque siguiente tiene por objetivo diagnosticar la madurez de nuestro entorno para adoptar la solución basada en IA.

En la primera parte se han de diagnosticar los datos necesarios (si es que son necesarios): ¿qué datos hacen falta para aplicar el recurso de IA? ¿Es viable disponer de estos datos? ¿Cuál es la calidad de los datos de que se dispone?

En la segunda, deben identificarse los procesos educativos involucrados: ¿se han identificado los procesos educativos en los que este recurso de IA va a ser usado? ¿Cómo se utilizará la IA para mejorar estos procesos? ¿Se ha diseñado alguna estrategia de gestión del cambio para que la IA se emplee de manera efectiva?

A continuación, se define el efecto esperado de la intervención propuesta y se propone una evaluación de su impacto.

Respecto a la intervención, hemos de ser capaces de responder a esta pregunta: ¿se puede conceptualizar el uso del recurso como una intervención en los procesos?

En el caso del impacto, las preguntas son: ¿cómo se medirá, monitorizará y evaluará el impacto de la intervención? Si se aprecia que los impactos del uso de la IA no son satisfactorios según lo previsto, ¿cómo

identificarás si esto se debe a cómo se diseñó el recurso, a cómo se está aplicando o a una combinación de ambos factores?

Por último, tenemos el análisis de riesgos, que ha de cubrir todos estos aspectos:

- *Equidad*
  ¿Cómo se garantizará que todos los estudiantes puedan acceder y beneficiarse de la IA? ¿Se ha pensado en que el recurso pueda ser accesible y adecuado a las necesidades de los estudiantes con necesidades cognitivas o físicas especiales? ¿Discriminación?
- *Autonomía del alumno y del profesor*
  ¿Qué resultados desfavorables podría predecir un sistema de IA? ¿Qué acción perjudicial podría tomarse potencialmente basada en esta predicción? ¿Qué medidas positivas podrían adoptarse para prevenir que se dé el resultado predicho? ¿Se puede afirmar que el recurso no fomentará la adicción entre los estudiantes o que no los forzará a extender el uso de un recurso más allá de un punto que sea beneficioso para su aprendizaje? ¿En qué actividades curriculares o extracurriculares se informa a los alumnos sobre los beneficios y riesgos de la IA? ¿Cómo se capacitará a los educadores y/u otros profesionales relevantes? ¿Cómo se garantizará que la autoridad de los educadores no se vea socavada como resultado del uso de la IA?
- *Privacidad*
  ¿Se puede confirmar que la herramienta y su uso cumplen con todos los marcos legales relevantes? ¿Qué usos de la herramienta podrían considerarse como vigilancia de los estudiantes y cómo podrían estos beneficiar a los estudiantes, sea directa o indirectamente? En el caso de usar datos personales, ¿cómo se justifica que estos datos son necesarios para fines educativos y que procesar estos datos beneficiará a los estudiantes?
- *Transparencia y responsabilidad*
  ¿Se puede definir una estructura de responsabilidad clara sobre uso de la IA a nivel de los docentes y la organización? ¿Se puede implementar una estrategia de transparencia adecuada para el recurso de IA seleccionado?

La decisión sobre la viabilidad

Una vez se dispone de un Canvas con un contenido consensuado, podemos pasar a la fase final de decisión. La toma de decisiones grupales ofrece diversas ventajas importantes tanto para las organizaciones como para los individuos involucrados. Una de las ventajas es la diversidad de perspectivas que aporta la participación de personas con diferentes antecedentes y experiencias. Esto no solo genera una amplia gama de ideas y enfoques, sino que también permite una mejor solución de problemas, ya que los grupos pueden identificar y abordar los desafíos de manera más efectiva e innovadora.

Otra ventaja significativa es el aumento de la creatividad y la innovación. La sinergia de ideas que se produce a través de la interacción entre los miembros del grupo fomenta el pensamiento colaborativo, permitiendo que se construyan soluciones nuevas a partir de las contribuciones de todos. Además, la toma de decisiones grupales estimula una mayor aceptación y compromiso, dado que los individuos involucrados tienden a sentirse más comprometidos en las decisiones finales y más dispuestos a aceptarlas, al haber participado activamente en el proceso.

Por último, la evaluación de opciones en un entorno grupal tiende a ser más completa y equilibrada. La diversidad de opiniones permite un análisis más exhaustivo de las opciones disponibles, lo cual ayuda a identificar riesgos potenciales y desventajas. Asimismo, la toma de decisiones grupales puede mitigar los sesgos individuales, ofreciendo una visión más objetiva y reduciendo la influencia de prejuicios personales en el proceso de decisión.

Cuando los miembros de un grupo tienen antecedentes similares, es más probable que sobreestimen sus habilidades y cometan errores de manera similar, cosa que puede comportar riesgos injustificados. Aumentar la diversidad del grupo disminuye estos riesgos, porque los miembros se basan en experiencias diferentes para recopilar información y resolver problemas (Page, 2008; Hong, 2004).

Para una buena toma de decisiones grupales, primero hemos de acordar una regla de decisión, esto es, el método que se utilizará para tomar la decisión final. Es clave elegir el enfoque antes de comenzar y asegu-

rarse de que todas las personas involucradas saben cuál es su rol. Las opciones más comunes son:

- *Consenso*: todos los miembros del grupo están de acuerdo.
- *Democrático*: la mayoría de los miembros del grupo están de acuerdo.
- *Consentimiento*: ningún miembro del grupo se opone.
- *Consulta*: nos basamos en opiniones de varios expertos externos. Durante la consulta, el asesoramiento generalmente se obtiene de manera independiente y después se integra, pero también puede implicar invitar a expertos a proporcionar asesoramiento consensuado en un entorno de grupo pequeño.
- *Agregación*: se agregan múltiples juicios (mediante encuestas, entrevistas, etc.) de muchas personas externas diferentes.

La elección de la regla de decisión depende de las prioridades del grupo, así como del tiempo y de los recursos disponibles.

CONCLUSIONES

¿Qué herramientas y usos de la inteligencia artificial son importantes en educación?

La inteligencia artificial (IA) está transformando el ámbito educativo, al proporcionar herramientas y aplicaciones que ofrecen nuevas posibilidades para mejorar tanto la enseñanza como el aprendizaje. Entre las más relevantes se encuentran los *sistemas de tutoría inteligente*, que brindan instrucción personalizada adaptada al ritmo, estilo de aprendizaje y necesidades de cada estudiante. Estas herramientas no solo ayudan a identificar áreas de dificultad, sino que también proporcionan recursos adicionales para facilitar el aprendizaje, haciendo que la experiencia educativa sea más eficiente y efectiva.

Otra aplicación clave es el *análisis de aprendizaje*, implementado, generalmente, en *sistemas de gestión de aprendizaje* (LMS). Estas plataformas permiten a los educadores recopilar y analizar datos sobre el rendimiento de los estudiantes, identificar patrones y tendencias, y aplicar ajustes basados en la evidencia para optimizar los métodos de enseñanza. Este análisis propicia un aprendizaje más informado y centrado en los resultados.

Además, la *evaluación automatizada* ha emergido como una herramienta esencial, pues, gracias a ella, se puede evaluar el desempeño del alumno de manera rápida y objetiva. Esto aligera la carga administrativa de los docentes y asegura que el alumnado reciba retroalimentación inmediata, lo cual es vital para su progreso académico.

El *aprendizaje adaptativo* también desempeña un papel crucial al ajustar automáticamente el contenido y las actividades educativas en función del rendimiento y las necesidades individuales de cada estudiante. De este modo, se propicia una experiencia de aprendizaje más inclusiva y personalizada. Asimismo, los asistentes virtuales y los *chatbots* facilitan tanto el apoyo académico como el administrativo, ayudando a los estudiantes a navegar por los cursos, resolver dudas frecuentes y acceder a recursos adicionales de forma rápida y eficiente.

Estas herramientas se basan en modelos de IA predictiva o de IA generativa, cuya comprensión es fundamental para evaluar sus beneficios y sus riesgos potenciales. Es esencial considerar cómo estas tecnologías pueden influir en los procesos educativos a largo plazo y garantizar que se implementen de manera eficaz y responsable.

¿Qué quiere decir hacer un uso de la inteligencia artificial responsable en educación?

Hacer un uso responsable de la IA en el ámbito educativo pasa por diseñar, construir e implementar sistemas con un enfoque ético, transparente y orientado al impacto positivo en estudiantes, docentes y la sociedad en general. Este enfoque no solo es un imperativo ético, sino que también trae beneficios tangibles, como son acelerar la innovación, mejorar la competitividad o fortalecer la confianza en las instituciones educativas.

En este contexto, establecer estándares de IA responsable en la educación es indispensable por varias razones. En primer lugar, las consideraciones éticas garantizan que los sistemas de IA se desarrollen sin prácticas perjudiciales, como la discriminación, los sesgos algorítmicos o la invasión de la privacidad. La implementación de estándares éticos minimiza el riesgo de consecuencias negativas y asegura que las herramientas de IA se utilicen para beneficiar a todos los actores involucrados.

Por otra parte, priorizar la responsabilidad en la IA fortalece la confianza en las instituciones educativas y su reputación. Las organizaciones que adoptan prácticas responsables generan credibilidad entre el alumnado, los docentes, los padres y la sociedad en general. También es relevante el cumplimiento legal y regulatorio, ya que los marcos normativos están evolucionando rápidamente para abordar las implicaciones éticas y sociales de la IA. Cumplir con estas normativas, más allá de prevenir sanciones legales, también posiciona a las instituciones como líderes en innovación responsable.

¿Cómo aseguramos el uso responsable de la inteligencia artificial en educación?

Garantizar el uso responsable de la IA en el ámbito educativo requiere un enfoque integral y participativo. Este proceso no puede limitarse a seguir una lista estática de recomendaciones, sino que debe adaptarse a las particularidades de cada aula, institución o contexto educativo.

El punto de partida es un proceso deliberativo que aborde determinadas preguntas clave: ¿qué objetivos educativos se quieren alcanzar? ¿Qué valor aportará la IA a la enseñanza y al aprendizaje? ¿Es viable técnica y procedimentalmente? ¿Qué riesgos éticos o sociales podrían surgir? Estas cuestiones ayudan a guiar la implementación de la IA y a anticipar posibles desafíos.

El uso del modelo EduIA Canvas puede ser especialmente útil en este proceso, por cuanto facilita la planificación estructurada y colaborativa de proyectos de IA, permitiendo a las instituciones diseñar estrategias que consideren tanto los beneficios como los riesgos asociados.

Con una metodología adecuada, la IA posee el potencial de transformar la educación, al mejorar los resultados de aprendizaje, apoyar a los docentes en su labor y promover la equidad educativa. Por contra, la ausencia de un enfoque bien estructurado puede derivar en riesgos significativos, como la exclusión digital, la perpetuación de desigualdades y la dependencia excesiva de tecnologías no probadas.

En consecuencia, el compromiso con un uso responsable de la IA tiene que ser continuo, integrando principios éticos, evaluaciones rigurosas y adaptaciones constantes de cara a asegurar que estas tecnologías cumplan con su promesa de enriquecer la educación de manera inclusiva y sostenible.

BIBLIOGRAFÍA

Alpaydin, E. (2021). *Machine learning*. MIT Press.

Barocas, S., Hardt, M. y Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Boden, M. A. (2014). GOFAI. En: Frankish, K. y Ramsey W. M. (eds.). *The Cambridge Handbook of Artificial Intelligence* (pp. 89-107). Cambridge University Press.

Bottou, L. y Schölkopf, B. (2023). *Borges and AI*. arXiv preprint arXiv:2310.01425

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, 538(7623), 20.

Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*. Basic Books.

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG

EUR Lex (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Hao, K. (2020). *The UK exam debacle reminds us that algorithms can't fix broken systems*. MIT Technology Review.

Hastie, T., Tibshirani, R., Friedman, J. H. y Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (vol. 2, pp. 1-758). Springer.

Holmes, W., Persson, J., Chounta, I. A., Wasson, B. y Dimitrova, V. (2022). *Artificial intelligence and education: A critical view through the lens of human rights, democracy and the rule of law*. Consejo de Europa.

Hong, L. y Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46), 16385-16389.

Jackson, P. (1990). *Introduction to expert systems*. Addison-Wesley Longman.

Kennedy, B., Tyson, A. y Saks, E. (2023). *Public awareness of artificial intelligence in everyday activities*. https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities

Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Amodei, D. *et al.* (2020). *Language models are few-shot learners*. arXiv preprint arXiv:2005.14165

McCarthy, J., Minsky, M. L., Rochester, N. y Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, 27(4), 12-12.

Narayanan, A. y Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.

OpenAI (2022). *Introducing ChatGPT* [Large language model]. https://openai.com/es-ES/index/chatgpt

Page, S. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press.

Prats, M. A. y Sintes, E. (2021). *Com impulsar la transformació digital de l'escola*. https://fundaciobofill.cat/publicacions/educacio-hibrida

Ribera, M. y Díaz, O. (2024). *ChatGPT y educación universitaria: posibilidades y límites de ChatGPT como herramienta docente*. Universitat de Barcelona. IDP/ICE y Octaedro.

Richardson, R., Schultz, J. M. y Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev.*, en línea, 94, 15.

Unesco (2021). *Recomendación sobre la ética de la Inteligencia Artificial*. Unesco.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. y Polosukhin, I. (2017). Attention is all you need. En: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan y R. Garnett (eds.). *Advances in Neural Information Processing Systems* (vol. 30). Curran Associates.

Zhang, A., Lipton, Z. C., Li, M. y Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press.

NORMAS PARA LOS COLABORADORES

https://bit.ly/3DKFESh

EXTENSIÓN

Las propuestas de cada cuaderno no podrán exceder **la extensión de 50 páginas (en Word)** salvo excepciones, unos 105.000 caracteres; espacios, referencias, cuadros, gráficas y notas, inclusive.

PRESENTACIÓN DE ORIGINALES

Los textos han de incluir, en formato electrónico, un **resumen** de unas diez líneas y tres palabras clave, no incluidas en el título. Igualmente, han de contener el **título**, un ***abstract*** y tres ***keywords*** en inglés.

Respecto a la **manera de citar y a las referencias bibliográficas**, se han de remitir a las utilizadas en este cuaderno.

EVALUACIÓN

La aceptación de originales se rige por el **sistema de evaluación externa por pares**.

Los originales son leídos, en primer lugar, por el **Consejo de Redacción**, que valora la adecuación del texto a las líneas y objetivos de los cuadernos y si cumple los requisitos formales y el contenido científico exigido.

Los originales se someten, en segundo lugar, a la **evaluación de dos expertos** del ámbito disciplinar correspondiente, especialistas en la temática del original. Los autores reciben los comentarios y sugerencias de los evaluadores y la valoración final con las correcciones y cambios oportunos que se han de aplicar antes de ser aceptada su publicación.

Si los cambios exigidos son significativos o afectan a buena parte del texto, el nuevo original se somete a evaluación de dos expertos externos y de un miembro del Consejo de Redacción. El proceso se lleva a cabo como «doble ciego».

REVISORES

https://bit.ly/3oF4izw

CUADERNOS DE DOCENCIA UNIVERSITARIA 51

¿CÓMO HACER UN USO RESPONSABLE DE LA INTELIGENCIA ARTIFICIAL EN EDUCACIÓN? ASPECTOS TÉCNICOS Y ÉTICOS DE LA UTILIZACIÓN DE LA INTELIGENCIA ARTIFICIAL EN EL AULA

JORDI VITRIÀ MARCA

Octaedro Editorial

Institut de Desenvolupament Professional
UNIVERSITAT DE BARCELONA