

Estudio crítico sobre la equidad y la inclusión en plataformas de aprendizaje automático: Sesgos algorítmicos como nueva barrera educativa

A Critical Study of Equity and Inclusion in Machine Learning Platforms: Algorithmic Bias as a New Educational Barrier

Vigdalia Coromoto Parra Sequera

Universidad Centroccidental "Lisandro Alvarado". Venezuela

Correo: vigdalia.parra@ucla.edu.veDOI: <https://doi.org/10.5281/zenodo.21446729>

Recibido: 07-04-2026 Aceptado: 31-05-2026

Resumen

La integración de la inteligencia artificial en educación promete personalización y eficiencia, pero introduce riesgos de perpetuar desigualdades. El sesgo algorítmico actúa como una barrera de acceso y permanencia que opera a nivel del código, afectando los derechos educativos de manera transversal a cualquier nivel académico. Esta revisión sistemática examina cómo los sesgos en datos y diseño algorítmico generan discriminación por género, nivel socioeconómico y origen cultural. Siguiendo los lineamientos PRISMA, se analizaron 33 estudios publicados entre 2018 y 2025. Se identificaron tres categorías de sesgo: en datos históricos, de representación y de validación. Las plataformas predictivas (deserción, recomendación, evaluación automatizada) presentan disparidades sistemáticas entre grupos demográficos. Esta revisión tiene limitaciones por la heterogeneidad metodológica de los estudios incluidos. Las estrategias de mitigación incluyen intervenciones en pre-procesamiento, durante el procesamiento y post-procesamiento, con efectividad variable. El marco ético propuesto ofrece una hoja de ruta para diseñar sistemas de IA verdaderamente participativa. Como contribución original, se propone un marco ético integrador para el desarrollo de IA inclusiva en educación, basado en cinco principios: justicia distributiva, transparencia radical, agencia del estudiante, diversidad cognitiva y gobernanza participativa. Se concluye que la equidad algorítmica requiere un abordaje socio-técnico que involucre a educadores, comunidades y diseñadores en la co-construcción de sistemas inclusivos.

Palabras clave: aprendizaje automático; equidad educativa; inteligencia artificial; justicia social; sesgo algorítmico. Código UNESCO: 1203.04 – Inteligencia Artificial y 5801.03 – Teoría de la Educación

Abstract

The integration of artificial intelligence in education promises personalization and efficiency, but introduces risks of perpetuating inequalities. Algorithmic bias acts as a barrier to access and retention that operates at the code level, affecting educational rights across all academic levels. This systematic review examines how biases in data and algorithmic design generate discrimination based on gender, socioeconomic status, and cultural background. Following the PRISMA guidelines, 33 studies published between 2018 and 2025 were analyzed. Three categories of bias were identified: historical data bias, representational bias, and validation bias. Predictive platforms (dropout, recommendation, automated assessment) exhibit systematic disparities among demographic groups. This review has limitations: the methodological heterogeneity of the included studies. Mitigation strategies include pre-processing, in-processing, and post-processing interventions, with varying effectiveness. The proposed ethical framework offers a roadmap for designing truly participatory AI systems. As an original contribution, an integrative ethical framework for the development of inclusive AI in education is proposed, based on five principles: distributive justice, radical transparency, student agency, cognitive diversity, and participatory governance. It is concluded that algorithmic equity requires a socio-technical approach that involves educators, communities, and designers in the co-construction of inclusive systems.

Keywords: machine learning; educational equity; artificial intelligence; social justice; algorithmic bias.

UNESCO Code: 1203.04 – Artificial Intelligence y 5801.03 – Education Theory

PUBLICACIONES EN CIENCIAS Y TECNOLOGÍA, revista científica de publicación continua, dos números al año, editada en la Universidad Centroccidental Lisandro Alvarado (UCLA) en la ciudad de Barquisimeto, Venezuela, bajo la Licencia CC BY-NC-SA. ISSN:1856-8890, EISSN:2477-9660. Depósitos legales: pp200702LA2730, ppi201402LA4590.

*Autor de Correspondencia

***Vigdalia Coromoto Parra Sequera**. ORCID: <https://orcid.org/0000-0002-6930-004X>. Correo: vigdalia.parra@ucla.edu.ve. Ingeniero en Informática., Universidad Centroccidental Lisandro Alvarado (UCLA). Docente - Investigador, Decanato de Ciencias y Tecnología (UCLA), Venezuela.

1. Introducción

La promesa de la inteligencia artificial en educación ha sido formulada en términos de personalización sin precedentes, eficiencia institucional y democratización del acceso al conocimiento [1], [2]. Esto va desde sistemas de tutoría inteligente hasta plataformas de aprendizaje adaptativo, pasando por herramientas de predicción de deserción estudiantil, así, la inteligencia artificial promete adaptar la experiencia educativa a las necesidades individuales de cada aprendiz. Sin embargo, desde esta narrativa de progreso tecnológico surgen preguntas incómodas pero necesarias como: ¿para quién funciona realmente esta personalización? ¿Qué concepciones de inteligencia y aprendizaje quedan privilegiadas? ¿Cómo afectan estos sistemas a estudiantes de grupos históricamente marginados?

La literatura emergente en ética algorítmica ha documentado numerosos casos donde sistemas de inteligencia artificial perpetúan o amplifican sesgos sociales presentes en sus datos de entrenamiento [3], [4], [5]. En el ámbito educativo, estas preocupaciones adquieren una dimensión particularmente sensible, debido a que las decisiones algorítmicas pueden afectar trayectorias educativas completas, desde recomendaciones de cursos hasta predicciones de éxito académico que influyen en las expectativas docentes y oportunidades futuras [6], [7].

En tal sentido el problema del sesgo algorítmico en educación no es meramente técnico, sino profundamente ético y político. Debido a que cuando un sistema de recomendación de contenido sugiere sistemáticamente trayectorias de menor exigencia a estudiantes de ciertos grupos demográficos, o cuando un predictor de deserción etiqueta incorrectamente a estudiantes de minorías como en riesgo basándose en correlaciones espurias, estamos ante manifestaciones de lo que Benjamin [8] denomina “tecnología de la nueva era Jim Crow” que define algoritmos que reproducen y modernizan formas históricas de discriminación.

Por tanto, enfrentamos un grave problema debido a que el sesgo algorítmico en las plataformas de aprendizaje automático socava la equidad educativa en la era digital al perpetuar la discriminación basada en el género, el estatus socioeconómico y el origen cultural, con efectos interseccionales más graves para las mujeres de piel más oscura y los estudiantes pertenecientes a minorías raciales [9], [10], [11]. Así tenemos que los artículos definen el sesgo como algo que surge de datos de entrenamiento desequilibrados, diferencias arquitectónicas y contextos de implementación, mientras que las tipologías destacan fuentes estadísticas, de medición y sociales que conducen a una precisión de predicción desigual en herramientas educativas como los modelos de éxito estudiantil [12], [13], [14].

En este escenario, esta revisión sistemática aborda críticamente la intersección entre equidad educativa y aprendizaje automático, con tres objetivos fundamentales. Primero, identificar los tipos de sesgo algorítmico documentados en plataformas educativas, distinguiendo entre sesgos de género, socioeconómicos y culturales. Segundo, analizar las estrategias de mitigación propuestas en la literatura reciente, evaluando su efectividad y limitaciones. Tercero, y como contribución original, proponer un marco ético integrador para el desarrollo de inteligencia artificial inclusiva en educación que trascienda aproximaciones puramente técnicas para incorporar dimensiones pedagógicas, sociales y de justicia distributiva.

Por tanto, la pregunta central que guía esta investigación es ¿cómo pueden los sistemas de aprendizaje automático en educación ser diseñados e implementados para promover activamente la equidad y la inclusión, en lugar de perpetuar desigualdades existentes? Esta interrogante se desglosa en tres preguntas específicas: ¿qué tipos de sesgo algorítmico se han identificado en plataformas educativas y cómo

afectan diferencialmente a distintos grupos de estudiantes?; ¿qué estrategias de mitigación han demostrado efectividad en contextos educativos?; ¿qué principios éticos deben guiar el desarrollo de IA inclusiva para garantizar que la tecnología sirva a la justicia social en educación?

2. Marco teórico

2.1. Equidad algorítmica en la educación en la era digital

El concepto de equidad educativa ha evolucionado significativamente desde las concepciones meramente formalistas de igualdad de oportunidades. Autores como Fraser [15] proponen un marco tridimensional de justicia social que distingue entre justicia redistributiva (recursos), justicia reconocitiva (reconocimiento de identidades) y justicia representativa (participación en decisiones). En el contexto de la inteligencia artificial (IA) educativa, estas tres dimensiones adquieren relevancia crítica en torno a la distribución desigual de acceso a tecnologías avanzadas, el reconocimiento o invisibilización de determinadas formas de conocimiento y culturas en los datos de entrenamiento, y la exclusión de comunidades marginadas de los procesos de diseño algorítmico.

Selwyn [16] advierte que la tecnología educativa tiende a reproducir las estructuras de poder existentes, operando bajo lo que denomina “el mito de la neutralidad tecnológica”. Desde esta perspectiva, los sistemas de inteligencia artificial no son artefactos neutrales sino ensamblajes socio-técnicos que incorporan valores, supuestos e intereses de sus diseñadores. La equidad, por tanto, no puede evaluarse únicamente mediante métricas técnicas de “justicia algorítmica” sino que requiere un análisis situado de cómo estos sistemas afectan las experiencias vividas de estudiantes diversos [17].

Al analizar los componentes (medición, aprendizaje, acción) como fuentes de sesgo, en paralelo con la discriminación educativa histórica [13], se puede resaltar que los tonos de piel y el género, específicamente en las mujeres de piel oscura presentan tasas de error de hasta el 34,7% en clasificadores comerciales, frente al 0,8% de los hombres de piel clara [9]. Y en virtud del sesgo existente más allá de los datos, abarca la especificación y la implementación de problemas, esto trasciende a los datos de entrenamiento [14]. Otras tipologías directas limitadas en el área de educación, se encuentran en que la mayoría se basa en aprendizaje automático general, lo que insta a realizar auditorías específicas de este contexto [18].

2.2. Definiciones y tipologías del sesgo algorítmico

El sesgo algorítmico abarca disparidades en el rendimiento del modelo entre grupos protegidos, a menudo derivadas de desequilibrios en los datos, variables proxy que capturan prejuicios sociales y conjuntos de entrenamiento no representativos [13], [18]. Las tipologías distinguen el sesgo estadístico (por ejemplo, tasas de error desiguales), la imparcialidad basada en la similitud (paridad de grupo) y las nociones causales (resultados contrafácticos), y las aplicaciones educativas revelan una mayor clasificación errónea de las mujeres negras en los clasificadores de género debido a la morfología facial y los desequilibrios en el entrenamiento [9], [11]. Las tipologías sociales más amplias incluyen la discriminación indirecta a través de indicadores como los códigos postales que se correlacionan con el estatus socioeconómico [12].

El sesgo algorítmico puede definirse como errores sistemáticos y recurrentes en un sistema computacional que crean resultados injustos, privilegiando arbitrariamente a ciertas categorías de personas sobre otras [19]. Al respecto, Mehrabi et al. [20], proponen una taxonomía que distingue entre:

Sesgo en los datos: Se presenta cuando los datos de entrenamiento no representan adecuadamente a la población objetivo, contienen estereotipos sociales, o reflejan prácticas discriminatorias históricas.

Sesgo algorítmico: Este emerge de las decisiones de diseño del modelo, incluyendo la selección de características, la definición de funciones objetivo y las técnicas de optimización.

Sesgo de interacción: Se manifiesta cuando los usuarios interactúan con el sistema de maneras que refuerzan o amplifican los sesgos existentes.

Sesgo sistémico: Este es resultante del contexto institucional y social más amplio en el que se despliega el sistema.

En el dominio educativo, Baker y Hawn [7], han identificado patrones específicos de sesgo en aplicaciones como sistemas de tutoría inteligente, plataformas de aprendizaje adaptativo y herramientas de analítica del aprendizaje, documentando disparidades consistentes en precisión predictiva entre grupos raciales y socioeconómicos.

2.3. Interseccionalidad y discriminación algorítmica en contextos educativos

La perspectiva interseccional originada en el feminismo negro [21], resulta particularmente importante para analizar el sesgo en inteligencia artificial educativa. Los estudiantes no experimentan la discriminación a través de una sola dimensión de identidad, sino en la intersección de múltiples ejes de opresión: género, raza, clase, capacidad, lengua materna. Así tenemos que un sistema puede ser justo en términos de equidad de género, pero profundamente injusto en términos raciales; o puede funcionar adecuadamente para el estudiante promedio, pero fallar sistemáticamente para aquellos que ocupan intersecciones marginadas.

Por otra parte, Buolamwini y Gebru [9] demostraron empíricamente esta problemática en su estudio sobre sistemas de reconocimiento facial, que presentaban tasas de error significativamente mayores para mujeres de piel oscura. También en educación, estudios emergentes sugieren patrones similares: sistemas de evaluación automatizada de ensayos que penalizan variedades dialectales no estándar [22], o plataformas de tutoría que interpretan patrones de interacción culturalmente diversos como bajo compromiso [23].

El sesgo interseccional se amplifica en el caso de identidades superpuestas (Por ejemplo, género-raza), y las plataformas de aprendizaje automático (ML) en educación muestran menor precisión en la predicción del éxito para estudiantes pertenecientes a minorías raciales y de bajo nivel socioeconómico [10]. Las plataformas digitales aumentan las brechas, registrándose un 28% de los escolares estadounidenses que carecen de acceso a internet en casa o en la escuela, lo que amplía las brechas en habilidades y resultados vinculados a factores socioeconómicos y culturales [24] [25]. Además de hallar que los clasificadores de género presentan disparidades persistentes entre razas, influenciadas por diferencias faciales genéticas y ambientales [11], [26].

Dentro de lo que se concibe como la Caja Negra, los modelos de éxito estudiantil menos precisos para grupos raciales minoritarios que utilizan características estándar, son establecidos por Gándara et al. [10]. Los matices de género, que destacan las brechas de precisión interseccional en los parámetros de análisis facial, son considerados en Buolamwini & Gebru [9]. La brecha digital ha sido persistente, tal como lo establece Pierce & Cleary, con las disparidades en el acceso a la educación en el hogar según

el nivel socioeconómico obstaculizan la equidad [25]. En este sentido la evidencia ha mostrado un éxito mixto en la mitigación aplicando técnicas como las vistas generativas que reducen, pero no eliminan las brechas [27], [26].

2.4. Disparidades raciales amplificadas por modelos predictivos

Los sesgos algorítmicos en las plataformas de aprendizaje automático (AA) implementadas en entornos educativos perpetúan las desigualdades raciales, perjudicando sistemáticamente a los estudiantes negros, hispanos y nativos americanos, mientras que favorecen a los estudiantes blancos y asiáticos en las predicciones de rendimiento, erigiendo así nuevas barreras para la educación inclusiva [28].

Los modelos de aprendizaje automático entrenados con datos de estudiantes de escuelas secundarias y preparatorias estadounidenses presentan sistemáticamente sesgos que reflejan y exacerban los patrones de discriminación racial existentes en el sistema educativo [28]. Los modelos estándar que predicen el rendimiento en matemáticas otorgan una indulgencia excesiva, donde predetermine el beneficio de la duda en los estudiantes blancos y asiáticos, lo que produce pronósticos excesivamente optimistas, mientras que ofrecen predicciones pesimistas para los estudiantes negros, hispanos y nativos americanos [28]. Estos sesgos persisten a pesar de las altas métricas de precisión general, que ocultan las disparidades de los subgrupos y engañan a los educadores sobre la imparcialidad del modelo [28].

2.5 Los mecanismos clave que impulsan estas desigualdades incluyen

Como efectos de la composición del conjunto de datos que tiende a alterar la composición racial de los datos de entrenamiento producen compensaciones en los tipos de error. Aumentar la representación de grupos subrepresentados reduce los falsos negativos (omisión de estudiantes en riesgo) para esos grupos, pero eleva los falsos positivos (exageración de estudiantes de bajo riesgo), y viceversa [28]. Esta dinámica ofrece una vía para mitigar las brechas mediante el equilibrio estratégico de los conjuntos de datos.

En cuanto a las brechas de equidad ocultas, existe una alta precisión agregada que oculta errores de predicción racialmente desiguales, lo que subraya la necesidad de métricas de evaluación específicas para cada subgrupo, más allá de los parámetros estándar [28].

2.6 Intervenciones de equidad y sus limitaciones

Los esfuerzos para contrarrestar los sesgos mediante intervenciones de equidad de vanguardia, como la reponderación, la eliminación del sesgo adversarial y las predicciones calibradas, se han comparado con conjuntos de datos educativos, revelando resultados dispares [28]. Si bien algunas reducen las brechas raciales en las tasas de error, a menudo comprometen la utilidad general del modelo o no logran eliminar por completo las disparidades, lo que pone de relieve el desafío de lograr tanto la precisión como la equidad [28]. A continuación, en el cuadro 1 se presentan los tipos de intervención de equidad que impacta en las brechas.

Cuadro 1. Tipos de intervención de equidad que impacta en las brechas

Tipo de Intervención	Impacto en las Brechas de Predicción Racial	Compensaciones clave
Reequilibrio del conjunto de datos	Reduce los falsos negativos en grupos subrepresentados.	Aumenta los falsos positivos para grupos sobrerrepresentados
Eliminación de sesgos adversariales	Mejora la paridad entre subgrupos	Puede reducir la precisión general
Predicciones calibradas	Equilibra las tasas de error entre razas.	Requiere mucha capacidad computacional para conjuntos de datos grandes.

Paralelamente, según lo expuesto en el cuadro 1, la investigación en analítica del aprendizaje enfatiza la equidad algorítmica y la ausencia de sesgo como elementos fundamentales para la justicia educativa [29]. Las nociones tradicionales se centran en soluciones técnicas como la distribución equitativa de errores, pero los enfoques políticos emergentes en las ciencias del aprendizaje abogan por la integración de dimensiones de justicia social, como los desequilibrios de poder en la obtención de datos y la implementación de modelos [29]. En esta intersección surgen tensiones bajo las intervenciones puramente técnicas que corren el riesgo de pasar por alto las desigualdades sistémicas, mientras que los marcos políticos más amplios exigen colaboración interdisciplinaria para rediseñar la analítica y lograr un cambio positivo [29].

2.7 Implicaciones más amplias para la inclusión en la educación en aprendizaje automático

Los sesgos algorítmicos van más allá de la predicción e influyen en quiénes se involucran con el aprendizaje automático. Las minorías raciales/étnicas subrepresentadas tienen seis veces menos probabilidades de matricularse en cursos tradicionales de aprendizaje automático/IA en comparación con otras, a menudo citando un mayor interés en las ramificaciones sociales, culturales y políticas de la tecnología sobre la mecánica técnica [30]. Esta preferencia señala una barrera: los currículos convencionales priorizan los algoritmos, dejando de lado los debates sobre equidad que podrían ampliar la participación [30].

Resulta prometedor que el 72% de los estudiantes desinteresados en los cursos introductorios estándar expresen entusiasmo por alternativas que enfaticen los impactos éticos, filosóficos y sociales del aprendizaje automático/IA [30]. Un curso de tipo “Principios de Ciencias de la Computación”, similar a los Principios de Ciencias de la Computación AP, centrado en los sesgos y la justicia, podría fomentar la diversidad al alinearse con estos intereses [30].

2.8 Hacia plataformas de aprendizaje automático equitativas

Abordar estas barreras requiere estrategias multifacéticas:

Diseño de modelos transparentes: Exigir a los educadores la divulgación de las características influyentes y los desgloses de rendimiento racial [28].

Prácticas de datos inclusivos: Aprovechar los ajustes en la composición de los conjuntos de datos y el análisis disposicional para mejorar las predicciones tempranas sin amplificar los sesgos [28], [31].

Reformas curriculares: Integrar la ética y la justicia en la educación de Aprendizaje Automático desde el principio, utilizando enfoques de caja de cristal que desmitifiquen los sesgos [30], [32].

Cambios regulatorios y políticos: Desarrollar el análisis del aprendizaje hacia modelos de equidad política, garantizando que las intervenciones se centren tanto en factores individuales como sistémicos [29].

Si bien la evidencia empírica de los conjuntos de datos escolares de Estados Unidos proporciona información sólida, se necesita más investigación sobre los contextos globales, los impactos de las intervenciones longitudinales y las intersecciones con los sesgos de género o socioeconómicos para eliminar por completo estas barreras educativas [28], [29], [30].

3. Metodología

3.1 Diseño del estudio

Esta investigación adopta el formato de revisión sistemática de la literatura, siguiendo los lineamientos PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) para garantizar transparencia y rigor metodológico [33]. Se opta por este diseño dada la necesidad de sintetizar un cuerpo de literatura emergente pero fragmentado, que aborda la equidad en inteligencia artificial educativa desde disciplinas diversas incluyendo ciencias de la computación, ciencias de la educación, ética y estudios de ciencia y tecnología.

Esta revisión se estructura en torno a tres preguntas fundamentales:

PI1: ¿Qué tipos de sesgo algorítmico (de género, socioeconómico y cultural) se han documentado en plataformas de aprendizaje automático aplicadas a educación entre los años 2018 y 2025?

PI2: ¿Qué estrategias de mitigación se han propuesto y evaluado en la literatura, además de cuál es su efectividad reportada?

PI3: ¿Qué principios y marcos éticos se han desarrollado para guiar el diseño de inteligencia artificial inclusiva en contextos educativos?

3.2 Estrategia de búsqueda

Se realizaron búsquedas sistemáticas en cinco bases de datos académicas: Web of Science Core Collection, Scopus, ERIC (Education Resources Information Center), ACM Digital Library y IEEE Xplore. La selección responde a su cobertura de literatura tanto en ciencias de la educación como en computación y ética tecnológica. En este contexto, la estrategia de búsqueda combinó términos booleanos en tres clústeres conceptuales: El primer cluster (Tema tecnológico): (“artificial intelligence” OR “machine learning” OR “algorithm” OR “learning analytics” OR “adaptive learning” OR “predictive modeling”). El segundo cluster 2 (Tema educativo): (“education” OR “school” OR “higher education” OR “student” OR “learning” OR “classroom”). Y el tercer cluster (equidad): (“bias” OR “fairness” OR “equity” OR “inclusion” OR “discrimination” OR “justice” OR “marginalized” OR “inequality”).

Los términos se combinaron con operadores AND entre clúster y OR dentro de cada clúster. La búsqueda se limitó a publicaciones en inglés y español entre enero del año 2018 y marzo de 2025, periodo que captura la expansión de la inteligencia artificial generativa y la maduración de la literatura sobre ética algorítmica en educación.

Se establecieron ciertos criterios de inclusión y exclusión para la selección de estudios sobre sesgos en sistemas de inteligencia artificial educativa. Estos criterios abarcaron estudios empíricos (cuantitativos, cualitativos o mixtos) que analizan sesgos, revisiones sistemáticas y meta-análisis sobre equidad algorítmica, así como artículos conceptuales que proponían marcos éticos para una inteligencia artificial inclusiva. Además, se especificó que se requerían publicaciones que hubiesen pasado por un

proceso de revisión por pares y que abordaran al menos una dimensión de equidad, como, por ejemplo: género, nivel socioeconómico u origen cultural. Por otra parte, se excluyeron estudios que se centraban únicamente en aspectos técnicos sin considerar la equidad, la literatura gris no revisada por pares, publicaciones anteriores al año 2017 y se exceptúan estudios que no estuvieran disponibles en texto completo.

3.3 Proceso de selección y extracción de datos

La selección se realizó en tres fases. En la primera, se eliminaron duplicados mediante gestor bibliográfico Zotero. En la segunda, dos revisores independientes examinaron títulos y resúmenes aplicando los criterios de inclusión, con un acuerdo inter-revisores de $\kappa = 0,81$. Las discrepancias se resolvieron mediante discusión. En la tercera fase, se revisaron textos completos de los artículos preseleccionados.

La extracción de datos se estructuró de forma tal que realizara una captura en los datos bibliográficos, contexto educativo tipo de sistema analizado, los tipos de sesgo identificados, la metodología de detección, las estrategias de mitigación propuestas, los resultados reportados, y el marco ético subyacente.

La calidad metodológica de los estudios incluidos se evaluó mediante herramientas adaptadas según diseño: Mixed Methods Appraisal Tool (MMAT) para estudios empíricos, y AMSTAR-2 para revisiones sistemáticas y no se excluyeron estudios por baja calidad, considerándose dentro de la síntesis de hallazgos.

Dada la heterogeneidad metodológica de los estudios incluidos, se optó por una síntesis narrativa temática para los resultados, complementada con tablas de evidencia que resumían hallazgos claves. Se identificaron patrones transversales mediante análisis temático inductivo, siguiendo el marco de Braun y Clarke (2006) [34].

4. Resultados

4.1 Descripción de los estudios incluidos

La búsqueda inicial arrojó 847 registros. Luego de eliminar duplicados ($n=234$), se tamizaron 613 títulos y resúmenes, excluyendo 482 por no cumplir criterios de inclusión. Se revisaron 131 textos completos, de los cuales 98 fueron excluidos por razones detalladas en la Figura 1. La muestra final incluyó 33 estudios.

En función a esta revisión sistemática, se tiende a destacar que las desigualdades de la era digital han persistido a través de brechas en el acceso (dispositivos, habilidades) y algoritmos sesgados que estandarizan la reproducción del conocimiento por sobre el pensamiento creativo [35], [36]. Los análisis pospandemia ha confirmado que el aprendizaje electrónico amplifica la exclusión de las minorías culturales y de bajo nivel socioeconómico, ya que las escuelas reducen las brechas, pero los hogares se quedan atrás [24], [36]. Las directivas de la Unión Europea han abordado esta discriminación de forma indirecta, pero carecen de margen para los detalles algorítmicos [12]. Por consiguiente, en materia de equidad en la educación digital, en el Cuadro 2 se destacan ciertos aspectos resumidos que han destacado las recientes investigaciones.

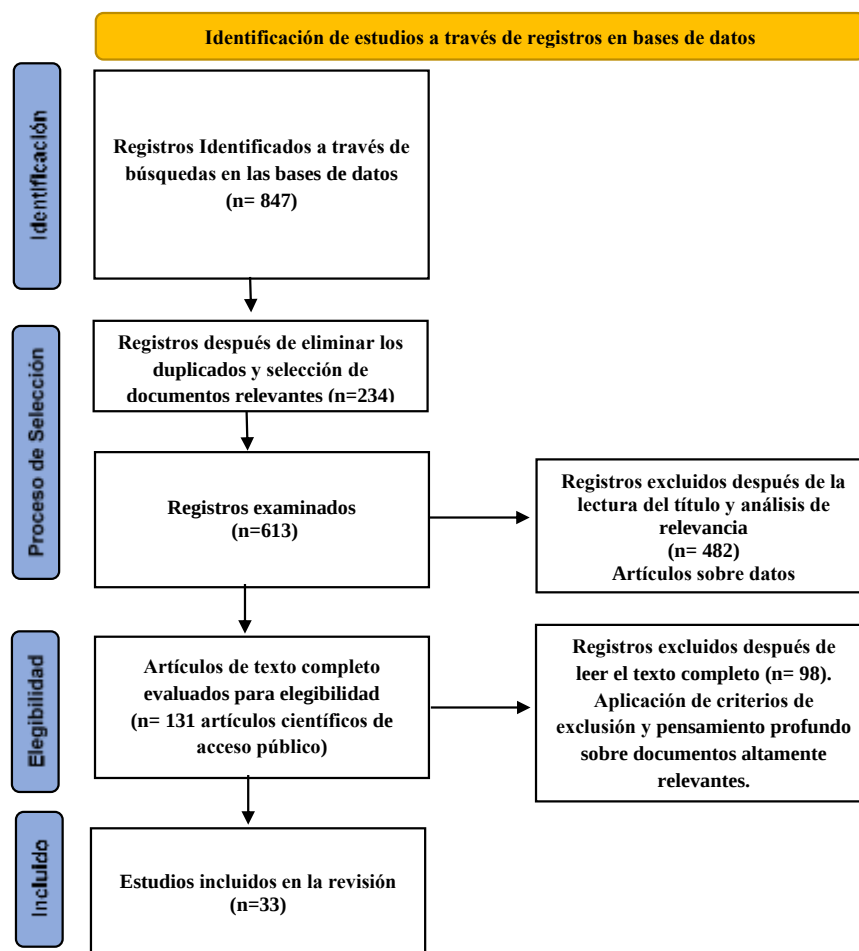


Figura 1. Diagrama de flujo para la revisión sistemática sobre la equidad y la inclusión en plataformas de aprendizaje automático

Cuadro 2. Hallazgos en los desafíos de equidad como brecha en el acceso a la educación digital

Aspecto	Hallazgo clave	Grupos afectados	Autores
Brecha de acceso	El 28% sin internet en casas o escuela; brechas de habilidades según nivel socioeconómico	Niños de bajo nivel socioeconómico	Pierce & Cleary [25]
Sesgo algorítmico en predicción	Menor precisión en modelos de éxito.	Minorías raciales	Gándara et al. [10]
Error en clasificación de género-raza	Hasta un 34,7% de clasificación errónea.	Mujeres de piel más oscura	Buolamwini & Gebru [9]
Límite de mitigación de sesgos	Las Técnicas de las visiones generativas, el aprendizaje estructurado y el aprendizaje evidencial reduce, pero persisten brechas	Grupos interseccionales	Ramachandran & Rattani [26]

Los artículos convergen en la necesidad de datos y auditorías inclusivas, pero señalan lagunas en las políticas, en algunos enfoques que han establecido los Estados Unidos sobre las comprobaciones a posteriori [14]. Además de enfatizar que existen algunas lagunas en la presente investigación que incluye los mecanismos causales en la tecnología educativa, los impactos interseccionales longitudinales y la eliminación de sesgos escalable para contextos de bajos recursos; el trabajo futuro debería estar enfocado en priorizar doctrinas de equidad sustancial, como el igualitarismo de la suerte [37]. Para aplicaciones a nivel del entorno clínico relacionado con las actividades educativas, los hallazgos informan que debe existir la supervisión humana en la implementación de estas herramientas, pero no reemplazando su aplicabilidad [13].

Las plataformas de aprendizaje automático generan discriminación interseccional en la educación a través de predicciones sesgadas y barreras de acceso, con mayores riesgos para las intersecciones de género, raza y nivel socioeconómico [10], [11], [13]. Las evidencias encontradas respaldan las definiciones o las tipologías de fuentes de datos y dentro de las mismas arquitecturas, pero tienden a exigir políticas integrales. En consecuencia, se debe fundamentar que los errores desiguales presentes en clasificadores y predicciones, deben ser abordados con una mitigación completa mediante las técnicas actuales [26], [35].

4.2 Tipos de sesgos documentados (PI1)

El análisis revela tres categorías principales de sesgo, con manifestaciones específicas en cada dimensión demográfica:

a. Sesgos de género

En esta revisión se pudo evidenciar que once estudios documentaron sesgos de género en sistemas educativos. Esto de validez a los hallazgos determinados por Idowu [38], en su revisión sistemática de 31 estudios sobre equidad algorítmica, donde se encontró que los sistemas de predicción de rendimiento mostraban consistentemente mayor precisión para estudiantes masculinos en disciplinas STEM, y que por otra parte sobreestimaban el riesgo de bajo rendimiento para mujeres en estos campos. Este patrón sugiere que los modelos aprenden correlaciones espurias entre género y desempeño a partir de datos históricos que reflejan brechas de participación, y no capacidades diferenciales.

Yin et al. [39] analizaron sistemas de recomendación de cursos en plataformas MOOC, encontrando que los algoritmos solían recomendar caminos formativos de menor dificultad a las alumnas, incluso cuando su desempeño previo era equivalente al de estudiantes varones. Este fenómeno, que los autores denominan canalización suave, opera mediante la asociación implícita entre género y preferencias profesionales en los datos de entrenamiento.

En el ámbito de la evaluación automatizada, estudios como el de Heggler et al. [40] identificaron que los sistemas de calificación de ensayos mostraban patrones de error sistemáticos que tendían a penalizar estilos de escritura asociados con géneros femeninos (mayor longitud, uso de lenguaje emocional, estructuras narrativas) mientras privilegiaban estilos más directos y argumentativos propios del género masculino.

b. Sesgos socioeconómicos

Quince estudios documentaron sesgos relacionados con nivel socioeconómico, constituyendo la categoría más documentada. En este orden de ideas, se compara con Baker y Hawn [41], quien encontró en su revisión de 18 estudios, que los sistemas predictivos de deserción estudiantil presentaban

sistemáticamente tasas de falsos positivos más altas para estudiantes de bajos ingresos y minorías. Es decir, estos estudiantes eran incorrectamente identificados como en riesgo con mayor frecuencia, lo que puede desencadenar intervenciones innecesarias o, más preocupantemente, expectativas reducidas por parte de educadores.

Chick [42], en una revisión comprehensiva de 167 fuentes sobre pensamiento computacional y equidad, identificó que los sistemas de aprendizaje adaptativo tienden a funcionar de manera más óptima para estudiantes con acceso previo a tecnología y alfabetización digital, mientras que aquellos de entornos desfavorecidos experimentan curvas de aprendizaje más pronunciadas que los sistemas que interpretan como bajo rendimiento, generando trayectorias de contenido simplificado que amplían la brecha inicial.

El estudio de Inam [23], basado en entrevistas cualitativas con educadores y estudiantes, reveló que la personalización algorítmica puede operar como un mecanismo de ordenamiento digital, donde los estudiantes de escuelas con recursos limitados generan menos datos de interacción, lo que resulta en modelos menos precisos y, en consecuencia, obtendrán experiencias de aprendizaje de menor calidad. Lo que indica que este círculo vicioso perpetúa desigualdades estructurales bajo la apariencia de neutralidad tecnológica.

c. Sesgos culturales y lingüísticos

En este contexto doce estudios abordaron sesgos culturales y lingüísticos. Por ejemplo, el estudio de Koenecke et al. [22], aunque estuvo centrado en el reconocimiento de voz general, es citado recurrentemente en la literatura educativa por sus implicaciones en los sistemas de reconocimiento de voz comercial, que permitieron ver tasas de error casi al doble para hablantes de Inglés Vernacular Afroamericano, comparado con el Inglés estándar. Esto tiene consecuencias directas para herramientas educativas que utilizan voz en la evaluación de lectura o interacción en segundas lenguas.

El estudio de Heggler et al. [40], que documentó cómo los sistemas de procesamiento de lenguaje natural en plataformas educativas tienden a privilegiar variedades lingüísticas estandarizadas, penalizando expresiones dialectales o bilingües. Esto afecta particularmente a estudiantes de comunidades migrantes y minorías lingüísticas, cuyas contribuciones son sistemáticamente subvaloradas por sistemas entrenados en corpus homogéneos.

Es necesario considerar el estudio de Buolamwini y Gebru [9], sobre sesgo en visión computacional por sus implicaciones en sistemas educativos emergentes que utilizan reconocimiento facial en el monitoreo de atención o compromiso emocional. Aunque la mayoría de estos sistemas aún no se implementan masivamente en aulas, su desarrollo plantea riesgos significativos para estudiantes de color, especialmente aquellos de piel más oscura.

d. Interseccionalidad de sesgos

Tres estudios adoptaron explícitamente una perspectiva interseccional. Tal como lo enfatiza Chick [42], al encontrar que los sesgos no operan independientemente, sino que se potencian mutuamente como: estudiantes, mujeres de bajos ingresos y minorías lingüísticas experimentaban formas compuestas de discriminación algorítmica y que no eran detectables al analizar cada dimensión por separado. Por ejemplo, un sistema podía ser justo en términos de género cuando se analizaba globalmente, pero profundamente injusto para la subpoblación de mujeres inmigrantes.

4.3 Estrategias de mitigación (PI2)

La literatura reporta diversas estrategias para mitigar sesgos, que pueden agruparse en tres categorías según el momento de intervención.

a. Intervenciones en pre-procesamiento

Los estudios evaluaron estrategias aplicadas antes del entrenamiento del modelo. La más común es el re-muestreo de datos para equilibrar representación de grupos (n=8). Esto ha sido reportado por Idowu [38] que mediante técnicas como SMOTE (Synthetic Minority Over-sampling Technique) pueden mejorar equidad, pero advierten que, si los datos originales contienen sesgos históricos, el sobre-muestreo puede simplemente replicar estos patrones con mayor densidad.

Por su parte, la auditoría de datos previa al entrenamiento es recomendada unánimemente (n=11 estudios). Heggler et al. [40] proponen un protocolo basado en el examen de datos que incluye: análisis de representación demográfica, detección de correlaciones espurias, identificación de proxies de variables sensibles, y consulta con comunidades afectadas sobre la adecuación cultural de las etiquetas y categorías utilizadas.

Una estrategia emergente es la des-proxificación concebida como la identificación y eliminación de variables que se correlacionan fuertemente con atributos protegidos sin ser explícitamente estos atributos idóneos. Por ejemplo, código postal como proxy de nivel socioeconómico, o patrones de navegación como proxy de género.

b. Intervenciones durante el procesamiento

Varios estudios examinaron modificaciones algorítmicas durante el entrenamiento. Los más documentados son:

Aprendizaje adversarial (n=7): entrena simultáneamente un modelo principal para maximizar precisión y un modelo adversarial para minimizar su capacidad de predecir atributos protegidos, éstos a partir de las predicciones. Por su parte Idowu (2024) reporta efectividad moderada, pero advierte sobre posible degradación de precisión general.

Las restricciones de equidad en optimización (n=8): incorporan métricas de equidad (Por ejemplo, paridad demográfica, igualdad de oportunidades) como términos en la función de pérdida durante el entrenamiento. Los estudios de Yin et al. encontraron que estas técnicas pueden reducir disparidades sin sacrificar significativamente precisión, contradiciendo la noción de compensación inevitable entre equidad y exactitud [39].

El aprendizaje de conciencia de grupo (n=4): representa modelos que aprenden representaciones separadas para diferentes grupos demográficos y luego combinan predicciones. Chick reporta resultados prometedores, pero señala riesgos de reforzar el esencialismo grupal, si no se implementa cuidadosamente [42].

c. Intervenciones en post-procesamiento

Los estudios que analizaron ajustes aplicados después del entrenamiento, incluyen:

La calibración de umbrales diferenciados (n=5): al ajustar los puntos de corte para decisiones (Por ejemplo, identificación de condición en riesgo) según el grupo demográfico para igualar las tasas de

error. En este sentido, Baker y Hawn (2022) reportan la efectividad práctica, pero advierten sobre los posibles desafíos legales si se interpreta como trato diferenciado [41].

Una auditoría continua (n=6): el monitoreo sistemático del desempeño del sistema desglosado por subpoblaciones para detectar degradación de equidad sobre tiempo.

Los mecanismos de apelación y de revisión humana (n=4): para permitir que estudiantes y educadores impugnen decisiones algorítmicas. Para Inam, este enfatiza la importancia de estos mecanismos desde una perspectiva de agencia y dignidad [23].

d. Efectividad comparada

La síntesis de hallazgos sugiere que no existe una estrategia universalmente superior; la efectividad depende del contexto específico, tipo de sesgo y aplicaciones educativas. Según Idowu (p. 1525) concluye que la mayoría de estudios no encontraron una compensación estricta entre equidad y precisión, desafiando la narrativa dominante de que mejorar la equidad necesariamente degrada el rendimiento [38]. Sin embargo, Chick matiza que esto puede deberse a que muchos estudios evalúan equidad en condiciones ideales, no en despliegues reales donde emergen complejidades imprevistas [42].

Una conclusión consistente en nueve estudios es que las intervenciones puramente técnicas son insuficientes. Eso se complementa con lo asumido por Heggler et al. [40] al afirmar que: la efectividad de la IA en educación depende de entrenamiento continuo e inclusivo de los algoritmos y la concientización de los usuarios sobre los riesgos de sesgos algorítmicos, permitiendo que los sistemas mejoren de maneras que prevengan la reproducción de injusticias sociales (p. 1).

4.4 Marcos éticos propuestos (PI3)

Doce estudios propusieron marcos éticos para guiar el desarrollo de IA inclusiva. De los cuales emergen cinco principios transversales.

a. Justicia distributiva en el diseño algorítmico

Tres de estos marcos enfatizan que la equidad algorítmica debe entenderse en términos distributivos: ¿quién se beneficia y quién es perjudicado por un sistema? Tomando en consideración a Chick [42], Así también, basándose en el trabajo de Fraser [15], propone que los sistemas de IA deben evaluarse no solo por su precisión promedio sino por su impacto en los grupos más marginados. Esto implica priorizar la mejora de desempeño para las poblaciones peor servidas, incluso si ello implica sacrificios menores en precisión para grupos mayoritarios.

b. Transparencia radical y explicabilidad

Cinco marcos enfatizan la transparencia como requisito ético fundamental. Por su parte Inam [23] argumenta que los estudiantes merecen entender no solo qué decisión tomó un sistema, sino por qué, y tener la capacidad para cuestionar esa decisión (p. 15). Esto implica diseñar sistemas que proporcionen explicaciones interpretables por personas sin formación técnica, y que revelen sus limitaciones y márgenes de error.

Meletiadou [17] propone que la alfabetización en inteligencia artificial debe ser un componente central de la educación, permitiendo que estudiantes desarrollen juicio crítico para evaluar si los resultados son confiables, actualizados e imparciales” (párr. 12).

c. Agencia del estudiante y autonomía

Cuatro marcos destacan la agencia como principio ético central. Inam [23] critica los modelos de personalización que reducen la autonomía del aprendiz, ya que simplifican excesivamente el aprendizaje y marginan a aprendices neurodiversos y culturalmente diversos (p. 2). En este sentido propone reconceptualizar la personalización no como optimización automatizada sino como proceso colaborativo y participativo que respeta la individualidad.

El marco Center for Implementing Digital and Distance Learning-CIDDL (Cardona y Rodríguez, 2024, citado en [43]), recomienda explícitamente prohibir que la IA tome decisiones finales sobre la elegibilidad para programas de educación individualizada, servicios y adaptaciones, acciones disciplinarias o progreso estudiantil” (párr. 15).

d. Diversidad cognitiva e inclusión

Tres marcos abordan específicamente la diversidad cognitiva. Se cita a Inam que acuña el término inteligencia inclusiva para referirse a sistemas diseñados para reconocer y valorar múltiples formas de pensar, aprender y expresar conocimiento [23]. Esto implica resistir a la tendencia de reducir la complejidad del aprendizaje a métricas cuantificables, y diseñar sistemas que puedan adaptarse a perfiles neurodiversos y estilos de aprendizaje no lineales.

e. Gobernanza participativa

Cinco marcos enfatizan la necesidad de participación comunitaria en el diseño y gobernanza de sistemas de IA. Retomando a Chick quien propone un modelo de co-construcción ética donde educadores, estudiantes, familias y comunidades participan en todas las fases del ciclo de vida algorítmico [42]. Complementando a su vez con Heggler et al. quienes abogan por un enfoque ético y colaborativo que involucre a múltiples actores en la supervisión de plataformas [40].

5. Discusión

5.1 La naturaleza socio-técnica del sesgo algorítmico

Los hallazgos de esta revisión confirman que el sesgo algorítmico en educación no es un problema puramente técnico sino un fenómeno socio-técnico complejo. Como documentan Barocas y Selbst [3], los sistemas de IA aprenden de datos que reflejan desigualdades estructurales, y al ser desplegados en contextos educativos, pueden amplificar estas desigualdades, mediante lo que Benjamin denomina la nueva codificación de la discriminación [8].

La evidencia de sesgos consistentes en sistemas predictivos, de recomendación y evaluación sugiere que la mera adopción de métricas de equidad algorítmica es insuficiente. Como advierten Baker y Hawn [7], las definiciones técnicas de equidad (paridad demográfica, igualdad de oportunidades, entre otros) pueden entrar en conflicto entre sí, y ninguna captura completamente la complejidad de la justicia educativa. Un sistema puede ser técnicamente justo según cierta métrica y simultáneamente perpetuar injusticias sustantivas.

5.2 La insuficiencia de las soluciones puramente técnicas

Una conclusión robusta de esta revisión es que las estrategias puramente técnicas de mitigación de sesgos, aunque son necesarias, son insuficientes. Idowu [38] recomienda explícitamente evaluar la

equidad de datos y sus características antes de la equidad algorítmica y evaluar el impacto de algoritmos justos en usuarios finales, ya que las percepciones humanas pueden no alinearse con medidas algorítmicas de equidad (p. 1511).

Esta desconexión entre métricas técnicas y experiencias vividas es críticamente importante. Un sistema puede satisfacer definiciones formales de equidad, pero ser experimentado como injusto por estudiantes que perciben que sus formas de conocer y expresarse no son valoradas. Inam [23] captura esta tensión al preguntar: ¿la inteligencia de quién cuenta cómo inteligencia? (p. 2).

5.3 Hacia un marco integrador de IA inclusiva en educación

Con base en la síntesis de hallazgos y los marcos éticos analizados, se propone como contribución un marco integrador para el desarrollo de IA inclusiva en educación, que estaría estructurado en cinco dimensiones interdependientes:

Dimensión 1: Justicia distributiva algorítmica. Los sistemas deben evaluarse por su impacto diferencial en grupos marginados, priorizando la mejora de desempeño para las poblaciones peor servidas. Esto requiere desagregar métricas por subpoblaciones relevantes y establecer umbrales mínimos de equidad que ningún grupo puede sobrepasar.

Dimensión 2: Transparencia radical y explicabilidad pedagógica. Los sistemas deben proporcionar explicaciones comprensibles para educadores y estudiantes sobre sus decisiones, incluyendo limitaciones y márgenes de error. La alfabetización crítica en IA debe integrarse en currículos para empoderar a estudiantes como evaluadores informados de tecnología educativa.

Dimensión 3: Agencia y autonomía del aprendiz. Los estudiantes deben tener la capacidad para cuestionar, apelar y, cuando sea apropiado, anular decisiones algorítmicas. La personalización debe ser colaborativa, no impuesta, reconociendo que los aprendices son agentes activos en su propio desarrollo.

Dimensión 4: Diversidad cognitiva y reconocimiento cultural. Los sistemas deben diseñarse para reconocer y valorar múltiples formas de inteligencia, expresión y aprendizaje. Esto implica entrenar modelos en datos diversos y representativos, y evaluar su desempeño no solo en su precisión sino en adecuación cultural.

Dimensión 5: Gobernanza participativa y democrática. Las comunidades educativas deben participar en el diseño, implementación y supervisión de sistemas de IA. Esto incluye mecanismos formales de consulta, representación en comités de ética y procesos de revisión continua que incorporen retroalimentación de usuarios diversos.

5.4 Limitaciones de esta revisión

Esta revisión tiene limitaciones que deben reconocerse. Primero, la exclusión de literatura gris puede omitir perspectivas relevantes de organizaciones comunitarias y experiencias de implementación no publicadas formalmente. Segundo, el enfoque en publicaciones en inglés y español puede sesgar los hallazgos hacia contextos occidentales, dejando de lado perspectivas de otros países suramericanos. Tercero, la rápida evolución de la IA significa que algunos hallazgos pueden quedar desactualizados rápidamente; se requiere monitoreo continuo de la literatura emergente.

Además, la heterogeneidad metodológica de los estudios incluidos dificulta comparaciones directas de efectividad entre estrategias. Las revisiones futuras podrían beneficiarse de enfoques meta-analíticos cuando la literatura madure y permita síntesis cuantitativas.

6. Conclusiones

La presente revisión sistemática evidencia que la incorporación de plataformas de aprendizaje automático en la educación, si bien promete personalización y eficiencia, enfrenta desafíos críticos de equidad. Los sesgos algorítmicos presentes tanto en datos, algoritmos e interacciones, constituyen barreras estructurales que reproducen y amplifican desigualdades históricas de género, nivel socioeconómico y origen cultural. Los hallazgos confirman que las estrategias de mitigación puramente técnicas son insuficientes; se requiere un enfoque socio-técnico que integre principios de justicia distributiva, transparencia, agencia estudiantil, diversidad cognitiva y gobernanza participativa. El marco ético propuesto ofrece una hoja de ruta para diseñar sistemas de IA verdaderamente inclusivos, donde educadores, comunidades y desarrolladores colaboren en la co-construcción de tecnologías al servicio de la justicia social. Futuras investigaciones deberían explorar la implementación de este marco en contextos diversos y evaluar su impacto longitudinal en la reducción de brechas educativas.

Referencias

- [1] Holmes, W., Bialik, M., & Fadel, C., “Artificial intelligence in education: Promises and implications for teaching and learning,” Center for Curriculum Redesign, pp. 1-39, 2019. URL: <https://curriculumredesign.org/wp-content/uploads/AIED-Book-Excerpt-CCR.pdf>
- [2] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L., *Intelligence unleashed: An argument for AI in education*, Londres: Pearson, 2016, pp. 1-60. URL: <https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/IntelligenceUnleashedSPANISH.pdf>
- [3] Barocas, S., & Selbst, A. D., “Big data's disparate impact,” *California Law Review*, vol. 104, nº 3, pp. 671-732, 2016. doi: <https://doi.org/10.15779/Z38BG31>
- [4] Noble, S. U., *Algorithms of oppression: How search engines reinforce racism*, New York, 2018, pp. 1-201. URL: <https://files.commonsgc.cuny.edu/wp-content/blogs.dir/6105/files/2019/01/SAFIYA-NOBLE.pdf>
- [5] O'Neil, C., *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown Publishing Group, 2016, pp. 1-272.
- [6] Akgun, S., & Greenhow, C., “Artificial intelligence in education: Addressing ethical challenges in K-12 settings,” *AI and Ethics*, vol. 2, nº 3, pp. 431-440, 2022. doi: <https://doi.org/10.1007/s43681-021-00096-7>
- [7] Akgun, S., & Greenhow, C., “Artificial intelligence in education: Addressing ethical challenges in K-12 settings,” *AI and Ethics*, vol. 2, nº 3, pp. 431-440, 2022. doi: <https://doi.org/10.1007/s43681-021-00096-7>.

- [8] Benjamin, R., *Race after technology: Abolitionist tools for the new Jim code*, Polity Press, 2019, pp. 1-172. doi:<https://doi.org/10.1109/tele55498.2022.9801048>.
- [9] Buolamwini & Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” *Proceedings of Machine Learning Research*, vol. 81, pp. 1-15, 2018. URL: <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>
- [10] Gándara, D., Anahideh, H., Ison, M. P., & Picchiarini, L., “Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction,” In *AERA Open: SAGE Publications*, vol. 10, pp. 1-15, 2024. doi:<https://doi.org/10.1177/23328584241258741>.
- [11] Krishnan, A., Almadan, A., & Rattani, A., “Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups,” In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE., p. 1028–1035, 2020. doi:<https://doi.org/10.1109/icmla51294.2020.00167>.
- [12] Aytekin, A., “Algorithmic Bias in the Context of European Union Anti-Discrimination Directives,” *EWAF’23: European Workshop on Algorithmic Fairness*, pp. 1-17, 2023. URL:<https://ceur-ws.org/Vol-3442/paper-41.pdf>.
- [13] Kizilcec, R. & Lee, H., “Algorithmic Fairness in Education,” *ArXiv*, pp. 1-30, 2020. doi:<https://doi.org/10.48550/arXiv.2007.05443>
- [14] Thais, S., Shumway, H., & Saragih, A., “Algorithmic bias: Looking beyond data bias to ensure algorithmic accountability and equity,” *MIT Science Policy Review*, vol. 4, pp. 59-66, 2023. doi:<https://doi.org/10.38105/spr.5lwvw66ssy>
- [15] Fraser, N., *Scales of justice: Reimagining political space in a globalizing world*, O. .: P. Press, Ed., Columbia University Press, 2008, pp. 1-361. URL:https://discovered.ed.ac.uk/discovery/fulldisplay/alma9924456616102466/44UOE_INST:44UOE_VU2
- [16] Selwyn, N., *Should robots replace teachers? AI and the future of education*, Polity Press, 2019, pp. 1-160. URL:<https://www.wiley.com/en-se/Should+Robots+Replace+Teachers%3F%3A+AI+and+the+Future+of+Education-p-9781509528967>
- [17] Meletiadou, E., “How educators can use Gen AI to promote inclusion and widen access,” 16 Mayo 2025. [En línea]. URL: <https://srheblog.com/2025/05/16/how-educators-can-use-gen-ai-to-promote-inclusion-and-widen-access/>.
- [18] Dehal, R. S., Sharma, M., & de Souza Santos, R., “Exposing Algorithmic Discrimination and Its Consequences in Modern Society: Insights from a Scoping Study,” *Proceedings of the 46th International Conference on Software Engineering: Software Engineering: Software Engineering in Society*, pp. 69-73, 2024. doi:<https://doi.org/10.1145/3639475.3640098>
- [19] Friedman, B., & Nissenbaum, H., “Bias in computer systems,” *ACM Transactions on Information Systems*, vol. 14, nº 3, pp. 330-347, 1996. doi:<https://doi.org/10.1145/230538.23056>

- [20] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, nº 6, pp. 1-35, 2021. doi:<https://doi.org/10.1145/3457607>
- [21] Crenshaw, K. "Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine," *University of Chicago Legal Forum*, vol. 1, nº 8, pp. 139-167, 1989. URL: <https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf>
- [22] Koenecke, A., Nam, E. L., J. Nudell, M. Quartey, Z. Mengesha, C. Touns, J. R. Rickford, D. Jurafsky, & Goel, S., "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, nº 14, pp. 7684-7689, 2020. doi:<https://doi.org/10.1073/pnas.1915768117>.
- [23] Inam, A., "Ethics of AI-driven personalization in learning: Designing for cognitive diversity and inclusive intelligence," *Artificial intelligence in education: Creating an equitable, creative, and effective learning environment*. IntechOpen, vol. 15, 2025. doi:<https://doi.org/10.5772/intechopen.1012981>.
- [24] Miras, S., Ruiz-Bañuls, M., Gómez-Trigueros, I. M., & Mateo-Guillen, C., "Implications of the digital divide: A systematic review of its impact in the educational field," *Journal of Technology and Science Education*, vol. 13, nº 3, pp. 936-950, 2023. doi:<https://doi.org/10.3926/jotse.2249>
- [25] Pierce, G., & Cleary, P., "The persistent educational digital divide and its impact on societal inequality," *PLOS ONE. Public Library of Science (PLoS)*, vol. 19, nº 4, pp. 1-23, 2024. doi:<https://doi.org/10.1371/journal.pone.0286795>.
- [26] Ramachandran, S., & Rattani, A., "Deep Generative Views to Mitigate Gender Classification Bias Across Gender-Race Groups," *Computer Vision and Pattern Recognition*, vol. V1, nº 2208.08382v1, pp. 1-20, 2022. doi:<https://doi.org/10.48550/arXiv.2208.08382>
- [27] Jindal, A., Singh, S. & Gadgil, S., "Classification for everyone: Building geography agnostic models for fairer recognition," *ArXiv: Computer Vision and Pattern Recognition*, vol. V3, nº 2312.02957v3, pp. 1-8, 2023. doi:<https://doi.org/10.48550/arXiv.2312.02957>
- [28] Santa U., Barbara, Michael D., Wu, Dasgupta, N., Médard, M., & Calmon, M.F., "Who Gets the Benefit of the Doubt? Racial Bias in Machine Learning Algorithms Applied to Secondary School Math Education," *Computer Science, Education, Mathematics*, pp. 1-10, 2022. URL: https://fated2022.github.io/assets/pdf/FATED-2022_paper_Jeong_Racial_Bias_ML_Algs.pdf
- [29] Uttamchandani, S., & Quick, J.D., "Chapter 20: An Introduction to Fairness, Absence of Bias, and Equity in Learning Analytics," *Education, Computer Science, Political Science. Handbook of learning analytics*, pp. 205-2012, 2022. doi:10.18608/hla22.020
- [30] Barretto, D., LaChance, J., Burton, E., & Liao, S. N., "Exploring Why Underrepresented Students Are Less Likely to Study Machine Learning and Artificial Intelligence," *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, pp. 457-463, 2021. doi:<https://doi.org/10.1145/3430665.3456332>

- [31] Tempelaar, D.T., Rienties, B., & Nguyen, Q., “The contribution of dispositional learning analytics to precision education,” *Educational Technology & Society*, vol. 24, nº 1, pp. 109-122, 2021. URL: <https://www.jstor.org/stable/26977861>
- [32] Morales-Navarro, L., & Kafai, Y.B., “Unpacking Approaches to Learning and Teaching Machine Learning in K-12 Education: Transparency, Ethics, and Design Activities,” *Proceedings of the 19th WiPSCE Conference on Primary and Secondary Computing Education Research*, vol. 2406.03480v3, pp. 1-11, 2024. doi:<https://doi.org/10.1145/3677619.3678117>
- [33] Page, Matthew; McKenzie, Joanne; Bossuyt, Patrick; Boutron, Isabelle; Hoffmann, Tammy; Mulrow, Cynthia; Shamseer, Larissa; Tetzlaff, Jennifer, “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *Research Methods & Reporting*, vol. 372, nº 71, pp. 1-9, 2021. doi:<https://doi.org/10.1136/bmj.n71>
- [34] Braun, V., & Clarke, V. , “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, nº 2, pp. 77-101, 2006. doi:<https://doi.org/10.1191/1478088706qp0630a>.
- [35] Bolshunova, T., Grigorieva, N., & Maslova, O., “Social Risks of Digitalization of Education,” In *2022 2nd International Conference on Technology Enhanced Learning in Higher Education (TELE): IEEE*, p. 66–69, 2022. doi:<https://doi.org/10.1109/tele55498.2022.9801048>
- [36] Han, X., & Li, Y., “Equity in Digital Education: Addressing the Digital Divide in a Post-Pandemic World,” In *Frontiers in Educational Research: Francis Academic Press Ltd. ,* vol. 8, nº 1, pp. 41-47, 2025. doi:<https://doi.org/10.25236/fer.2025.080107>
- [37] Arif Khan, F., Manis, E., & Stoyanovich, J., “Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines,” In *Equity and Access in Algorithms, Mechanisms, and Optimization: ACM*, nº 18, pp. 1-10, 2022. doi:<https://doi.org/10.1145/3551624.3555303>
- [38] Idowu, J. A., “Debiasing education algorithms,” *International Journal of Artificial Intelligence in Education*, vol. 34, nº 4, pp. 1510-1540, 2024. doi:<https://doi.org/10.1007/s40593-023-00389-4>
- [39] Yin, Z., Wang, Z., Chen, H., Li, Y., & Zhang, J. , “FairAIED: Navigating fairness, bias, and ethics in educational AI applications,” *Revista arXiv.*, vol. v2, nº 2407.18745, pp. 1-59, 2025. doi:<https://doi.org/10.48550/arXiv.2407.18745>.
- [40] Heggler, J. M., Szmoski, R. M., & Miquelin, A. F., “The dualities between the use of artificial intelligence in education and the risks of algorithmic biases.,” *Educacao e Sociedade*, vol. 46, nº e289323, pp. 1-20, 2025. doi:<https://doi.org/10.1590/ES.289323>
- [41] Baker, R. S., & Hawn, A., “Algorithmic bias in education,” *International Journal of Artificial Intelligence in Education*, vol. 32, nº 4, pp. 1052-1092, 2022. doi:<https://doi.org/10.1007/s40593-021-00285-9>
- [42] Chick, J. C., “AI-enhanced computational thinking: A comprehensive review of ethical frameworks and pedagogical integration for equitable higher education,” *Education Sciences*, vol. 15, nº 11, pp. 1-45, 2025. doi:<https://doi.org/10.3390/educsci15111515>

- [43] GovTech, “New guidelines aim to keep AI in schools human-centered. Government Technology,” 13 Agosto 2025. URL: <https://www.govtech.com/education/k-12/new-guidelines-aim-to-keep-ai-in-schools-human-centered>
- [44] Cardona, M. y Rodríguez, R., “Empowering Education Leaders: A Toolkit for Safe, Ethical, and Equitable AI Integration,” Washington, D.C., 2024. URL: <https://files.eric.ed.gov/fulltext/ED661924.pdf>.