La inteligencia artificial en la educación: potencial transformador, riesgos de sesgo y desafíos éticos

Inteligência artificial na educação: potencial transformador, riscos de discriminação e desafios éticos

Artificial Intelligence in Education: Transformative Potential, Bias Risks, and Ethical Challenges

Elena Del Valle https://orcid.org/0000-0002-8759-6171
Universidad Europea (UE), España

Resumen. La inteligencia artificial (IA) se ha posicionado como una herramienta disruptiva en múltiples ámbitos, siendo la educación uno de los sectores donde su impacto puede ser más transformador. Desde la personalización del aprendizaje hasta la automatización de procesos administrativos, las aplicaciones de la IA en el ámbito educativo plantean oportunidades inéditas, pero también riesgos considerables, en especial relacionados con el sesgo algorítmico. Este artículo explora, desde una perspectiva crítica, el potencial de la IA en la educación, las implicaciones del sesgo algorítmico y la necesidad urgente de marcos éticos y regulatorios. Se integran hallazgos recientes, experiencias prácticas y debates contemporáneos, con el objetivo de fomentar una implementación responsable, inclusiva y centrada en el ser humano.
Palabras clave: inteligencia artificial; educación; sesgo algorítmico; ética digital; personalización del aprendizaje.

Resumo. A inteligência artificial (IA) tem se posicionado como uma ferramenta inovadora em muitas áreas, sendo a educação um dos setores em que seu impacto pode ser mais transformador. Da personalização do aprendizado à automação de processos administrativos, os aplicativos de IA na educação apresentam oportunidades sem precedentes, mas também riscos consideráveis, especialmente relacionados ao viés algorítmico. Este artigo explora, a partir de uma perspectiva crítica, o potencial da IA na educação, as implicações do viés algorítmico e a necessidade urgente de estruturas éticas e regulatórias. São integradas descobertas recentes, experiências práticas e debates contemporâneos, com o objetivo de promover uma implementação responsável, inclusiva e centrada nas pessoas.
Palavras chave: inteligência artificial; educação; viés algorítmico; ética digital; personalização do aprendizado.

Abstract. Artificial Intelligence (AI) has emerged as a disruptive tool across multiple domains, with education standing out as one of the sectors where its impact can be most transformative. From personalized learning to the automation of administrative tasks, AI applications in education present unprecedented opportunities, but also considerable risks—especially those related to algorithmic bias. This article critically examines the potential of AI in education, the implications of algorithmic bias, and the urgent need for ethical and regulatory frameworks. Recent findings, practical experiences, and current debates are integrated to promote a responsible, inclusive, and human-centered implementation of AI technologies in educational settings.
Keywords: artificial intelligence, education, algorithmic bias, digital ethics, personalized learning

1. Introducción

La inteligencia artificial (IA) ha dejado de ser una promesa futurista para convertirse en una realidad cotidiana, con aplicaciones transversales que impactan desde el ámbito industrial hasta el doméstico. En la educación, la IA abre la posibilidad de reconfigurar modelos pedagógicos tradicionales, permitiendo personalizar trayectorias de aprendizaje y automatizar procesos administrativos. Sin embargo, su implementación trae consigo importantes interrogantes éticos, en particular cuando se trata de sesgos implícitos en los algoritmos (Luckin et al., 2016). Estos sesgos pueden replicar o incluso intensificar desigualdades estructurales preexistentes, afectando directamente la equidad educativa (Holstein et al., 2019). Por tanto, el análisis de su aplicación debe ir más allá del entusiasmo tecnológico para considerar implicaciones sociales profundas.

Desde un enfoque sociotécnico, la IA no puede ser comprendida como un mero instrumento neutral, sino como una tecnología moldeada por las relaciones de poder que atraviesan su diseño, entrenamiento y aplicación (Noble, 2018). En el campo educativo, esto adquiere una relevancia crítica, dado que las decisiones algorítmicas pueden condicionar trayectorias académicas, diagnosticar supuestas competencias o incluso mediar el acceso a oportunidades formativas. La aparente objetividad de los sistemas basados en datos suele invisibilizar los valores y supuestos ideológicos que subyacen a su funcionamiento. Como señala Williamson (2020), la educación impulsada por datos tiende a reducir al estudiante a una unidad cuantificable, desplazando dimensiones éticas, culturales y afectivas que son inherentes al acto educativo.

Por otro lado, la narrativa tecnocrática dominante ha promovido una visión de la IA como solución mágica a problemas educativos complejos, sin considerar los contextos socioculturales en los que se inserta. Este enfoque ahistórico y despolitizado desconoce que las brechas digitales no son solo técnicas, sino profundamente estructurales y vinculadas con desigualdades de género, etnia, clase y discapacidad (Eubanks, 2018). En este sentido, aplicar IA en la educación sin una mirada crítica implica el riesgo de reforzar dispositivos de exclusión, esta vez mediada por algoritmos. De allí la necesidad urgente de abordar la inteligencia artificial desde marcos epistemológicos éticos, feministas y de justicia social que reconozcan la diversidad de saberes, subjetividades y trayectorias que habitan la escuela y sus periferias.

2. Metodología

El presente trabajo se inscribe en una **metodología cualitativa de carácter documental y analítico**, orientada a comprender la inteligencia artificial (IA) en educación no como un fenómeno meramente técnico, sino como una construcción sociotécnica que refleja relaciones de poder, ideologías y valores culturales (Noble, 2018; Crawford, 2021). Desde esta perspectiva, el análisis no busca medir impactos cuantificables, sino **interpretar los discursos, tensiones y posibilidades éticas** que emergen en la intersección entre tecnología, pedagogía y justicia social.

La **recopilación de información** se sustentó en una revisión exhaustiva de literatura académica y documentos institucionales recientes, publicados entre 2016 y 2023. Se consultaron bases de datos internacionales como *Scopus*, *ERIC* y *Google Scholar*, priorizando investigaciones que abordaran el papel de la IA en los procesos educativos desde una mirada crítica, así como reportes de organismos multilaterales –entre ellos la UNESCO (2021b)– que orientan la gobernanza ética de estas tecnologías. La selección de fuentes respondió a **criterios de relevancia temática, rigor metodológico y diversidad geográfica**, con el propósito de evitar la reproducción de perspectivas eurocéntricas o tecnocráticas que dominan gran parte del debate (Mohamed et al., 2020).

Asimismo, dentro del proceso de revisión y análisis documental, se prestó especial atención a las investigaciones sobre el uso de **chatbots educativos** como herramienta de mediación pedagógica. Estos sistemas conversacionales se examinaron no solo desde su eficacia técnica, sino también desde su potencial para reconfigurar las interacciones entre docentes, estudiantes y conocimiento. En particular, se analizaron estudios que problematizan la aparente neutralidad de los chatbots y sus limitaciones

para sostener diálogos empáticos o culturalmente situados (Følstad y Brandtzæg, 2017; Selwyn, 2019). Este enfoque permitió considerar a los chatbots no solo como ejemplos de aplicación de IA, sino como **artefactos discursivos** que materializan ciertas visiones sobre la enseñanza, la automatización y la relación educativa.

El **análisis de la información** se desarrolló mediante una lectura hermenéutica e interpretativa, buscando identificar patrones discursivos y dilemas éticos recurrentes en la literatura revisada. Este proceso se apoyó en una **triangulación teórica** entre tres marcos principales: (a) la justicia algorítmica y los estudios de equidad en sistemas de IA (Binns, 2018; D'Ignazio y Klein, 2020), (b) la pedagogía crítica freireana, entendida como praxis emancipadora frente a la tecnificación del aprendizaje (Freire, 1970/2014), y (c) los debates contemporáneos sobre ética digital y gobernanza educativa (Selwyn, 2019; Williamson y Eynon, 2020).

Esta articulación permitió **analizar la IA educativa como campo de disputa política y epistémica**, donde se entrelazan promesas de personalización con riesgos de exclusión estructural. El proceso metodológico no se limitó a describir usos de la tecnología, sino que buscó **comprender las condiciones de posibilidad** de su adopción: quién diseña los algoritmos, desde qué lógicas y con qué implicaciones para la equidad educativa.

Finalmente, los hallazgos fueron **organizados temáticamente** en torno a tres ejes que estructuran la reflexión: (1) las aplicaciones actuales de la IA en los ecosistemas educativos, (2) los riesgos de sesgo algorítmico y su impacto en la equidad, y (3) la necesidad de marcos éticos y de gobernanza democrática. Este enfoque metodológico posibilitó un diálogo entre la teoría crítica y la práctica educativa, con el objetivo de promover una **inteligencia artificial humanizada**, que reconozca la pluralidad de saberes, trayectorias y subjetividades que habitan la escuela.

3. Aplicaciones actuales de la IA en el ecosistema educativo

Las aplicaciones de la IA en educación se han diversificado a lo largo de la última década. Entre las más utilizadas se encuentran los sistemas de tutoría inteligente, los chatbots educativos, las plataformas de aprendizaje adaptativo y las analíticas de aprendizaje (Zawacki-Richter et al., 2019). Estas herramientas permiten recoger datos en tiempo real sobre el rendimiento y comportamiento de los estudiantes, ajustando así los contenidos y actividades propuestas. Si bien estas tecnologías prometen una mejora en la eficiencia del aprendizaje, su eficacia depende de la calidad y diversidad de los datos utilizados para entrenar los modelos. Cuando estos datos no representan a toda la población estudiantil, se corre el riesgo de generar recomendaciones sesgadas que benefician solo a ciertos grupos (Baker y Smith, 2019).

El despliegue de sistemas de tutoría inteligente (ITS) ha sido uno de los desarrollos más destacados de la inteligencia artificial en el ámbito educativo. Estas plataformas, al adaptar automáticamente las rutas de aprendizaje a partir del desempeño del estudiante, buscan emular ciertos aspectos de la enseñanza personalizada. Investigaciones recientes destacan mejoras significativas en el aprendizaje de conceptos abstractos, particularmente en áreas STEM (Woolf et al., 2021). Sin embargo, la eficacia de estos sistemas sigue siendo desigual, especialmente cuando se aplican en contextos con baja conectividad, escasa capacitación docente o sin sensibilidad

cultural. El riesgo de que estas soluciones refuercen modelos pedagógicos prescriptivos y conductistas también ha sido señalado por críticos de la "automatización de la enseñanza" (Selwyn, 2019).

Los chatbots educativos, por su parte, han ganado popularidad en entornos virtuales de aprendizaje por su capacidad de ofrecer respuestas instantáneas, simular tutorías y apoyar tareas administrativas como el recordatorio de fechas o seguimiento de progresos. Si bien su implementación puede reducir la carga cognitiva del profesorado y aumentar la disponibilidad de ayuda para los estudiantes, su efectividad pedagógica aún es objeto de debate. Diversos estudios indican que la interacción con bots puede generar frustración si no se alcanzan niveles adecuados de procesamiento del lenguaje natural, o si se percibe una falta de empatía o personalización (Følstad y Brandtzæg, 2017). Además, como toda tecnología basada en datos, los chatbots corren el riesgo de operar sobre patrones estadísticos que no siempre reflejan trayectorias diversas ni contextos marginalizados.

Otro de los pilares de la IA educativa contemporánea son las plataformas adaptativas, que ajustan en tiempo real los contenidos, secuencias y niveles de dificultad. Estas plataformas utilizan modelos probabilísticos para inferir el nivel de competencia del estudiante y adaptar la experiencia de aprendizaje en consecuencia. No obstante, como advierten Holmes et al. (2021), estas adaptaciones pueden simplificar excesivamente la complejidad del aprendizaje humano, omitiendo variables como el interés, la creatividad, o el entorno familiar. Asimismo, si las decisiones de la plataforma no son transparentes, los estudiantes y docentes pueden quedar atrapados en circuitos automatizados sin posibilidad de agencia o ajuste manual, lo que refuerza la lógica de caja negra que muchas veces caracteriza a los sistemas de IA.

4. Ventajas de la IA en procesos de enseñanza-aprendizaje

La capacidad de la IA para personalizar el aprendizaje constituye uno de sus aportes más significativos. Al analizar patrones de interacción, errores frecuentes y ritmos de avance, los algoritmos pueden ajustar los contenidos y niveles de dificultad de manera individualizada (Chen et al., 2020). Esto resulta particularmente beneficioso en contextos de diversidad, donde las necesidades educativas son heterogéneas. Además, la IA puede liberar a los docentes de tareas repetitivas como la corrección de ejercicios estructurados o la gestión de plataformas, permitiéndoles enfocarse en la dimensión pedagógica y afectiva del proceso educativo (Selwyn, 2019). Sin embargo, para que estas ventajas se materialicen equitativamente, es indispensable un diseño ético y contextualizado de estas herramientas.

La promesa de personalización algorítmica ha sido ampliamente promovida por desarrolladores y organismos internacionales como un mecanismo para mejorar el rendimiento académico y la retención estudiantil. En efecto, al procesar grandes volúmenes de datos sobre desempeño, participación e incluso emociones, los sistemas de IA pueden ofrecer itinerarios personalizados, lo que se alinea con los principios del Universal Design for Learning (UDL) y el enfoque de atención a la diversidad (Rose y Meyer, 2002). Sin embargo, la efectividad de estas estrategias depende de múltiples factores, entre ellos, la calidad del diseño instruccional, la contextualización cultural de los contenidos y la forma en que se comunica la retroalimentación. De

no cuidarse estos elementos, la personalización corre el riesgo de convertirse en una segmentación superficial basada en etiquetas y no en verdaderos procesos de diferenciación pedagógica.

Además, la automatización de tareas como la retroalimentación inmediata o la generación de informes de progreso puede liberar tiempo valioso para que los docentes se concentren en el acompañamiento emocional, la motivación intrínseca y el pensamiento crítico de sus estudiantes. Esta reconfiguración del tiempo docente es clave, sobre todo en sistemas educativos con altas ratios profesor-alumno o escasez de recursos humanos. Sin embargo, también implica un cambio profundo en las competencias profesionales requeridas. Como señalan Holmes et al. (2021), los docentes no solo deben aprender a "usar" estas tecnologías, sino también a comprender sus fundamentos, detectar sesgos, interpretar outputs algorítmicos y tomar decisiones informadas a partir de ellos. En este sentido, la personalización efectiva no es un proceso automatizado, sino una colaboración inteligente entre humanos y máquinas.

Finalmente, el uso intensivo de datos para personalizar el aprendizaje plantea interrogantes relevantes sobre privacidad, consentimiento y vigilancia. Cuando los algoritmos trazan perfiles de comportamiento a partir de datos recogidos continuamente, existe el riesgo de configurar una pedagogía basada en la predicción, donde los estudiantes son definidos por sus trayectorias pasadas y no por su potencial de transformación. Tal como advierten Williamson y Eynon (2020), el aprendizaje personalizado puede devenir en "aprendizaje predeterminado", si no se establecen límites claros sobre qué se mide, cómo se interpreta y quién decide. La personalización significativa, por tanto, exige no solo sofisticación técnica, sino también claridad ética y compromiso político con la equidad y la dignidad de cada estudiante.

5. Sesgo algorítmico: un riesgo sistémico

El sesgo algorítmico es uno de los principales desafíos asociados al uso de IA en educación. Los algoritmos aprenden de datos históricos que, en muchos casos, reflejan estructuras de poder, discriminación y desigualdad (Binns, 2018). Así, una herramienta predictiva puede perpetuar estereotipos de género, raza o clase si no se diseña con criterios de equidad. Por ejemplo, un sistema que predice riesgo de abandono escolar basado en datos pasados puede penalizar injustamente a estudiantes provenientes de entornos vulnerables. La complejidad adicional radica en que estos sesgos son frecuentemente invisibles para los usuarios finales, lo que dificulta su identificación y cuestionamiento (Eubanks, 2018). En contextos educativos, donde la equidad es un principio rector, este tipo de errores no son triviales: tienen impacto directo en trayectorias vitales.

El origen del sesgo algorítmico no se encuentra únicamente en los datos, sino también en las decisiones humanas que configuran los sistemas: qué variables se priorizan, cómo se clasifican los comportamientos, qué objetivos se optimizan. Como señala Crawford (2021), los algoritmos son artefactos sociales que cristalizan visiones del mundo, muchas veces impregnadas de prejuicios inconscientes y lógicas excluyentes. En educación, esto se vuelve especialmente problemático cuando los modelos predictivos se utilizan para distribuir oportunidades, asignar recursos o tomar decisiones

sobre la permanencia estudiantil. Un algoritmo que predice "fracaso" puede funcionar como profecía autocumplida si determina la exclusión o el etiquetado anticipado de ciertos perfiles estudiantiles, reforzando estigmas institucionales ya existentes.

A esto se suma el problema de la opacidad algorítmica. Muchos sistemas utilizados en plataformas educativas funcionan como **cajas negras**, cuyos procesos internos no son accesibles ni comprensibles para docentes, estudiantes o familias. Esta falta de explicabilidad impide cuestionar o corregir decisiones automatizadas que puedan ser injustas. Como advierten Selbst et al. (2019), los marcos legales actuales no siempre contemplan la rendición de cuentas algorítmica, lo cual genera una zona de impunidad técnica que resulta incompatible con los principios de justicia educativa. En este contexto, la ética algorítmica no puede limitarse a "corregir errores" a posteriori, sino que debe implicar una vigilancia activa desde el diseño, con participación plural, enfoques interseccionales y auditorías sociales continuas.

Por otra parte, el sesgo algorítmico no solo afecta a los sujetos, sino también a los contenidos y prácticas pedagógicas. La curaduría automatizada de recursos educativos, las sugerencias de evaluación o los sistemas de detección de plagio pueden invisibilizar saberes no hegemónicos, limitar el pensamiento divergente o imponer modelos homogéneos de excelencia. Este fenómeno de invisibilización ocurre cuando los algoritmos priorizan ciertos patrones lingüísticos, epistemológicos o culturales que corresponden a matrices de conocimiento dominantes –generalmente eurocéntricas y anglófonas–, dejando fuera formas alternativas de producir y validar saberes. En los entornos educativos, esto se traduce en la exclusión de perspectivas locales, saberes comunitarios y pedagogías no lineales que no se ajustan a las lógicas de estandarización de los sistemas digitales (Mohamed et al., 2020). De este modo, la inteligencia artificial no solo selecciona qué aprender, sino también **quién tiene el derecho de ser reconocido como sujeto de conocimiento**, reproduciendo jerarquías epistémicas que la educación crítica busca precisamente desmontar (D'Ignazio y Klein, 2020). Como han señalado autores desde el sur global, la IA educativa corre el riesgo de reproducir formas de colonialismo digital si no se problematiza desde una perspectiva crítica decolonial (Mohamed et al., 2020). En este sentido, resistir al sesgo algorítmico no es solo una cuestión técnica o legal, sino también epistémica y política: se trata de defender el derecho de cada sujeto a ser educado fuera del prejuicio y dentro de la dignidad.

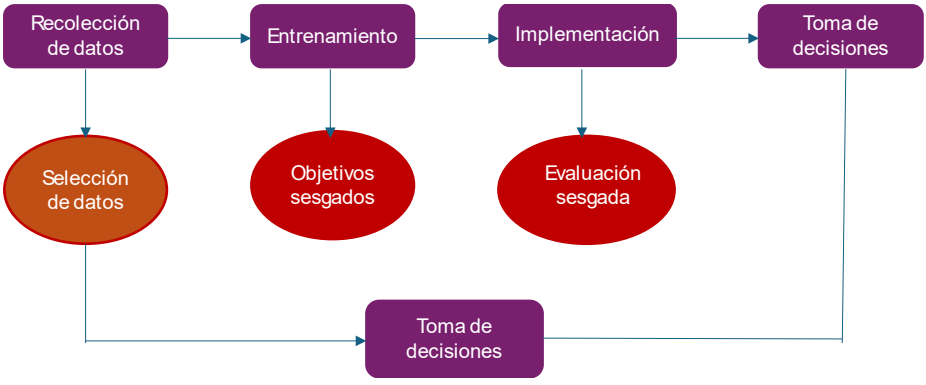

Figura 1. Ciclo de decisión algorítmica en educación y puntos críticos de sesgo
Fuente: adaptado de Crawford (2021).

La figura 1 ilustra el **ciclo de decisión algorítmica aplicado a contextos educativos**, destacando los puntos críticos donde pueden emerger sesgos estructurales. Este flujo comienza con la **recolección de datos**, etapa en la que las decisiones sobre qué información se incluye o excluye ya reflejan visiones del mundo y prioridades institucionales. Luego, durante el **entrenamiento del modelo**, los algoritmos aprenden patrones que pueden contener prejuicios implícitos si los datos de entrada no son diversos o si replican inequidades históricas (Crawford, 2021). La fase de **implementación** representa el momento en que el sistema opera dentro de entornos educativos reales –como plataformas de evaluación o recomendación–, generando outputs que influencian trayectorias formativas. Finalmente, en la etapa de **toma de decisiones**, el sistema puede reforzar exclusiones si los resultados se aplican sin mediación crítica. Este diagrama permite visualizar que el sesgo no es un error puntual, sino una posibilidad sistémica que debe ser abordada en cada fase del ciclo de vida del sistema.

6. Impacto del sesgo algorítmico en la equidad educativa

La falta de diversidad en los conjuntos de datos de entrenamiento no es un problema técnico menor, sino una cuestión política y ética. Las decisiones automatizadas –como la asignación de recursos, la admisión a programas especiales o la evaluación de desempeño– pueden consolidar barreras históricas si no se examinan críticamente los modelos utilizados (Noble, 2018). En particular, las minorías étnicas, lingüísticas o funcionales suelen ser subrepresentadas en los datos, lo que las convierte en las principales víctimas de errores algorítmicos. La consecuencia directa es una profundización de las desigualdades, disfrazada de neutralidad tecnológica. Por tanto, la equidad no puede ser una consecuencia colateral del uso de IA, sino un criterio central desde su diseño.

Sabemos que las minorías étnicas, lingüísticas y funcionales son particularmente vulnerables al sesgo algorítmico porque los datos utilizados para entrenar los sistemas de IA reflejan, de forma estructural, los privilegios y exclusiones del mundo social del que proceden. Diversas investigaciones han demostrado que los modelos de aprendizaje automático tienden a **subrepresentar o malinterpretar** a estos grupos, ya sea por su escasa presencia en los conjuntos de datos o por el uso de categorías clasificatorias que no reconocen su diversidad cultural y lingüística (Buolamwini y Gebru, 2018; Blodgett et al., 2020). En el ámbito educativo, esto se traduce en algoritmos que penalizan acentos, gramáticas no estándar o trayectorias escolares disidentes, configurando un entorno donde la diferencia es interpretada como error. La evidencia empírica acumulada revela, por tanto, que el sesgo no es una hipótesis abstracta, sino una forma concreta de desigualdad automatizada.

La infrarepresentación de ciertos grupos en los conjuntos de datos con los que se entrenan sistemas de inteligencia artificial no es solo un descuido técnico, sino el reflejo de relaciones históricas de exclusión. Como argumenta Benjamin (2019), los datos no existen en el vacío: están cargados de contextos, decisiones, omisiones y estructuras de poder. Por ello, cuando los algoritmos educativos se entrenan sobre bases de datos que privilegian ciertos cuerpos, saberes y trayectorias, lo que se automatiza no es la equidad, sino la injusticia. En este escenario, las poblaciones ra-

cializadas, con discapacidad o que no responden a los estándares dominantes corren el riesgo de ser sistemáticamente invisibilizadas o malinterpretadas por tecnologías que, paradójicamente, se promueven como neutrales.

Además, la falsa objetividad de la IA puede dificultar la intervención humana en decisiones educativas que deberían estar mediadas por criterios pedagógicos, éticos y afectivos. La automatización de procesos como la admisión a programas de excelencia, la distribución de becas o la identificación de "alumnos problema" puede llevar a una delegación excesiva del juicio educativo en sistemas que carecen de sensibilidad contextual. Como afirman Latonero y Yeung (2021), esta lógica de gobierno automatizado no solo reduce la agencia institucional, sino que contribuye a una "despolitización algorítmica", donde las decisiones aparecen como técnicas cuando en realidad son profundamente ideológicas. En educación, esta neutralidad es inaceptable: cada algoritmo debe ser tratado como un actor político que opera sobre vidas reales.

Por tanto, el diseño ético de sistemas de IA debe partir del reconocimiento de las desigualdades estructurales que atraviesan los datos. Esto implica no solo mejorar la representatividad de los conjuntos de entrenamiento, sino también **incluir activamente a las comunidades afectadas** en la definición de los problemas que se pretende resolver, en los criterios de éxito y en los mecanismos de rendición de cuentas. La equidad no puede quedar relegada a una fase posterior del desarrollo tecnológico: debe ser el eje articulador desde la concepción misma del sistema. Solo así podremos evitar que la inteligencia artificial se convierta en una herramienta de exclusión elegante y comenzar a construir tecnologías verdaderamente inclusivas y democratizadoras (D'Ignazio y Klein, 2020).

7. Ética y gobernanza de la IA en educación

Frente a estos desafíos, diversos organismos internacionales han comenzado a proponer marcos normativos para el uso ético de la IA. La UNESCO, por ejemplo, ha establecido principios como la equidad, la inclusión, la transparencia y la responsabilidad como pilares para una IA al servicio de la educación (UNESCO, 2021a). No obstante, la implementación práctica de estos principios requiere voluntad política, recursos y, sobre todo, participación activa de las comunidades educativas. La gobernanza algorítmica debe ser democrática y descentralizada, incluyendo a docentes, estudiantes, familias y expertos en derechos digitales (Williamson y Eynon, 2020). Solo así se podrá garantizar que las decisiones que afectan a los procesos educativos no queden en manos de lógicas opacas o intereses corporativos.

La gobernanza algorítmica en educación no puede quedar restringida al campo de la ingeniería ni a los marcos legales tradicionales, ya que lo que está en juego son decisiones que afectan derechos fundamentales. Por ello, resulta indispensable incorporar la perspectiva de la pedagogía crítica, entendiendo que todo diseño tecnológico con implicaciones educativas debe ser discutido en clave política y pedagógica. Como afirma Selwyn (2019), pensar la IA en educación exige recuperar la pregunta por el para qué educamos, y no solo por cómo optimizamos los procesos. Esta visión contrasta con los discursos dominantes de eficiencia y escalabilidad, que suelen relegar la dimensión humana, afectiva y situada de los aprendizajes.

Además, resulta problemático que muchas de las soluciones basadas en IA para el sector educativo provengan de empresas tecnológicas privadas, cuyas lógicas responden más al mercado que al bien común. Esta "plataformización" de la educación implica una creciente dependencia de sistemas cerrados, cuya transparencia, auditabilidad y rendición de cuentas son limitadas (Williamson, 2022). La falta de normativas claras sobre la gestión y uso de datos estudiantiles, así como la opacidad en los criterios algorítmicos de personalización o evaluación, plantea serias preocupaciones sobre vigilancia digital, autonomía pedagógica y consentimiento informado (van Dijck et al., 2018). Sin mecanismos públicos de control y evaluación crítica, corremos el riesgo de ceder la soberanía educativa a actores que no comparten los principios de la educación como derecho social.

Por tanto, avanzar hacia una IA ética en educación implica no solo diseñar regulaciones técnicas, sino promover una cultura digital democrática, donde los sujetos educativos puedan cuestionar, reinterpretar y apropiarse críticamente de la tecnología. Esto requiere fortalecer las capacidades de las comunidades educativas para entender cómo funcionan los sistemas de IA, quién los diseña, con qué datos y con qué fines. Tal como sugiere Holmes et al. (2021), se trata de desarrollar una alfabetización algorítmica que no solo forme usuarios competentes, sino también ciudadanos capaces de intervenir en la producción de tecnologías más justas y emancipadoras. En esta tarea, la educación no es solo un campo de aplicación de la IA, sino un terreno estratégico para democratizarla desde sus raíces.

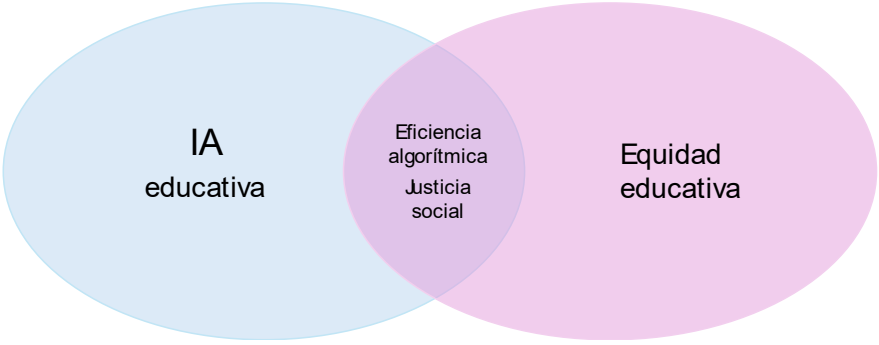

Figura 2. Intersección entre IA educativa y equidad educativa.
Fuente: inspirado en Williamson y Eynon (2020).

La figura 2 representa un **diagrama de Venn** que cruza dos grandes ejes en tensión: por un lado, el desarrollo de **sistemas de inteligencia artificial aplicados a la educación** y, por otro, los principios de **equidad, inclusión y justicia social** que deben regir los sistemas educativos. En la intersección se sitúa el espacio de **posibilidad transformadora**, donde la eficiencia algorítmica se encuentra con el imperativo de no reproducir las desigualdades estructurales. Este espacio no surge de forma automática, sino que debe ser deliberadamente construido mediante políticas públicas, marcos éticos robustos y participación activa de las comunidades escolares (Williamson y Eynon, 2020). El diagrama hace visible la necesidad de diseñar sistemas educativos inteligentes, pero también éticos, que no sacrifiquen justicia por automatización.

8. El rol del docente en la era de la IA

Contrario a la narrativa que augura la obsolescencia del profesorado frente a la automatización, la IA exige un fortalecimiento del rol docente como mediador crítico. La alfabetización digital, la comprensión de los fundamentos éticos de la IA y la capacidad de evaluar críticamente herramientas tecnológicas se convierten en competencias clave (Holmes et al., 2021). Además, el docente debe ser un actor protagónico en la selección e implementación de soluciones tecnológicas, no un mero consumidor. La inteligencia pedagógica y contextual que posee el profesorado no puede ser replicada por ningún algoritmo. Por tanto, en lugar de reemplazar, la IA debe complementar y empoderar la labor docente.

La afirmación de que los algoritmos pueden "enseñar mejor" que los docentes no solo es técnicamente cuestionable, sino epistemológicamente peligrosa. Esta idea de que los algoritmos podrían "enseñar mejor" que los docentes no surge de la investigación pedagógica, sino de los discursos tecnocráticos y corporativos que acompañan el desarrollo de la llamada *edtech industry*. Empresas tecnológicas y organismos promotores de la digitalización educativa han difundido la narrativa de la "eficiencia algorítmica" como promesa de objetividad y escalabilidad, desplazando el valor del juicio humano en el proceso de enseñanza (Luckin et al., 2016; Selwyn, 2019). Sin embargo, como advierte Williamson (2022), este tipo de retórica responde menos a una preocupación educativa que a intereses de mercado, al presentar la automatización como solución universal a problemas estructurales de los sistemas escolares. Cuestionar esa premisa es esencial para recuperar la centralidad del docente como mediador crítico y garante de la dimensión ética del aprendizaje. Enseñar no es solo transmitir información ni adaptar contenidos: es sostener vínculos, generar sentido, cuidar trayectorias y habilitar subjetividades. La pedagogía es, como afirma Paulo Freire, un acto profundamente ético y político, que no puede ser automatizado ni reducido a lógicas de eficiencia (Freire, 1970/2014). Frente a los sistemas de IA que predicen, corrigen o recomiendan, el docente aporta interpretación, escucha situada y juicio profesional, dimensiones insustituibles que hacen posible una educación verdaderamente humana.

Además, el desplazamiento simbólico del rol docente puede tener efectos nocivos en la identidad profesional y en el clima institucional. Las decisiones tecnológicas que se toman sin participación del profesorado tienden a producir resistencias, desmotivación o usos meramente instrumentales de las herramientas disponibles. Diversos estudios advierten que la apropiación significativa de la IA en las escuelas solo es posible cuando los docentes sienten que su saber es respetado, que sus contextos son comprendidos y que su experiencia profesional es valorizada como criterio legítimo de evaluación e implementación (Selwyn, 2019; Holmes et al., 2021). Incluir al profesorado desde el inicio de los procesos de diseño no es solo una cuestión técnica, sino una condición de justicia epistémica.

Finalmente, empoderar al profesorado frente a la IA implica también repensar la formación docente inicial y continua. La alfabetización digital debe ir más allá del uso operativo de plataformas, e incluir la capacidad de analizar críticamente los discursos que promueven la tecnificación de la educación, los modelos de negocio que la sostienen y las implicaciones sociales que conllevan. Como plantea Williamson y Eynon

(2020), se trata de formar docentes con pensamiento crítico algorítmico, capaces de interrogar las promesas de la IA desde una mirada pedagógica y democrática. Solo así será posible construir una cultura digital educativa donde la tecnología esté al servicio del cuidado, el pensamiento y la emancipación.

9. Ejemplos concretos de sesgo algorítmico y su relevancia para el ámbito educativo

Uno de los casos más conocidos y documentados de sesgo algorítmico es el del sistema COMPAS, utilizado en tribunales estadounidenses para predecir el riesgo de reincidencia criminal. Investigaciones demostraron que el algoritmo atribuía un mayor riesgo a personas negras que a personas blancas con antecedentes similares, perpetuando así patrones racistas en la justicia penal (Angwin et al., 2016). Aunque este caso no pertenece al ámbito educativo, sirve como advertencia clara: si algoritmos entrenados con datos históricos replican sesgos sociales, su uso en decisiones educativas puede tener consecuencias igualmente graves. Un sistema que evalúe la "probabilidad de éxito académico" podría, de manera análoga, subvalorar a estudiantes de comunidades marginadas si se basa en datos cargados de prejuicio estructural.

En el contexto educativo, se han reportado sesgos en plataformas de aprendizaje adaptativo que penalizan ciertos estilos cognitivos o perfiles lingüísticos. Por ejemplo, algoritmos de corrección automática pueden mostrar menor tolerancia a estructuras gramaticales no estándar, afectando de manera desproporcionada a estudiantes bilingües o hablantes de variedades no hegemónicas del idioma (Blodgett et al., 2020). Este tipo de sesgo no solo reduce la calidad del *feedback* pedagógico, sino que transmite un mensaje de invalidez cultural. Cuando estas herramientas se integran en procesos de evaluación sumativa, pueden impactar directamente en la calificación del estudiante y en su autoestima académica.

Otro ejemplo preocupante es el del algoritmo de predicción de calificaciones implementado por el gobierno del Reino Unido en 2020 para reemplazar los exámenes A-level cancelados por la pandemia. Este sistema asignó calificaciones en función del rendimiento histórico de las escuelas, lo que llevó a una infraevaluación masiva de estudiantes provenientes de escuelas públicas y una sobrevaloración de quienes asistían a instituciones privadas (Crawford et al., 2021). La protesta social generada por este sesgo algorítmico forzó al gobierno a revertir el modelo, evidenciando el potencial daño de tomar decisiones de alto impacto basadas en datos descontextualizados. El caso demuestra que el sesgo no es un defecto técnico menor, sino una amenaza real a los principios de justicia educativa.

También se han documentado sesgos de género en sistemas de recomendación vocacional basados en IA. Un estudio realizado por UNESCO (2021b) y el Instituto de Datos Abiertos reveló que algunas plataformas utilizadas para orientar a estudiantes hacia carreras universitarias sugieren de forma desproporcionada áreas STEM a varones y áreas de cuidado o educación a mujeres, replicando estereotipos profesionales profundamente arraigados. Estos algoritmos, al guiar expectativas y decisiones formativas, pueden limitar el acceso a trayectorias no tradicionales, especialmente en niñas y jóvenes de contextos rurales o con menor acceso a orientación vocacional presencial.

Finalmente, cabe mencionar los riesgos en el uso de sistemas de vigilancia algorítmica en entornos escolares, como herramientas de **reconocimiento facial para el control de asistencia o disciplina**. Diversos estudios han señalado que estas tecnologías presentan mayores tasas de error en personas con piel oscura o rasgos no eurocéntricos (Buolamwini y Gebru, 2018), lo cual puede traducirse en identificaciones erróneas, sanciones injustas o prácticas de hipercontrol sobre ciertos grupos estudiantiles. Más allá del sesgo técnico, se plantea una cuestión ética de fondo: ¿qué tipo de cultura escolar estamos promoviendo al delegar el cuidado, la presencia y la autoridad en sistemas de vigilancia automatizada?

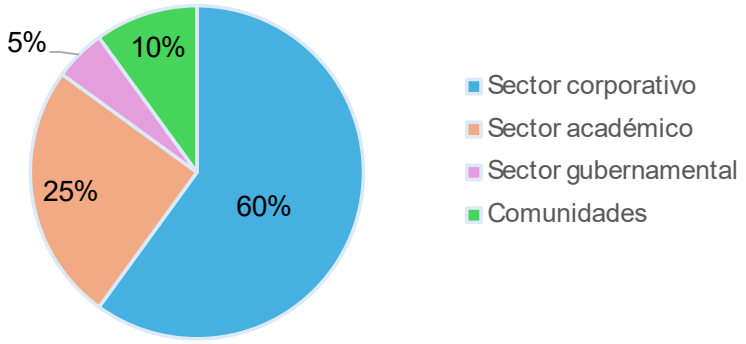

Figura 3. ¿Quién diseña la IA educativa?
Fuente: basado en van Dijck et al. (2018).

La figura 3 expone de manera visual la distribución estimada de actores involucrados en el diseño de sistemas de inteligencia artificial aplicados a la educación. El gráfico revela que el **sector corporativo** concentra la mayor parte del desarrollo tecnológico, seguido por la **academia**, los **gobiernos** y, en último lugar, las **comunidades educativas**. Esta desproporción refleja una asimetría preocupante en la toma de decisiones, ya que quienes experimentan las consecuencias cotidianas de la IA educativa –docentes, estudiantes y familias– apenas participan en su creación. Como argumentan van Dijck et al.l (2018), esta lógica de "plataformización" educativa transfiere el control pedagógico a actores privados, cuya prioridad no es necesariamente la equidad o el desarrollo humano. El gráfico invita a repensar la gobernanza de la IA desde un enfoque más participativo, democrático y situado.

10. Conclusiones y recomendaciones

La IA representa una oportunidad inédita para transformar los sistemas educativos hacia modelos más flexibles, personalizados y eficientes. Sin embargo, su implementación no está exenta de riesgos. El sesgo algorítmico, en particular, plantea desafíos profundos a los principios de equidad, justicia y democracia que deben regir la educación. Frente a ello, es urgente avanzar hacia marcos regulatorios que garanticen una IA ética, inclusiva y centrada en el ser humano. Esto implica desarrollar políticas públicas que promuevan la participación activa de las comunidades educativas en el diseño y evaluación de tecnologías, fomentar la formación docente en competencias

digitales críticas, y exigir a las empresas transparencia y responsabilidad en el desarrollo de sistemas algorítmicos. Solo así, la inteligencia artificial podrá ser una aliada real en la construcción de una educación más justa y humanizante (Cios y Zapala, 2021).

La promesa transformadora de la IA en educación no puede evaluarse únicamente en función de indicadores de rendimiento, escalabilidad o personalización. Criterios como **justicia algorítmica**, **inclusión epistémica** y **cuidado de datos** deben ocupar un lugar central en cualquier política pública orientada a su implementación. Como plantea Eubanks (2018), la automatización mal regulada corre el riesgo de reforzar mecanismos de control y exclusión, bajo la apariencia de eficiencia. La educación no puede permitirse reproducir estos patrones, pues está llamada a ser un espacio de igualdad sustantiva y emancipación. Por eso, toda estrategia tecnológica debe someterse a principios éticos que garanticen la participación y el bienestar de quienes históricamente han sido marginados por el sistema.

Asimismo, urge una vigilancia activa sobre las empresas proveedoras de soluciones de IA educativa. El creciente desembarco del sector privado en la gobernanza de datos escolares plantea serias preocupaciones sobre **soberanía digital** y **seguridad de la información**. La captura de datos estudiantiles para fines comerciales o predictivos sin consentimiento claro puede vulnerar derechos fundamentales, especialmente en contextos donde los marcos normativos son débiles o inexistentes (van Dijck et al., 2018). Es imprescindible exigir transparencia en los algoritmos, acceso a explicaciones comprensibles para usuarios no técnicos, y mecanismos independientes de auditoría y reparación ante errores o discriminaciones automatizadas.

Por último, el horizonte debe ser más ambicioso que simplemente regular los riesgos de la IA: necesitamos imaginar formas de inteligencia artificial **socialmente situada**, diseñadas desde y para los territorios. Esto implica fomentar la investigación participativa, el desarrollo de modelos abiertos y auditables, y el fortalecimiento de capacidades locales para que la producción tecnológica no sea monopolio de unos pocos. Tal como afirma Selwyn (2019), democratizar la IA en educación es una tarea política, que requiere no solo voluntad institucional, sino también el compromiso activo de comunidades educativas críticas, creativas y éticamente posicionadas. Solo así podremos avanzar hacia una inteligencia artificial al servicio de una pedagogía del cuidado, de la justicia y de la transformación.

Referencias

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. https://go.oei.int/wz8qf1pg

Baker, T., & Smith, L. (2019). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges.* NESTA. https://www.nesta.org.uk/report/education-rebooted/

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code.* Polity Press.

Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy*. Proceedings of the 2018 Conference on Fairness, Accountability and Transparency, 149-159.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454-5476. https://doi.org/10.18653/v1/2020.acl-main.485

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cios, K. J., & Zapala, M. (2021). Ethics of AI and Big Data in Education. In *Ethics of Artificial Intelligence and Robotics*. Stanford Encyclopedia of Philosophy. https://go.oei.int/iquqdnio

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crawford, R., Kallitsis, M., & McKenna, L. (2021). *Algorithmic injustice in education: The UK A-level grading scandal.* Data & Society Institute. https://go.oei.int/arhp1l53

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press. https://data-feminism.mitpress.mit.edu/

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions, 24*(4), 38-42. https://doi.org/10.1145/3085558

Freire, P. (2014). *Pedagogía del oprimido* (30.ª ed.). Siglo XXI Editores. (Obra original publicada en 1970)

Holmes, W., Bialik, M., & Fadel, C. (2021). *Artificial intelligence in education: Promises and implications for teaching and learning.* Center for Curriculum Redesign. https://go.oei.int/urx6xpik

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3290605.3300830

Latonero, M., & Yeung, K. (2021). *Governing artificial intelligence: Upholding human rights & dignity.* Data & Society Research Institute. https://datasociety.net/wp-content/uploads/2021/07/Governing-AI.pdf

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education.* Pearson Education.

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*(4), 659-684. https://doi.org/10.1007/s13347-020-00405-8

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning.* ASCD.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68. https://doi.org/10.1145/3287560.3287598

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education.* Polity Press.

UNESCO. (2021a). *Artificial intelligence and education: Guidance for policy-makers.*

UNESCO. (2021b). *AI and gender equality: A global study on the use of artificial intelligence to support women's empowerment.* https://unesdoc.unesco.org/ark:/48223/pf0000377250

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world.* Oxford University Press.

Williamson, B. (2022). Education platforms and the platformization of education policy. *Learning, Media and Technology, 47*(1), 12–25. https://doi.org/10.1080/17439884.2021.1987302

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(3), 223–235. https://doi.org/10.1080/17439884.2020.1798995

Woolf, B. P., Burleson, W., Arroyo, I., Dragon, T., Cooper, D. G., & Picard, R. W. (2021). Affect-aware tutors: Recognising and responding to student affect. In *Advances in Intelligent Tutoring Systems* (pp. 157–168). Springer. https://doi.org/10.1007/978-3-030-64452-9_10

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Cómo citar en APA:

Del Valle, E. (2025). La inteligencia artificial en la educación: potencial transformador, riesgos de sesgo y desafíos éticos. *Revista Iberoamericana de Educación,* 99(1), 79-93. https://doi.org/10.35362/rie9916838