

Sesgos algorítmicos en sistemas educativos con inteligencia artificial: Implicaciones éticas, jurídicas y pedagógicas desde la literatura reciente

Algorithmic bias in educational systems with artificial intelligence: Ethical, legal, and pedagogical implications from recent literature

Albano Torres-Gómez**

RESUMEN

Este artículo analiza la literatura reciente sobre sesgos algorítmicos en sistemas educativos con inteligencia artificial (IA), con énfasis en sus implicaciones éticas, jurídicas y pedagógicas. Mediante una revisión crítica de fuentes académicas especializadas y documentos normativos internacionales, se examinan los modos en que los modelos predictivos, la analítica del aprendizaje, la evaluación automatizada, los tutores inteligentes y la supervisión remota pueden reproducir o intensificar desigualdades previas. El argumento central sostiene que el sesgo algorítmico no constituye solo un problema de precisión técnica, sino una cuestión social y jurídica vinculada con derechos humanos, no discriminación, protección de datos, transparencia, rendición de cuentas y justicia educativa. El trabajo identifica vacíos de investigación en contextos iberoamericanos y propone criterios para una gobernanza educativa basada en auditoría, explicabilidad situada, participación de la comunidad y supervisión humana.

PALABRAS CLAVE

Inteligencia artificial educativa, sesgo algorítmico, derechos humanos, no discriminación, gobernanza algorítmica.

ABSTRACT

This article examines recent literature on algorithmic bias in educational systems with artificial intelligence (AI), emphasizing its ethical, legal, and pedagogical implications. Through a critical review of specialized academic sources and international policy documents, it analyzes how predictive models, learning analytics, automated assessment, intelligent tutoring systems, and remote monitoring may reproduce or intensify prior inequalities. The central argument is that algorithmic bias is not merely a matter of technical accuracy, but a social and legal issue connected to human rights, non-discrimination, data protection, transparency, accountability, and educational justice. The article identifies research gaps in Iberoamerican contexts and proposes criteria for educational governance based on auditing, situated explainability, community participation and human oversight.

KEYWORDS

Artificial intelligence in education, algorithmic bias, human rights, non-discrimination, algorithmic governance.

* Artículo de investigación

** Doctor en educación y maestro en comunicación por la Benemérita Universidad Autónoma de Puebla, con una estancia posdoctoral en la Universidad Nacional Autónoma de México (albano.torres@comunidad.unam.mx). <https://orcid.org/0000-0002-8451-031X>.

SUMARIO

1. Introducción
2. Metodología de revisión
3. Del sesgo técnico al perjuicio educativo
4. Implicaciones éticas y jurídicas
5. Implicaciones pedagógicas
6. Hacia una gobernanza educativa de la justicia algorítmica
7. Conclusiones
8. Referencias Bibliograficas

1. INTRODUCCIÓN

La inteligencia artificial aplicada a la educación ha dejado de ser una promesa del futuro para convertirse en infraestructura cotidiana de administración escolar, diagnóstico, evaluación, tutoría y personalización del aprendizaje; en donde los diversos sistemas tecnológicos emergentes puede ser usados por las instituciones como, plataformas adaptativas, evaluadores automáticos, asistentes generativos, tutores inteligentes y herramientas de supervisión remota que operan sobre datos de desempeño, interacción, tiempo de conexión, comportamiento y, en ocasiones, rasgos biométricos. Esa expansión no solo transforma la técnica educativa; también altera la distribución de oportunidades, apoyos, reconocimientos y cargas institucionales; ante este escenario, el sesgo algorítmico se vuelve una posibilidad real que sirve para analizar si la automatización educativa amplía derechos o consolida desigualdades ya existentes.¹

La literatura sobre inteligencia artificial educativa muestra un crecimiento sostenido en los últimos años, pero también se ha dado de forma irregular, gran parte de la investigación ha sido impulsada desde ciencias de la computación, ingeniería y minería de datos, mientras que las preguntas pedagógicas, éticas y jurídicas aparecen con frecuencia como consecuencias posteriores, no como condiciones de diseño. Esta brecha importa porque la educación no es un dominio neutral, en ella se decide quién recibe apoyo, quién es clasificado como estudiante en riesgo, quién es considerado competente, quién es vigilado con mayor intensidad y quién puede quedar excluido por datos incompletos o categorías mal definidas.²

El problema no se reduce a preguntar si un modelo acierta; en el ámbito educativo, una predicción correcta para el promedio puede resultar injusta para grupos específicos; una intervención técnicamente eficiente puede desplazar la responsabilidad institucional hacia el estudiante; y una métrica de equidad puede ser incompatible con otra cuando existen desigualdades estructurales en las tasas de acceso, permanencia o logro. La justicia algorítmica en educación exige considerar todos los pasos del proceso formativo; es decir, medición, selección de datos, definición de etiquetas, entrenamiento,

¹BAKER, RYAN S. Y HAWN, AARON, «Algorithmic Bias in Education», *International Journal of Artificial Intelligence in Education*, vol. 32, 2022, pp. 1052-1092.

²ZAWACKI-RICHTER, OLAF; MARÍN, VICTORIA I.; BOND, MELISSA y GOUVERNEUR, FRANZISKA, «Systematic Review of Research on Artificial Intelligence Applications in Higher Education: Where Are the Educators?», *International Journal of Educational Technology in Higher Education*, vol. 16, art. 39, 2019.

implementación, interpretación docente, reacción institucional y efectos acumulativos sobre trayectorias educativas.³

El objetivo de este artículo es construir un estado del arte crítico sobre sesgos algorítmicos en sistemas educativos con inteligencia artificial (IA) y derivar sus implicaciones éticas, jurídicas y pedagógicas. La pregunta que guía el análisis es ¿qué muestra la literatura reciente sobre las fuentes, manifestaciones y riesgos del sesgo algorítmico en educación, y qué criterios de gobernanza pueden orientar su prevención desde una perspectiva de derechos humanos y justicia educativa?

La premisa que este trabajo sostiene es que el sesgo algorítmico educativo debe entenderse como un fenómeno sociotécnico, ya que sus causas no se localizan únicamente en datos incompletos o modelos defectuosos, sino en la interacción entre historias de desigualdad, decisiones de diseño, finalidades institucionales, opacidad comercial, regulaciones insuficientes y prácticas pedagógicas que pueden asumir el resultado de un algoritmo como si fuera evidencia objetiva. Por ello, las respuestas puramente técnicas, aunque necesarias, son insuficientes, es por esto que se requieren mecanismos de auditoría, documentación, participación, protección de datos, supervisión humana y deliberación pedagógica.

2. METODOLOGÍA DE REVISIÓN

El artículo se desarrolla mediante un proceso de revisión crítica de literatura, en la que se pretende realizar un análisis integrador de fuentes científicas y normativas que permita identificar consensos, tensiones, vacíos y criterios de acción. La recuperación bibliográfica privilegió artículos indexados o publicados por editoriales académicas reconocidas en las bases de datos Google Academic, SciELO, Redalyc y Dialnet, además de documentos normativos de organismos internacionales y fuentes oficiales.

La estrategia de búsqueda se organizó en los siguientes grupos conceptuales; inteligencia artificial educativa, sesgo algorítmico en educación, justicia o equidad algorítmica, indicadores de aprendizaje, predicción de riesgo académico, evaluación automatizada, supervisión remota, gobernanza digital, derechos humanos, auditoría y documentación de sistemas de IA. Se priorizaron fuentes de alto impacto por citación y centralidad conceptual, así como trabajos recientes que conectan explícitamente justicia algorítmica con sistemas educativos con IA.

Los criterios de inclusión fueron: a) pertinencia directa para educación, algoritmos, equidad, derechos o gobernanza; b) disponibilidad con acceso abierto, página editorial, repositorio institucional u organismo oficial; c) contribución conceptual, empírica o normativa al estado del arte; y d) preferencia por artículos revisados por pares, libros académicos o documentos de política pública con autoridad institucional. Se excluyeron textos sin datos bibliográficos claros, fuentes duplicadas y materiales que sólo repetían hallazgos ya consolidados. Al final del proceso de búsqueda se identificaron 26 fuentes pertinentes para este trabajo.

La revisión reconoce como limitación relevante, que la literatura disponible está geográficamente concentrada, con predominio de estudios de Estados Unidos y Europa,

³KIZILCEC, RENÉ F. Y LEE, HANSOL, «Algorithmic Fairness in Education», en Holmes, Wayne y Porayska-Pomsta, Kaska (eds.), *The Ethics of Artificial Intelligence in Education*, Londres, Routledge, 2022, pp. 174-202.

mientras que América Latina e Iberoamérica aparecen subrepresentadas. Esta diferencia reconoce la existencia de sistemas educativos con desigualdades, categorías de análisis, infraestructuras digitales y marcos de protección distintos; lo que enfatiza la necesidad de investigación local contextualizada.⁴

Para ordenar la revisión se construyó una tipología de análisis con cuatro niveles; el primer nivel corresponde a sesgos de entrada, tales como, datos incompletos, muestras no representativas y registros producidos por plataformas que capturan actividad digital pero no necesariamente aprendizaje. El segundo nivel corresponde a sesgos de modelado; en particular, selección de objetivos, optimización por precisión promedio, etiquetas históricas y métricas de justicia no fundamentadas pedagógicamente. El tercer nivel corresponde a sesgos de implementación; por ejemplo, interpretación acrítica, automatización de decisiones, opacidad contractual y ausencia de canales de reclamación. El cuarto nivel corresponde a daños educativos; entre los que se encuentra, exclusión, estigmatización, vigilancia desproporcionada, pérdida de autonomía, disminución de oportunidades y reproducción de desigualdades.

Esta tipología permite una lectura reflexiva que busca responder la pregunta de cómo el sesgo se formula en términos de trayectoria; es decir, cómo un dato se convierte en predicción, cómo una predicción se convierte en intervención y cómo esa intervención altera la experiencia educativa de una persona o grupo. En consecuencia, un sistema puede ser problemático, aunque no discrimine explícitamente en su código; basta con que sus insumos de información, objetivos o usos institucionales produzcan efectos diferenciados injustificados.

3. DEL SESGO TÉCNICO AL PERJUICIO EDUCATIVO

Luego de la lectura de los materiales de información identificados y recolectados, un primer consenso es que el sesgo algorítmico no se origina en un único punto del sistema; este puede introducirse en la definición del problema, en la recolección de datos, en la medición de variables, en la elección de etiquetas, en el aprendizaje del modelo, en la evaluación, en el despliegue y en el modo en que las personas interpretan y actúan frente a los resultados. En el ámbito educativo, estos puntos son especialmente sensibles porque muchas variables usadas como indicadores de aprendizaje son en realidad un reflejo de condiciones sociales específicas; por ejemplo, conectividad, disponibilidad de tiempo, familiaridad con el lenguaje académico, capital cultural, escuela de procedencia o trayectorias previas.⁵

La literatura crítica también advierte que el término sesgo suele usarse de manera ambigua, ya que en algunos textos designa error estadístico; en otros, disparidad de desempeño entre grupos; en otros más, daño distributivo, estereotipación, exclusión o reproducción de jerarquías. Esta ambigüedad no es menor, si una universidad afirma que su sistema no tiene sesgos porque mantiene una precisión promedio, puede ignorar

⁴OECD, «Algorithmic Bias: The State of the Situation and Policy Recommendations», en OECD Digital Education Outlook 2023, París, OECD Publishing, 2023.

⁵SURESH, HARINI y GUTTAG, JOHN, «A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle», Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2021

que comete más errores con estudiantes indígenas, hablantes no nativos, personas con discapacidad o estudiantes de bajo nivel socioeconómico; de esta manera, la definición de sesgo es, por tanto, una decisión normativa.⁶

La teoría formal de la equidad algorítmica aporta herramientas importantes, como paridad demográfica, igualdad de oportunidades, calibración, equidad individual o restricciones de error por grupo. Sin embargo, la misma literatura demuestra que estas métricas pueden entrar en conflicto, tratar de igual manera a estudiantes que en apariencia son similares exige una métrica de similitud que no es evidente en educación; garantizar tasas iguales de intervención puede reducir precisión; y calibrar un modelo puede no eliminar disparidades cuando los grupos parten de condiciones estructurales distintas; por ello, la equidad educativa no puede resolverse eligiendo una métrica aislada.^{7,8}

En cuanto a indicadores de aprendizaje, la preocupación central se ubica en los sistemas de predicción de riesgo académico. Estos modelos prometen identificar tempranamente a estudiantes que podrían reprobado, abandonar o requerir apoyo. Su atractivo institucional es evidente, porque permite anticipar problemas, asignar recursos y mejorar retención; pero la literatura muestra que los algoritmos pueden funcionar de forma desigual entre grupos y que la optimización por precisión puede invisibilizar errores diferenciados. Un falso positivo puede estigmatizar; un falso negativo puede privar de apoyo; y una intervención basada en riesgo puede convertirse en una meta institucional errónea si docentes o tutores reducen expectativas.^{9,10}

Los estudios sobre predicción de éxito estudiantil muestran que los ajustes posteriores para mejorar una dimensión de equidad pueden empeorar otra, esto obliga a abandonar la idea de que existe una solución universal de justicia educativa. En educación, la selección de la métrica debe justificarse con base en el daño que se busca evitar; por ejemplo, exclusión del acceso, pérdida de becas, vigilancia injustificada, ausencia de apoyo, clasificación estigmatizante, invisibilidad de necesidades o disminución de autonomía. La pregunta correcta no es solo si el modelo es justo en abstracto, sino justo respecto de qué derecho, qué grupo, qué decisión y qué consecuencia.¹¹

Los tutores inteligentes y plataformas adaptativas plantean otro tipo de riesgo, dado que, al personalizar contenidos, ritmos o retroalimentación, pueden asignar experiencias educativas distintas según patrones históricos de desempeño. Si los datos reflejan

⁶BLODGETT, SU LIN; BAROCAS, SOLON; DAUMÉ III, HAL y WALLACH, HANNA, «Language (Technology) is Power: A Critical Survey of 'Bias' in NLP», Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5454-5476.

⁷DWORK, CYNTHIA; HARDT, MORITZ; PITASSI, TONIANN; REINGOLD, OMER y ZEMEL, RICHARD, «Fairness through Awareness», Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012, pp. 214-226.

⁸KLEINBERG, JON; MULLAINATHAN, SENDHIL y RAGHAVAN, MANISH, «Inherent Trade-Offs in the Fair Determination of Risk Scores», 8th Innovations in Theoretical Computer Science Conference, LIPIcs, vol. 67, 2017, art. 43.

⁹BAYER, VACLAV; HLOSTA, MARTIN y FERNANDEZ, MIRIAM, «Learning Analytics and Fairness: Do Existing Algorithms Serve Everyone Equally?», en Artificial Intelligence in Education. AIED 2021, Lecture Notes in Computer Science, vol. 12749, Cham, Springer, 2021, pp. 71-75.

¹⁰RIAZY, SHIRIN; SIMBECK, KATHARINA y SCHRECK, VANESSA, «Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments», Proceedings of the 12th International Conference on Computer Supported Education, vol. 1, 2020, pp. 15-25.

¹¹LEE, HANSOL y KIZILCEC, RENÉ F., «Evaluation of Fairness Trade-Offs in Predicting Student Success», arXiv, 2020.

desigualdades previas, el sistema puede ofrecer tareas menos desafiantes a estudiantes que requieren más oportunidades, no menos. Es por esto que, la personalización puede convertirse en segmentación silenciosa, tal es el caso de estudiantes considerados de bajo rendimiento que reciben trayectorias empobrecidas, mientras quienes ya cuentan con ventajas reciben retos de mayor complejidad; lo que implica que la justicia pedagógica debe exigir una revisión de la exactitud del diagnóstico, aunado a la calidad de la experiencia que se deriva de él.

Por otra parte, el monitoreo remoto de exámenes muestra con claridad la convergencia entre sesgo, vigilancia y derechos, a causa de las capacidades de las herramientas de vigilancia que pueden detectar rostros, movimientos, sonidos, miradas o patrones de navegación para inferir conductas sospechosas. Estos sistemas han sido cuestionados por riesgos de privacidad, autonomía, libertad, transparencia, confianza y no discriminación. En términos pedagógicos, introducen un clima de sospecha que puede afectar de manera diferenciada a estudiantes con discapacidad, estudiantes neurodivergentes, personas con condiciones domésticas precarias o usuarios de equipos y conexiones inestables.¹²

De forma similar, la evaluación automatizada, en particular la calificación de ensayos, respuestas abiertas o producción oral concentra riesgos adicionales porque traduce expresiones complejas de aprendizaje en puntuaciones. Si un sistema penaliza variedades lingüísticas, estilos argumentativos, acentos, estructuras discursivas o formas culturalmente situadas de escritura, el daño no es solo estadístico, también afecta el reconocimiento académico de la voz estudiantil. La literatura psicométrica reciente propone marcos para evaluar la función algorítmica diferencial, análoga a preocupaciones existentes sobre sesgo en pruebas, pero adaptada a decisiones automatizadas.¹³

Ahora bien, un hallazgo transversal que se pudo identificar en esta revisión es la falta de evidencia robusta sobre grupos históricamente poco estudiados, en donde el género, la raza y la etnicidad concentran buena parte de la literatura, especialmente en Estados Unidos. En cambio, discapacidad, condición migrante, lengua indígena, ruralidad, conectividad, primera generación universitaria, maternidad o paternidad estudiantil, trabajo remunerado, edad, orientación sexual o interseccionalidad aparecen de manera dispersa. En América Latina, el vacío es más grave, los sistemas educativos tienen desigualdades profundas, pero existen pocos estudios que midan cómo los algoritmos educativos funcionan en esos contextos.

Es así como, los escenarios donde se pueden manifestar sesgos descritos anteriormente permiten plantear dos tipos de brechas, la primera es epistémica, derivada de la manera en que los sistemas se entrenan, evalúan o justifican con evidencia producida en otros contextos, se corre el riesgo de invisibilizar desigualdades locales. La segunda es de naturaleza regulatoria, sin evidencia situada, las políticas públicas pueden adoptar fórmulas generales de transparencia o protección de datos sin atender daños concretos de lengua, territorio, modalidad educativa o infraestructura, tal es el caso de regiones

¹²COGHLAN, SIMON; MILLER, TIM y PATERSON, JEANNIE, «Good Proctor or 'Big Brother'? Ethics of Online Exam Supervision Technologies», *Philosophy & Technology*, vol. 34, 2021, pp. 1581-1606.

¹³SUK, YOUNG y HAN, KYUNG T., «A Psychometric Framework for Evaluating Fairness in Algorithmic Decision Making: Differential Algorithmic Functioning», *Applied Measurement in Education*, vol. 49, núm. 2, 2024.

con alta desigualdad digital, donde un modelo que usa frecuencia de conexión como indicador de compromiso puede confundir precariedad material con desinterés académico.

También existe un vacío sobre impactos acumulativos, muchos estudios evalúan un modelo en un momento específico, pero las trayectorias educativas se componen de decisiones sucesivas; tales como, admisión, diagnóstico, nivelación, asignación de grupos, becas, evaluación, tutoría, alertas, permanencia y egreso. Pequeñas disparidades en cada etapa pueden acumularse hasta producir exclusiones significativas, por ello, la investigación futura debería analizar procesos de decisión y no solo herramientas aisladas. En síntesis, la IA educativa puede ayudar a detectar barreras, ofrecer retroalimentación o ampliar apoyos, pero solo bajo condiciones estrictas de diseño y gobernanza. El riesgo no es la automatización por sí misma, sino su uso acrítico, opaco y descontextualizado en instituciones que ya distribuyen oportunidades de manera desigual.

4. IMPLICACIONES ÉTICAS Y JURÍDICAS

Desde una perspectiva ética, el sesgo algorítmico en el ámbito educativo compromete dignidad, autonomía, justicia, transparencia, beneficencia, no maleficencia y responsabilidad institucional. Desde una perspectiva jurídica, se vincula con igualdad y no discriminación, derecho a la educación, protección de datos personales, debido proceso, transparencia, acceso a información significativa y disponibilidad de medios. La UNESCO y el Consejo de Europa colocan estos principios en el centro de la gobernanza de IA, porque los sistemas deben respetar derechos humanos, igualdad, privacidad, supervisión y rendición de cuentas durante todo su ciclo de vida.^{14,15}

El reglamento europeo de inteligencia artificial ofrece un referente comparado relevante porque clasifica como de alto riesgo los sistemas de IA usados en educación cuando pueden determinar acceso, evaluación o trayectoria profesional, e impone obligaciones sobre gestión de riesgos, calidad de datos, documentación, información al usuario, supervisión humana, robustez y trazabilidad. Además, prohíbe ciertas prácticas como el reconocimiento de emociones en instituciones educativas. Aunque este marco no es directamente aplicable en América Latina, su propuesta para analizar el riesgo permite identificar qué decisiones educativas requieren mayor control jurídico.¹⁶

En el contexto mexicano, el análisis puede anclarse en el derecho a la educación, la igualdad sustantiva, la no discriminación y la protección de datos personales. La educación debe orientarse al respeto de la dignidad, los derechos humanos y la igualdad; por ello, una plataforma que distribuya apoyos o evalúe estudiantes mediante criterios opacos no puede tratarse únicamente como herramienta administrativa. Si procesa datos personales, su operación debe observar finalidad, proporcionalidad, información, seguridad, consentimiento cuando proceda, responsabilidad y derechos de acceso,

¹⁴UNESCO, Recommendation on the Ethics of Artificial Intelligence, París, UNESCO, 2022.

¹⁵COUNCIL OF EUROPE, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024.

¹⁶EUROPEAN COMMISSION, «AI Act», Shaping Europe's Digital Future, 2026. Disponible en: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>; Reglamento (UE) 2024/1689

rectificación, cancelación u oposición. Si actúa en instituciones públicas, el estándar de legalidad y rendición de cuentas se vuelve más importante.¹⁷

Una primera implicación jurídica es la necesidad de pasar de la transparencia formal a la explicabilidad significativa, entendida como el requerimiento por parte de las personas afectadas para conocer, en términos comprensibles, qué tipo de datos se usan, con qué finalidad, qué decisión se apoya o automatiza, qué margen de intervención humana existe, cómo puede controvertirse el resultado y qué consecuencias puede producir. La opacidad es especialmente problemática cuando el algoritmo influye en admisión, becas, permanencia, evaluación, disciplina o asignación de apoyos.¹⁸

Una segunda implicación es la rendición de cuentas, bajo el entendido de que las guías éticas de IA suelen coincidir en transparencia, justicia, no maleficencia, responsabilidad y privacidad, pero la literatura advierte que esos principios pueden volverse retórica si no se traducen en obligaciones, responsables, evidencia y remedios. En el contexto educativo, la responsabilidad no puede desplazarse al proveedor tecnológico ni diluirse en la complejidad del modelo, a razón de que la institución que adopta el sistema conserva deberes de selección diligente, evaluación de impacto, monitoreo, información a la comunidad y suspensión cuando el riesgo sea injustificado.¹⁹

De esta manera, la auditoría algorítmica interna y la documentación de modelos son instrumentos necesarios para cerrar la brecha de responsabilidad. Ante este escenario, registros de entrenamiento, reportes de desempeño por subgrupo, bitácoras de incidentes y evaluaciones de impacto permiten superar declaraciones genéricas de confianza y producir evidencia verificable. Estos documentos deben incluir finalidades pedagógicas, población prevista, variables sensibles, limitaciones, condiciones de uso, pruebas de equidad, protocolos de apelación y criterios para retiro del sistema.^{20,21}

Una tercera implicación se ubica en la relación entre privacidad y equidad, ya que para medir sesgos suele ser necesario conocer variables sensibles o de contexto; pero recopilar esos datos puede aumentar riesgos de vigilancia, perfilamiento o uso secundario. Por esto, la respuesta no debe ser ignorar las diferencias, porque eso impide detectar discriminación, ni recolectar todo sin límites, porque eso vulnera privacidad. La salida jurídicamente razonable, de acuerdo con especialistas, consiste en la gobernanza de datos; mediante la finalidad explícita, minimización, separación de bases, controles

¹⁷CÁMARA DE DIPUTADOS DEL H. CONGRESO DE LA UNIÓN, Constitución Política de los Estados Unidos Mexicanos, texto vigente; Ley General de Educación; Ley Federal de Protección de Datos Personales en Posesión de los Particulares; Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados.

¹⁸MITTELSTADT, BRENT DANIEL; ALLO, PATRICK; TADDEO, MARIAROSARIA; WACHTER, SANDRA y FLORIDI, LUCIANO, «The Ethics of Algorithms: Mapping the Debate», *Big Data & Society*, vol. 3, núm. 2, 2016.

¹⁹JOBIN, ANNA; IENCA, MARCELLO y VAYENA, EFFY, «The Global Landscape of AI Ethics Guidelines», *Nature Machine Intelligence*, vol. 1, 2019, pp. 389-399.

²⁰RAJI, INIOLUWA DEBORAH; SMART, ANDREW; WHITE, REBECCA N.; MITCHELL, MARGARET; GEBRU, TIMNIT; HUTCHINSON, BEN; SMITH-LOUD, JAMILA; THERON, DANIEL y BARNES, PARKER, «Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing», *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

²¹MITCHELL, MARGARET; WU, SIMONE; ZALDIVAR, ANDREW; BARNES, PARKER; VASSERMAN, LUCY; HUTCHINSON, BEN; SPITZER, ELENA; RAJI, INIOLUWA DEBORAH y GEBRU, TIMNIT, «Model Cards for Model Reporting», *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220-229.

de acceso, anonimización o seudonimización cuando sea posible, evaluación de impacto y participación de grupos afectados.

Una cuarta implicación es el derecho a impugnar decisiones, dado que los sistemas educativos inteligentes no deberían producir decisiones finales e incontestables, especialmente cuando afectan trayectoria académica. La revisión humana debe ser significativa, es decir, no un trámite en que la persona revisora confirma automáticamente el resultado, sino una evaluación con autoridad real para corregir, contextualizar, suspender o descartar la recomendación algorítmica. Esto exige capacitación docente y administrativa, así como reglas claras sobre evidencia, carga argumentativa y tiempos de respuesta.

5. IMPLICACIONES PEDAGÓGICAS

La discusión pedagógica en cuanto a los sesgos que pueden cometer los algoritmos no debe limitarse a preguntar si la IA únicamente mejora la eficiencia del proceso de aprendizaje. La cuestión de fondo es qué concepción de conocimiento, estudiante, docente y evaluación se manifiesta en el sistema. Un modelo que reduce aprendizaje a número de clics, permanencia en pantalla o patrones de respuesta puede ser útil para ciertas tareas, pero pobre para valorar comprensión, creatividad, colaboración, ética o pensamiento crítico. Por esto, la pedagogía aporta criterios para decidir qué puede automatizarse, qué debe permanecer bajo juicio docente y qué nunca debe delegarse sin garantías reforzadas.²²

La analítica del aprendizaje suele justificarse como forma de prevención institucional, bajo el entendido de que conocer mejor al estudiante permitiría intervenir antes. Sin embargo, la misma práctica puede convertirse en vigilancia si no existe transparencia, consentimiento informado, proporcionalidad y participación. La frontera entre cuidado y control depende de quién define el problema, qué datos se observan, qué inferencias se hacen, cómo se comunican los resultados y qué oportunidades reales tiene el estudiante de comprender o cambiar su perfil.²³

La ética en torno a los indicadores de aprendizaje propone principios de transparencia, control, seguridad, consentimiento, acceso y mejora educativa. Desde una mirada pedagógica, estas bases deben traducirse en prácticas concretas; por ejemplo, explicar a estudiantes cómo se usa su información, permitirles ver y corregir datos relevantes, evitar inferencias que excedan la finalidad educativa, diferenciar recomendaciones de decisiones obligatorias y asegurar que las intervenciones algorítmicas amplíen oportunidades en lugar de restringir trayectorias.²⁴

Al mismo tiempo, las percepciones de los estudiantes deben tomarse en cuenta, más aún, porque la aceptación de este tipo de sistemas no puede suponerse por el hecho de que estén orientados a mejorar resultados. Los estudiantes pueden valorar

²²AKGUN, SELIN y GREENHOW, CHRISTINE, «Artificial Intelligence in Education: Addressing Ethical Challenges in K-12 Settings», *AI and Ethics*, vol. 2, 2022, pp. 431-440.

²³SLADE, SHARON y PRINSLOO, PAUL, «Learning Analytics: Ethical Issues and Dilemmas», *American Behavioral Scientist*, vol. 57, núm. 10, 2013, pp. 1510-1529.

²⁴PARDO, ABELARDO y SIEMENS, GEORGE, «Ethical and Privacy Principles for Learning Analytics», *British Journal of Educational Technology*, vol. 45, núm. 3, 2014, pp. 438-450.

ciertos usos de datos y rechazar otros; pueden aceptar alertas académicas y oponerse al monitoreo continuo; pueden considerar legítimo recibir retroalimentación y, a la vez, demandar límites frente a inferencias sobre motivación, salud mental, honestidad o riesgo de abandono; de manera ideal, una gobernanza pedagógica debe incorporar estas diferencias y no tratar al estudiante como simple fuente de datos.²⁵

Por otra parte, el papel docente también cambia, a partir de una reflexión, una plataforma puede producir recomendaciones, pero es el docente quien puede asumirlas como apoyo interpretativo o como autoridad superior. En el primer caso, el algoritmo enriquece la deliberación profesional; en el segundo, desplaza el juicio pedagógico. Por ello, la formación docente en IA debe incluir la lectura crítica de métricas, comprensión de errores, análisis de sesgos, diseño de intervenciones equitativas y criterios para explicar a estudiantes las razones de una decisión apoyada por tecnología.

Desde esta mirada, la alfabetización algorítmica se perfila como dimensión emergente del derecho educativo en la que no basta enseñar a usar herramientas; es necesario comprender cómo clasifican, qué datos requieren, qué sesgos pueden reproducir, qué límites tienen y cómo cuestionarlos. En una sociedad donde algoritmos influyen en admisión, crédito, empleo, salud, seguridad y educación, formar ciudadanía democrática exige que estudiantes puedan interpretar, auditar y disputar sistemas automatizados.

Sin embargo, los avances no se detienen, la irrupción de IA generativa intensifica estos dilemas, porque a diferencia de sistemas predictivos tradicionales, los modelos generativos producen texto, retroalimentación, explicaciones y materiales que pueden influir directamente en el aprendizaje. Sus sesgos pueden aparecer como estereotipos, omisiones, alucinaciones, traducciones deficientes, recomendaciones culturalmente inadecuadas o estilos de corrección que favorecen ciertos registros lingüísticos. Ante esto, la guía de la UNESCO insiste en una aproximación centrada en la persona, con regulación, privacidad, edad apropiada, validación ética y diseño pedagógico significativo.²⁶

En términos generales, las implicaciones pedagógicas pueden sintetizarse en cuatro exigencias. Primera, no usar IA para decisiones educativas de alto impacto sin evidencia suficiente de validez, equidad y pertinencia contextual. Segunda, preservar espacios de interacción humana, especialmente en evaluación, orientación y apoyo socioemocional. Tercera, convertir la transparencia en práctica didáctica, no solo en aviso legal. Cuarta, diseñar intervenciones que reparen desigualdades, no que administren de manera más eficiente su reproducción.

6. HACIA UNA GOBERNANZA EDUCATIVA DE LA JUSTICIA ALGORÍTMICA

La gobernanza educativa basada en la justicia algorítmica debe partir de la siguiente premisa formal, los sistemas no se vuelven justos por cumplir una métrica aislada, porque las decisiones técnicas operan dentro de contextos institucionales, culturales y jurídicos. La literatura sobre sistemas sociotécnicos advierte contra las trampas de abstraer demasiado el problema, ignorar el contexto, formalizar de manera estrecha

²⁵IFENTHALER, DIRK y SCHUMACHER, CLARA, «Student Perceptions of Privacy Principles for Learning Analytics», *Educational Technology Research and Development*, vol. 64, 2016, pp. 923-938.

²⁶MIAO, Fengchun y Holmes, Wayne, *Guidance for Generative AI in Education and Research*, París, UNESCO, 2023.

la equidad o suponer que el diseño técnico puede reemplazar procesos deliberativos. En las instituciones educativas, estas trampas son especialmente visibles cuando se implementa una herramienta sin revisar currículo, población, infraestructura y cultura institucional.²⁷

De tal manera que se plantean los siguientes criterios como base para una gobernanza en este contexto, el primero es la evaluación de impacto algorítmico antes de la adopción. Toda institución que pretenda usar IA para admisión, evaluación, permanencia, becas, disciplina, vigilancia, orientación o asignación de apoyos debería documentar finalidad, base jurídica, necesidad, proporcionalidad, alternativas menos invasivas, datos usados, riesgos para grupos vulnerables, pruebas de desempeño, mecanismos de supervisión y procedimientos de reclamación. Esta evaluación debe actualizarse durante el ciclo de vida del sistema, porque los modelos pueden degradarse o producir efectos no previstos.

El segundo criterio es la documentación verificable, esto comienza desde el proveedor, que debe entregar información técnica y funcional suficiente para que la institución comprenda límites, población objetivo, variables, entrenamiento, supuestos, errores conocidos y condiciones de uso. Si el secreto comercial impide auditar sesgos, la herramienta no debería usarse en decisiones de alto impacto. La contratación pública o institucional de tecnología educativa debe incluir cláusulas de acceso a evidencia, auditoría independiente, seguridad de datos, eliminación al terminar el servicio y prohibición de usos secundarios no autorizados.

El tercer criterio es la evaluación por subgrupos y por daños en la que no basta reportar precisión global; deben medirse falsos positivos, falsos negativos, sensibilidad, especificidad, calibración y consecuencias por grupos relevantes para el contexto. En México e Iberoamérica, esos grupos no pueden definirse copiando categorías de países desarrollados, estos deben dialogar con desigualdades locales, como lengua indígena, ruralidad, discapacidad, brecha digital, tipo de institución, modalidad educativa, trabajo remunerado, género, condición migrante y nivel socioeconómico.

El cuarto criterio es la participación de la comunidad, en donde estudiantes, docentes, familias, personal de orientación, áreas jurídicas, especialistas en datos y representantes de grupos potencialmente afectados deben intervenir en la definición de finalidades, riesgos aceptables, indicadores de equidad y procedimientos de revisión. La participación no es un mero trámite; mejora la calidad epistémica del sistema porque permite identificar daños que no aparecen en métricas técnicas.

El quinto criterio es el control humano significativo, comenzando por entender que no toda intervención humana es relevante; la supervisión exige capacitación, acceso a explicaciones, tiempo institucional, autoridad para corregir resultados y responsabilidad por decisiones finales. También requiere evitar automatización de facto, la cual puede darse cuando la carga de apartarse del proceso automatizado es demasiado alta y el resultado algorítmico termina gobernando aunque formalmente exista revisión humana.

²⁷SELBST, ANDREW D.; BOYD, DANAH; FRIEDLER, SORELLE A.; VENKATASUBRAMANIAN, SURESH y VERTESI, JANET, «Fairness and Abstraction in Sociotechnical Systems», FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 2019, pp. 59-68.

El sexto criterio es la proporcionalidad pedagógica, es aquí donde vale la pena preguntarse si toda ventaja tecnológica justifica el tratamiento de datos o el riesgo de sesgo. Una plataforma de recomendación de lecturas no demanda el mismo escrutinio que un sistema de asignación de becas; un chatbot de apoyo general no plantea el mismo riesgo que un evaluador automático de exámenes finales. A partir de este principio, la intensidad de la regulación institucional debe aumentar con el impacto sobre derechos, trayectoria, evaluación y dignidad del estudiante.

Por último, el séptimo criterio es la reparación, si un sistema ha perjudicado a un grupo, la institución debe contar con medidas de corrección, entre las que se pueden encontrar, la revisión individual, rectificación de expedientes, reentrenamiento o retiro del modelo, comunicación a personas afectadas, ajuste de políticas, medidas compensatorias cuando procedan y aprendizaje institucional. Sin remedios, la auditoría se convierte en ejercicio descriptivo sin impacto real.

Estos criterios deben integrarse en políticas institucionales, no quedar como recomendaciones voluntarias de un comité técnico. Una universidad o sistema escolar que adopta IA debería contar con un inventario público de sistemas automatizados, responsables internos, clasificación de riesgo, periodicidad de auditoría, reglas de contratación, protocolos de incidentes, canales de queja y mecanismos de participación. La gobernanza algorítmica educativa requiere continuidad administrativa, no puede depender de la buena voluntad de un equipo o de la promesa comercial de un proveedor. En instituciones con capacidades técnicas limitadas, la cooperación interinstitucional puede ser decisiva, esto reduciría costos y evitaría que cada institución enfrente sola problemas que son estructurales.

7. CONCLUSIONES

El análisis de literatura permitió sostener que los sesgos algorítmicos en sistemas educativos que usan inteligencia artificial constituyen potenciales problemas de derechos y justicia en el contexto educativo. estas herramientas pueden apoyar la enseñanza, diagnóstico y gestión, pero también pueden reproducir desigualdades cuando aprenden de datos históricos, operan sin transparencia, carecen de evaluación por subgrupos o desplazan el juicio pedagógico hacia modelos opacos de trabajo.

La primera conclusión es de naturaleza conceptual, el sesgo algorítmico debe comprenderse como fenómeno social, el cual no nace exclusivamente en el código, sino en decisiones institucionales sobre qué medir, para qué predecir, qué consecuencias asociar a la predicción y quién puede cuestionarla. Por eso, la justicia algorítmica educativa requiere combinar métricas técnicas con principios jurídicos, deliberación pedagógica y participación de comunidades afectadas.

La segunda conclusión es jurídica, los sistemas que influyen en acceso, evaluación, permanencia, apoyos o vigilancia deben tratarse como tecnologías de alto impacto. Su adopción exige evaluación de necesidad y proporcionalidad, protección robusta de datos personales, información significativa, documentación, auditoría, revisión humana y remedios, en donde la opacidad contractual o técnica no puede justificar decisiones que afecten derechos educativos.

La tercera conclusión es pedagógica, porque la IA educativa debe subordinarse a fines formativos, si una herramienta reduce aprendizaje a datos conductuales, personaliza hacia trayectorias empobrecidas o convierte la intervención en vigilancia, su eficiencia puede ser pedagógicamente regresiva. El criterio de calidad no debe ser solo predicción correcta, sino ampliación efectiva de oportunidades de aprendizaje, reconocimiento de diversidad y fortalecimiento de autonomía estudiantil.

Finalmente, se afirma que existe un vacío importante de evidencias contextualizadas en este tema, debido a que la mayor parte de la literatura sobre sesgo algorítmico educativo proviene de países con categorías, datos e instituciones distintas. Esto obliga a impulsar estudios locales, especialmente sobre discapacidad, lengua, ruralidad, nivel socioeconómico, brecha digital e interseccionalidad. La ausencia de evidencia no debe interpretarse como ausencia de sesgo; al contrario, indica una zona de riesgo no observada.

El artículo propone, en suma, una base de criterios de gobernanza educativa para la justicia algorítmica que incluyen evaluación de impacto, documentación verificable, auditoría diferenciada por grupos, participación, control humano significativo, proporcionalidad pedagógica, protección de datos y reparación. Estos principios permiten alinear la innovación educativa con derechos humanos, no discriminación y Estado de derecho.

8. BIBLIOGRAFÍA

- AKGUN, SELIN y GREENHOW, CHRISTINE, «Artificial Intelligence in Education: Addressing Ethical Challenges in K-12 Settings», *AI and Ethics*, vol. 2, 2022, pp. 431-440. En línea. [Disponible en: <https://doi.org/10.1007/s43681-021-00096-7>].
- BAKER, RYAN S. Y HAWN, AARON, «Algorithmic Bias in Education», *International Journal of Artificial Intelligence in Education*, vol. 32, 2022, pp. 1052-1092. En línea. [Disponible en: <https://doi.org/10.1007/s40593-021-00285-9>].
- BAYER, VACLAV; HLOSTA, MARTIN y FERNANDEZ, MIRIAM, «Learning Analytics and Fairness: Do Existing Algorithms Serve Everyone Equally?», en *Artificial Intelligence in Education. AIED 2021, Lecture Notes in Computer Science*, vol. 12749, Cham, Springer, 2021, pp. 71-75. En línea. [Disponible en: https://doi.org/10.1007/978-3-030-78270-2_12].
- BLODGETT, SU LIN; BAROCAS, SOLON; DAUMÉ III, HAL y WALLACH, HANNA, «Language (Technology) is Power: A Critical Survey of 'Bias' in NLP», *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5454-5476. En línea. [Disponible en: <https://doi.org/10.18653/v1/2020.acl-main.485>].
- CÁMARA DE DIPUTADOS DEL H. CONGRESO DE LA UNIÓN, *Leyes federales vigentes*, México. En línea. [Disponible en: <https://www.diputados.gob.mx/LeyesBiblio/index.htm>].
- COGHLAN, SIMON; MILLER, TIM y PATERSON, JEANNIE, «Good Proctor or 'Big Brother'? Ethics of Online Exam Supervision Technologies», *Philosophy & Technology*, vol. 34, 2021, pp. 1581-1606. En línea. [Disponible en: <https://doi.org/10.1007/s13347-021-00476-1>].
- COUNCIL OF EUROPE, *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, Estrasburgo, Council of Europe, 2024. En línea. [Disponible en: <https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>].
- DWORK, CYNTHIA; HARDT, MORITZ; PITASSI, TONIANN; REINGOLD, OMER y ZEMEL, RICHARD, «Fairness through Awareness», *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 214-226. En línea. [Disponible en: <https://doi.org/10.1145/2090236.2090255>].

- EUROPEAN COMMISSION, «AI Act», Shaping Europe's Digital Future, Bruselas, 2026. En línea. [Disponible en: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>].
- IFENTHALER, DIRK y SCHUMACHER, CLARA, «Student Perceptions of Privacy Principles for Learning Analytics», *Educational Technology Research and Development*, vol. 64, 2016, pp. 923-938. En línea. [Disponible en: <https://doi.org/10.1007/s11423-016-9477-y>].
- JOBIN, ANNA; IENCA, MARCELLO y VAYENA, EFFY, «The Global Landscape of AI Ethics Guidelines», *Nature Machine Intelligence*, vol. 1, 2019, pp. 389-399. En línea. [Disponible en: <https://doi.org/10.1038/s42256-019-0088-2>].
- KIZILCEC, RENÉ F. Y LEE, HANSOL, «Algorithmic Fairness in Education», en Holmes, Wayne y Porayska-Pomsta, Kaska (eds.), *The Ethics of Artificial Intelligence in Education*, Londres, Routledge, 2022, pp. 174-202. En línea. [Disponible en: <https://doi.org/10.4324/9780429329067-10>].
- KLEINBERG, JON; MULLAINATHAN, SENDHIL y RAGHAVAN, MANISH, «Inherent Trade-Offs in the Fair Determination of Risk Scores», 8th Innovations in Theoretical Computer Science Conference, LIPIcs, vol. 67, 2017, art. 43. En línea. [Disponible en: <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>].
- LEE, HANSOL y KIZILCEC, RENÉ F., «Evaluation of Fairness Trade-Offs in Predicting Student Success», arXiv, 2020. En línea. [Disponible en: <https://doi.org/10.48550/arXiv.2007.00088>].
- MIAO, FENGCHUN y HOLMES, WAYNE, *Guidance for Generative AI in Education and Research*, París, UNESCO, 2023. En línea. [Disponible en: <https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research>].
- MITCHELL, MARGARET; WU, SIMONE; ZALDIVAR, ANDREW; BARNES, PARKER; VASSERMAN, LUCY; HUTCHINSON, BEN; SPITZER, ELENA; RAJI, INIOLUWA DEBORAH y GEBRU, TIMNIT, «Model Cards for Model Reporting», *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220-229. En línea. [Disponible en: <https://doi.org/10.1145/3287560.3287596>].
- MITTELSTADT, BRENT DANIEL; ALLO, PATRICK; TADDEO, MARIAROSARIA; WACHTER, SANDRA y FLORIDI, LUCIANO, «The Ethics of Algorithms: Mapping the Debate», *Big Data & Society*, vol. 3, núm. 2, 2016. En línea. [Disponible en: <https://doi.org/10.1177/2053951716679679>].
- OECD, «Algorithmic Bias: The State of the Situation and Policy Recommendations», en *OECD Digital Education Outlook 2023*, París, OECD Publishing, 2023. En línea. [Disponible en: https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023_c74f03de-en/full-report/algorithmic-bias-the-state-of-the-situation-and-policy-recommendations_a0b7cec1.html].
- PARDO, ABELARDO y SIEMENS, GEORGE, «Ethical and Privacy Principles for Learning Analytics», *British Journal of Educational Technology*, vol. 45, núm. 3, 2014, pp. 438-450. En línea. [Disponible en: <https://doi.org/10.1111/bjet.12152>].
- RAJI, INIOLUWA DEBORAH; SMART, ANDREW; WHITE, REBECCA N.; MITCHELL, MARGARET; GEBRU, TIMNIT; HUTCHINSON, BEN; SMITH-LOUD, JAMILA; THERON, DANIEL y BARNES, PARKER, «Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing», *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33-44. En línea. [Disponible en: <https://doi.org/10.1145/3351095.3372873>].
- RIAZY, SHIRIN; SIMBECK, KATHARINA y SCHRECK, VANESSA, «Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments», *Proceedings of the 12th International Conference on Computer Supported Education*, vol. 1, 2020, pp. 15-25. En línea. [Disponible en: <https://doi.org/10.5220/0009324100150025>].
- SELBST, ANDREW D.; BOYD, DANAH; FRIEDLER, SORELLE A.; VENKATASUBRAMANIAN, SURESH y VERTESI, JANET, «Fairness and Abstraction in Sociotechnical Systems», *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 2019, pp. 59-68. En línea. [Disponible en: <https://doi.org/10.1145/3287560.3287598>].
- SLADE, SHARON y PRINSLOO, PAUL, «Learning Analytics: Ethical Issues and Dilemmas», *American Behavioral Scientist*, vol. 57, núm. 10, 2013, pp. 1510-1529. En línea. [Disponible en: <https://doi.org/10.1177/0002764213479366>].

- SUK, YOUMI y HAN, KYUNG T., «A Psychometric Framework for Evaluating Fairness in Algorithmic Decision Making: Differential Algorithmic Functioning», *Applied Measurement in Education*, vol. 49, núm. 2, 2024. En línea. [Disponible en: <https://doi.org/10.3102/10769986231171711>].
- SURESH, HARINI y GUTTAG, JOHN, «A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle», *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021. En línea. [Disponible en: <https://doi.org/10.1145/3465416.3483305>].
- UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, París, UNESCO, 2022. En línea. [Disponible en: <https://unesdoc.unesco.org/ark:/48223/pf0000380455>].
- ZAWACKI-RICHTER, OLAF; MARÍN, VICTORIA I.; BOND, MELISSA y GOUVERNEUR, FRANZISKA, «Systematic Review of Research on Artificial Intelligence Applications in Higher Education: Where Are the Educators?», *International Journal of Educational Technology in Higher Education*, vol. 16, art. 39, 2019. En línea. [Disponible en: <https://doi.org/10.1186/s41239-019-0171-0>].