Diálogos Educativos
Revista especializada en Ciencias de la Educación

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

Tendencias investigativas sobre la relación entre Inteligencia Artificial y acoso escolar en entornos educativos: revisión teórica

Research trends on the relationship between Artificial Intelligence and school bullying: a theoretical review

Jennyffer Astrid Piratova Carvajal
University of Technology and Education (UTE), Colombia.
https://orcid.org/0009-0009-7403-1380
jennyffer.piratovac2025@uted.us

Dr. Alonso José Larreal Bracho
University of Technology and Education (UTE), Colombia.
https://orcid.org/0009-0007-2270-2547
alonso.larreal@uted.us

Recibido: 1 abril 2026
Aceptado: 29 junio 2026
URL: https://relaticpanama.org/_journals/index.php/dialogoseducativos/article/view/181
DOI: https://doi.org/10.66707/9y0t1v73

Resumen

Este artículo presenta una revisión teórica de las tendencias investigativas que articulan el uso de la inteligencia artificial (IA) con el acoso y ciberacoso escolar en contextos educativos. Se realizó una búsqueda en Scopus, Web of Science (Core Collection), ERIC, SciELO, Redalyc y Dialnet, considerando publicaciones entre 2018 y 2025. La estrategia combinó consultas booleanas en español e inglés, normalización de descriptores y ampliación por referencias en cadena; se incluyeron estudios con aplicación educativa explícita (básica y media) y metodología reportada, y se excluyeron duplicados, textos de opinión sin sustento y trabajos fuera del ámbito escolar. Tras

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

el cribado en dos fases (título/resumen y texto completo) y la extracción mediante ficha estandarizada, la síntesis temática identificó cinco tendencias: (1) detección de ciberacoso mediante procesamiento de lenguaje natural y minería de texto; (2) analítica de aprendizaje y modelos predictivos para alerta temprana; (3) chatbots y agentes conversacionales para psicoeducación, reporte y acompañamiento; (4) visión por computador y sensado en espacios escolares, con debates sobre privacidad, proporcionalidad y consentimiento; y (5) percepciones de estudiantes y docentes sobre la IA (confianza, utilidad y ética). El análisis discute condiciones de validez educativa —corpora representativos, explicabilidad, evaluación justa y mitigación de sesgos— y propone un modelo conceptual por capas para integrar estas herramientas con protocolos y rutas institucionales. Se recomienda fortalecer la formación docente y la participación estudiantil informada. Finalmente, se delimitan vacíos para investigación futura, destacando evaluaciones longitudinales, estudios de eficacia en escenarios reales y análisis de efectos no intencionados sobre la convivencia escolar.

Palabras clave: acoso escolar; ciberacoso; educación ; inteligencia artificial; prevención

Abstract

This article presents a theoretical review of research trends that link the use of artificial intelligence (AI) with school bullying and cyberbullying in educational settings. A search was conducted in Scopus, Web of Science (Core Collection), ERIC, SciELO, Redalyc, and Dialnet, covering publications from 2018 to 2025. The strategy combined Boolean queries in Spanish and English, descriptor normalization, and snowballing through chained references; studies were included when they reported an explicit educational application (primary and secondary education) and a stated methodology, and duplicates, opinion pieces without evidence, and studies outside the school context were excluded. After a two-stage screening process (title/abstract followed by full text) and data extraction using a standardized form, the thematic synthesis identified five trends: (1) cyberbullying detection using natural language processing and text mining; (2) learning analytics and predictive models for early warning; (3) chatbots and conversational agents for psychoeducation, reporting, and support; (4) computer vision and sensing in

Diálogos Educativos
Revista especializada en Ciencias de la Educación

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

school spaces, alongside debates on privacy, proportionality, and consent; and (5) students' and teachers' perceptions of AI (trust, usefulness, and ethics). The analysis discusses conditions for educational validity—representative corpora, explainability, fair evaluation, and bias mitigation—and proposes a layered conceptual model to integrate these tools with institutional protocols and response pathways. It emphasizes human oversight, responsible data governance, and contextual adaptation to local coexistence policies and resources. Strengthening teacher professional development and informed student participation is recommended. Finally, research gaps are outlined, highlighting longitudinal evaluations, effectiveness studies in real-world scenarios, and analyses of unintended effects on school climate.

Keywords: bullying; cyberbullying; education; artificial intelligence; prevention

Introducción

El acoso escolar y su variante digital, el ciberacoso, continúan siendo retos de alta prioridad para los sistemas educativos por su impacto acumulativo en el rendimiento académico, el bienestar socioemocional y el clima de aula (Menesini y Salmivalli, 2017). Lejos de constituir fenómenos aislados, ambos se despliegan como continuos de conducta que transitan entre lo presencial y lo mediado por tecnologías, con dinámicas de persistencia, amplificación y difusión propias de entornos digitales —por ejemplo, el reenvío, la replicabilidad y la escalabilidad del daño— que complejizan la prevención y la respuesta , tal como lo afirman la Organización para la Cooperación y el Desarrollo Económicos (OCDE, 2021) y para la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, 2023). Esta complejidad obliga a replantear enfoques preventivos y de intervención, articulando acciones pedagógicas, protocolos institucionales y herramientas tecnológicas bajo criterios de proporcionalidad, debida diligencia y resguardo de derechos de niñas, niños y adolescentes (Council of Europe, 2022; UNESCO, 2023).

En paralelo, la inteligencia artificial (IA) ha madurado en capacidades aplicables al ámbito escolar —procesamiento de lenguaje natural, analítica de aprendizaje, visión por computador y agentes conversacionales— que posibilitan detección temprana, monitoreo de señales de riesgo, orientación

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

psicoeducativa y soporte a la toma de decisiones (Crompton y Burke, 2022; Paul y Saha, 2022; Elsafoury et al., 2021). No obstante, su adopción responsable exige más que desempeño algorítmico: requiere corpora representativos de la comunicación escolar, explicabilidad que dote de sentido a las señales que activan alertas, evaluaciones justas ante datos desbalanceados, y marcos de gobernanza que minimicen el uso de datos, protejan la privacidad y garanticen supervisión humana significativa. En instituciones públicas, además, las soluciones deben alinearse con marcos normativos, manuales de convivencia y procesos formativos del profesorado, así como con la participación estudiantil que legitime su uso pedagógico y su utilidad social.

En este marco, el presente artículo ofrece una revisión teórica de la literatura reciente que relaciona IA y acoso/ciberacoso escolar. El objetivo es sintetizar las principales tendencias investigativas, identificar resultados robustos y delimitar riesgos y vacíos de conocimiento para orientar líneas de acción en la práctica docente y en la política institucional. Como aportes operativos, se propone un modelo conceptual por capas para la integración de herramientas de IA con los protocolos escolares de prevención y respuesta, y delineamos una agenda de investigación centrada en evaluaciones longitudinales, auditorías de sesgo y análisis de efectos no intencionados. Finalmente, se describen implicaciones pedagógicas que desplazan el foco desde la mera detección técnica hacia ecosistemas preventivos en los que la IA complementa —sin sustituir— la dimensión humana del trabajo escolar.

Método de revisión teórica

Se realizó una revisión teórica de corte narrativo, sin pretensión meta-analítica, destinada a mapear tendencias y vacíos en el cruce IA–acoso escolar.

En primer lugar, se ejecutó una búsqueda estructurada en Scopus, Web of Science–Core Collection, ERIC, SciELO, Redalyc y Dialnet para el periodo 2018–2025, combinando consultas booleanas en español/inglés (en campos de título, resumen y palabras clave), normalización de descriptores y ampliación por referencias en cadena (rastreo *backward* y *forward*). En segundo término, se definieron criterios de inclusión que exigieron aplicación educativa explícita en educación básica y media, diseño metodológico reportado y resultados o propuestas sustentadas; y, de forma complementaria, criterios de exclusión para duplicados, textos de opinión sin respaldo y estudios fuera

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

del ámbito escolar.

Posteriormente, el cribado se realizó en dos fases (título/resumen y texto completo), registrando motivos de exclusión para asegurar trazabilidad. A continuación, la extracción se operativizó mediante una ficha estandarizada con variables sobre enfoque de IA (PLN, analítica, visión, chatbots), tarea pedagógica/preventiva, nivel y contexto educativo, métricas de evaluación e implicaciones éticas. Seguidamente, la codificación combinó enfoques inductivo-deductivos para alinear categorías emergentes con el marco teórico y, cuando fue pertinente, triangular evidencias cuantitativas y cualitativas. La síntesis fue temática y narrativa, apoyada en matrices comparativas para describir convergencias y divergencias entre estudios y sin estimación de tamaños de efecto, dada la heterogeneidad de diseños, poblaciones y medidas; esta decisión metodológica prioriza la validez interpretativa y la utilidad pedagógica de los hallazgos sobre su agregación estadística.

Finalmente, se documentaron limitaciones (p. ej., posible omisión de literatura de baja visibilidad y variabilidad terminológica) y se preservó la reproducibilidad del flujo de trabajo mediante cuaderno de bitácora y registro de decisiones. Para fortalecer la transparencia del reporte, se tomaron como guía recomendaciones para revisiones narrativas y para la presentación clara del proceso de búsqueda y selección (Snyder, 2019; Page et al., 2021), describiendo el flujo con lógica de identificación, cribado, elegibilidad e inclusión. El corpus final de fuentes clave revisadas fue de n = 22 (artículos académicos, revisiones/surveys y documentos normativos), seleccionadas por su pertinencia directa para comprender y orientar el uso de IA frente al acoso y ciberacoso escolar. De manera complementaria, se incorporaron referencias seminales previas cuando resultaron necesarias para sustentar el marco metodológico y ético del análisis (p. ej., privacidad y explicabilidad).

Extracción y síntesis

Se elaboró una matriz de extracción para normalizar la información clave de cada estudio, incorporando campos de autoría, año, país y/o muestra, nivel educativo, enfoque de IA (PLN, analítica de aprendizaje, visión por computador, chatbots), variable específica de acoso o ciberacoso, diseño metodológico, métricas empleadas (precisión, exhaustividad, F1, AUC, entre otras), principales hallazgos y notas sobre consideraciones éticas y de equidad. En segundo término, se aplicó una codificación inductivo-deductiva: partió de un esquema inicial derivado del marco teórico y se ajustó a

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

medida que emergieron categorías durante la lectura a texto completo; un subconjunto de artículos fue doble codificado para verificar convergencia y las discrepancias se resolvieron por consenso, dejando trazabilidad de las decisiones.

A continuación, se procedió a la triangulación entre resultados cuantitativos (rendimiento de modelos, sensibilidad/especificidad, análisis de error) y cualitativos (percepciones, contextos de implementación, condiciones institucionales), con el fin de mitigar sesgos propios de un único enfoque y fortalecer la validez interpretativa. Seguidamente, los estudios se agruparon en cinco tendencias temáticas y se compararon alcances y límites por tipo de contexto escolar (urbano/rural, público/privado), por nivel educativo (básica primaria y secundaria) y por madurez tecnológica (prototipo, piloto, despliegue), registrando supuestos y condiciones de transferencia. Por último, la síntesis siguió un enfoque narrativo orientado a la coherencia temática y a la utilidad pedagógica: se redactaron relatos comparativos apoyados en matrices de contraste y en resúmenes tabulares, evitando la agregación estadística de efectos debido a la heterogeneidad de diseños, medidas y contextos; con ello, se priorizó la claridad para la toma de decisiones educativas y la identificación de vacíos y riesgos operativos.

RESULTADOS: **Tendencias investigativas**

Detección automatizada de ciberacoso con PLN y minería de texto

En los últimos ciclos de investigación, los modelos basados en transformadores (BERT y derivados monolingües, multilingües y adaptados al dominio) han desplazado de forma consistente a enfoques clásicos como n-gramas con SVM o NB al clasificar mensajes hostiles u ofensivos, principalmente por su capacidad para modelar dependencias de largo alcance y representar el contexto discursivo a nivel de secuencia (Paul y Saha, 2022; Elsafoury et al., 2021; Rosa et al., 2019).

Este salto de desempeño, sin embargo, suele degradarse cuando se evalúa fuera del dominio de entrenamiento o en muestras reales de aula: la variación diatópica del español, la jerga juvenil y los códigos híbridos con anglicismos o emoticonografía introducen ambigüedades que erosionan la portabilidad del modelo (Hasan et al., 2023; Van Hee et al., 2018). Por ello, las prácticas recomendadas incluyen adaptación al dominio mediante reentrenamiento ligero, ajuste fino con ejemplos locales y uso

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

de estrategias de aumento de datos —traducción inversa, sinónimos controlados, perturbaciones léxicas— para robustecer la generalización (Hasan et al., 2023; Rosa et al., 2019).

Un reto persistente es la ambigüedad pragmática: ironía, sarcasmo y humor. Los sistemas basados exclusivamente en superficie textual tienden a confundir bromas entre pares con agresiones o a ignorar hostilidades veladas; la investigación reciente recurre a señales paralingüísticas y multimodales para desambiguar —fusión tardía o temprana de texto-imagen-emoji, incorporación de metadatos de interacción (quién habla a quién, frecuencia, latencia) y representación de la conversación como grafo— (Yi y Zubiaga, 2023), con ganancias dependientes de la calidad de los datos (Elsafoury et al., 2021).

Aun así, estos beneficios dependen críticamente de contar con corpora anotados de calidad; la falta de lineamientos claros de etiquetado y la baja concordancia interanotador conducen a modelos que optimizan contra etiquetas ruidosas. De ahí la necesidad de publicar guías de anotación, medir y reportar acuerdos (*Cohen's kappa*, *Krippendorff's alpha*) y documentar resolución de discrepancias (Van Hee et al., 2018). La identificación de roles —víctima, agresor, observador— añade complejidad técnica y conceptual. Los enfoques más sólidos combinan clasificación por turnos con etiquetado secuencial y modelado conversacional; otros adoptan marcos multitarea donde, además de toxicidad, se predicen intencionalidad, severidad y rol (Hasan et al., 2023).

En contextos escolares la distribución de clases es marcadamente desbalanceada; por tanto, métricas como macro-F1 y AUC-PR son preferibles a la exactitud, junto con técnicas de manejo del desbalance (ponderación de clases, *focal loss*, *oversampling* cuidadoso) y análisis de errores (Hasan et al., 2023; Rosa et al., 2019). La calibración probabilística y la selección informada de umbrales deben acompañar cualquier reporte: sin calibración, puntos de corte mal ajustados amplifican falsos positivos con consecuencias disciplinarias injustas (OECD, 2021).

Se observa, además, un viraje hacia modelos instruccionables y de pocas muestras: uso de grandes modelos de lenguaje en modo cero/pocas tomas con plantillas de *prompting* y verificación de consistencia. Aunque prometedores para dominios de baja cobertura, su comportamiento es sensible al *prompt* y puede introducir sesgos opacos; cuando se utilicen, conviene anclarlos con reglas de seguridad y capas de verificación, y compararlos contra *baselines* finamente ajustados. En paralelo, se exploran esquemas de aprendizaje activo y supervisión débil (lexicones, patrones de emojis) para reducir costos de anotación, siempre con revisiones humanas periódicas para contener derivas (Elsafoury et al., 2021).

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

Desde la perspectiva de despliegue responsable, la trazabilidad y la explicabilidad son condiciones de uso: técnicas pos-hoc como *SHAP*, inspección de *attention* y reglas de decisión ayudan a comprender qué rasgos activan la alerta y a formar al profesorado en su interpretación pedagógica. La integración de historial conversacional, centralidad en redes y contigüidad temporal mejora la detección de patrones relacionales y reduce falsos positivos, pero exige salvaguardas: minimización y seudonimización de datos, ventanas de retención limitadas, control de acceso por roles y auditorías periódicas de equidad para evitar impactos diferenciales por género, etnia, curso o discapacidad. Finalmente, toda solución debe someterse a validación externa entre centros, protocolos de evaluación realista en línea y revisión por comité de convivencia; solo así los avances del PLN se traducen en sistemas preventivos útiles, calibrados y justos en entornos escolares (Crompton y Burke, 2022).

Analítica de aprendizaje y modelos predictivos de riesgo

La analítica de aprendizaje integra fuentes heterogéneas que, una vez depuradas y sincronizadas temporalmente, permiten construir indicadores compuestos de vulnerabilidad. Entre ellos destacan: trazas de plataforma (tiempos de conexión, frecuencia de entregas, reincidencia en retrasos), patrones de interacción digital (densidad y reciprocidad en foros, latencia de respuesta, centralidad en redes de clase) y mediciones socioemocionales procedentes de autorreportes, observaciones docentes o registros de convivencia.

La utilidad de estos datos depende de su normalización por calendario escolar (cortes, evaluaciones, semanas especiales), cohortes y asignaturas, así como de la construcción de series de tiempo que capten tendencias y rupturas mediante ventanas móviles y diferencias intersemanales (Crompton y Burke, 2022). En este marco, los modelos supervisados convencionales (árboles de gradiente, regresiones penalizadas) y los enfoques secuenciales (HMM, LSTM o transformadores livianos) permiten detectar trayectorias atípicas que suelen anteceder episodios de hostigamiento o retraimiento, tales como caídas abruptas de participación, cambios en la posición de un estudiante dentro de la red social de aula, y picos anómalos de reportes o de lenguaje afectivo negativo.

Dado el desbalance estructural de clases pocos casos positivos de acoso frente a muchos negativos, la evaluación debe priorizar métricas robustas como F1 macro y AUC-PR, además de analizar sensibilidad a distintos umbrales más allá de la exactitud promedio.

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

La calibración probabilística (por ejemplo, *isotonic regression* o *Platt scaling*), junto con curvas de confiabilidad y *Brier score*, es necesaria para transformar probabilidades en decisiones accionables tipo semáforo y para fundamentar discusiones con equipos psicoeducativos sobre riesgos y costos de error. Asimismo, el uso de curvas precisión-recobrado y del análisis de ganancias de decisión ayuda a seleccionar puntos de corte acordes con recursos disponibles y con la asimetría de consecuencias entre falsos positivos y falsos negativos. De forma complementaria, la validación cruzada por centros o cursos y la evaluación temporal "prospectiva" entrenar en meses previos y probar en meses posteriores ofrecen estimaciones más realistas que particiones aleatorias.

La explicabilidad no es un accesorio, sino un requisito de adopción escolar. Herramientas de explicación local y reglas pos-hoc deben traducirse en narrativas comprensibles: qué combinación de señales activó la alerta (por ejemplo, descenso sostenido de participación, aislamiento en la red de foros y aumento de incidentes reportados) y qué acciones pedagógicas sugiere cada patrón contacto tutor-familia, mediación, refuerzo de habilidades socioemocionales o derivación. Este tipo de explicaciones facilita que el profesorado detecte errores de modelado como confundir timidez con riesgo de acoso, identifique sesgos inadvertidos y ajuste la respuesta con criterio, preservando el control humano significativo en la toma de decisiones (Council of Europe, 2022).

La eficacia real depende del acoplamiento con protocolos institucionales.

Se recomienda un flujo de *triage* con tres capas: primero, cribado automático y registro de la alerta con su explicación; segundo, revisión por un equipo humano (docente orientador y/o comité de convivencia) que valide evidencia, descarte artefactos y clasifique la urgencia; y tercero, asignación de rutas de apoyo graduadas —psicoeducación, tutoría, mediación o derivación clínica cuando proceda— con tiempos de respuesta máximos. Este flujo debe acompañarse de salvaguardas: minimización y seudonimización de datos, políticas de retención limitada, control de acceso por roles, bitácoras de auditoría y pruebas de equidad que monitoreen paridad de error por género, edad, curso o pertenencia a grupos minoritarios. Además, el seguimiento del *drift* —cambios en las distribuciones por calendario, cohortes o plataformas— y los ciclos de reentrenamiento programados previenen degradaciones silenciosas del desempeño (Crompton y Burke, 2022).

Más allá de predecir, es clave evaluar impacto preventivo. Diseños cuasi-experimentales, análisis

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

de series temporales interrumpidas y, cuando sea posible, ensayos controlados por conglomerados permiten estimar si las alertas reducen reincidencia, acortan tiempos hasta la intervención o mejoran indicadores de clima escolar. La analítica debe cerrarse con retroalimentación bidireccional: paneles discretos y no estigmatizantes para equipos docentes y directivos, y mensajes formativos para estudiantes y familias cuando corresponda; a la vez, la política institucional ha de fijar condiciones de salida del seguimiento para evitar etiquetados prolongados y promover la restitución plena una vez mitigado el riesgo. En conjunto, estos elementos convierten a los modelos predictivos de simples tableros de alerta en sistemas preventivos calibrados, explicables y útiles pedagógicamente.

Chatbots y agentes conversacionales para prevención y apoyo

Los chatbots educativos constituyen una interfaz de primera línea para psicoeducación, orientación y canalización de denuncias relacionadas con acoso y ciberacoso. En la práctica, coexisten tres familias tecnológicas: sistemas basados en reglas y árboles de decisión, asistentes de recuperación con base de conocimiento institucional (FAQ y rutas locals), y modelos generativos ajustados con salvaguardas; estas variantes permiten despliegues multimodales (texto/voz) y multicanal (intranet escolar, LMS, mensajería), ampliando el acceso fuera del aula y reduciendo barreras de búsqueda de ayuda (Lafrance y Villeneuve, 2024; Gabrielli et al., 2020). La literatura reciente subraya que la simple disponibilidad técnica no garantiza valor educativo: el rendimiento sostenido depende de la alineación con objetivos formativos y de la integración con los flujos de atención escolar existentes.

El diseño efectivo requiere cuatro pilares operativos: lenguaje claro y centrado en el estudiante; privacidad y minimización de datos, por ejemplo, diálogos efímeros cuando sea viable, seudonimización en registros de incidentes y políticas de retención acotadas; límites de actuación transparentes, el *bot* no diagnóstica, no sanciona y no sustituye a profesionales; y acoplamiento explícito con protocolos institucionales de convivencia y rutas de atención. En contextos escolares, es indispensable implementar detección de señales críticas, menciones de daño inminente, ideación suicida, violencia sexual o de género con umbrales de escalamiento automático hacia personal humano autorizado, registro de la alerta y trazabilidad de decisiones; estas funciones deben complementarse con barreras de seguridad frente a desinformación, sesgos y alucinaciones, así como con filtros de contenido y listas de exclusión que eviten respuestas inadecuadas.

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

En términos de efectividad, los estudios reportan incrementos en alfabetización sobre convivencia, mayor disposición a reportar y reducción del tiempo entre incidente y notificación cuando el *chatbot* se integra al ecosistema escolar con funciones de identificación de roles, guías de afrontamiento, formularios de reporte y seguimiento del caso (Gabrielli et al., 2020; Lafrance y Villeneuve, 2024). No obstante, se requieren evaluaciones más robustas con diseños longitudinales, análisis de series temporales y, cuando sea posible, ensayos por conglomerados que midan resultados intermedios —intención de búsqueda de ayuda, autoeficacia, uso de rutas y finales, disminución de reincidencia, mejora del clima escolar; además, deben incorporarse auditorías de equidad para detectar asimetrías por género, edad, curso, lengua o discapacidad.

Desde la gobernanza, conviene un marco operativo que combine consentimiento informado acorde con la normativa de protección de datos de menores, control de acceso por roles, bitácoras de auditoría, revisión periódica de contenidos, capacitación docente para interpretar y acompañar recomendaciones del *bot* y planes de continuidad del servicio. Asimismo, se recomiendan versiones lingüísticas y culturales localizadas, variedades del español y jergas escolares y criterios de accesibilidad, lectura en voz, contraste, navegación por teclado, a fin de asegurar inclusión y pertinencia. Finalmente, el *chatbo*t debe actuar como puerta de entrada no estigmatizante: acoger, orientar y derivar; la responsabilidad de valoración, intervención y seguimiento recae siempre en equipos humanos formados, quienes validan la información, contextualizan el caso.

Visión por computador y sensado en espacios escolares

Las soluciones basadas en visión por computador y sensado ambiental comprenden desde cámaras con inferencia en el borde hasta módulos especializados para reconocimiento de poses, detección de objetos, estimación de densidad de multitudes y análisis acústico de eventos anómalos. En términos funcionales, estos sistemas buscan identificar aglomeraciones inusuales, trayectorias compatibles con persecución, gestualidades asociadas a agresión o incrementos súbitos de ruido que podrían indicar peleas.

La inferencia local y el procesamiento por eventos, en lugar de grabación continua reducen exposición y volumen de datos, pero el rendimiento real se ve afectado por oclusiones, variaciones de luz, zonas ciegas y conductas ambiguas propias del juego, lo que exige revisión humana contextualizada

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

y ciclos periódicos de recalibración. En suma, su promesa técnica depende tanto de la calidad del diseño como de la integración con prácticas escolares y de la disponibilidad de protocolos de uso específicos para recreos, pasillos y canchas.

La adopción en centros con población menor de edad plantea dilemas de vigilancia, consentimiento y proporcionalidad. Por defecto, deben descartarse funcionalidades de identificación biométrica y reconocimiento facial, y restringir la finalidad del sistema a seguridad y convivencia bajo criterios de minimización de datos, retención corta y acceso diferenciado por roles. Antes de cualquier piloto, conviene realizar evaluaciones de impacto en derechos y protección de datos (DPIA, por sus siglas en inglés de *Data Protection Impact Assessment*; en español, evaluación de impacto en la protección de datos [EIPD]), con participación del comité de convivencia/ética y mecanismos de rendición de cuentas (señalética visible, políticas públicas de uso, registro de incidentes y vías de reclamación). La proporcionalidad exige demostrar que no existe una alternativa menos intrusiva de eficacia comparable y que los beneficios superan los riesgos previsibles para la privacidad y la autonomía de estudiantes y personal (Council of Europe, 2022).

Desde la perspectiva operativa, es indispensable un flujo de *triage* con control humano significativo: toda alerta debe ser verificada in situ por personal autorizado antes de cualquier decisión disciplinaria; quedan proscritas acciones automáticas derivadas únicamente de la señal algorítmica. La evaluación debe reportar curvas precisión-recobrado, análisis por umbrales y costos de error, así como auditorías de equidad que monitoricen paridad de desempeño por género, edad, curso o pertenencia a grupos minoritarios; además, se recomienda validación temporal y entre centros para estimar generalización fuera del dominio de entrenamiento. En paralelo, se deben documentar planes de continuidad (fallas, cortes de energía, canales alternos), ciberseguridad y criterios de actualización de modelos, junto con bitácoras de auditoría y formación docente para interpretar correctamente las alertas.

Por último, las políticas institucionales deberían priorizar intervenciones educativas y soluciones menos intrusivas, utilizando la analítica espacial en forma agregada para rediseñar zonas de conflicto, escalonar recreos, mejorar la supervisión adulta y fortalecer competencias socioemocionales. En este enfoque, los sistemas de visión/sensado actúan como apoyo focal, calibrado y temporalmente acotado, nunca como sustituto de la acción formativa ni de la deliberación humana; su legitimidad depende de transparencia, participación de la comunidad educativa y revisión periódica de impactos y

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

externalidades.

Percepciones de estudiantes y docentes

La aceptación de la IA en la escuela depende de creencias de utilidad, facilidad de uso y control humano, pero también de la confianza que el profesorado deposita en que las herramientas no vulneran derechos ni desplazan su juicio pedagógico. La evidencia comparativa en seis países muestra que la confianza docente aumenta cuando crece la autoeficacia con IA y la comprensión de cómo funcionan los sistemas; esas variables median la percepción de beneficios y preocupaciones (p. ej., sesgo, opacidad) y, en conjunto, predicen mayor disposición a usar IA en *K-12* (Viberg et al., 2024).

De forma convergente, se ha argumentado que avanzar en la adopción requiere entender primero a los educadores, sus necesidades, restricciones de tiempo y criterios de responsabilidad, más que insistir solo en capacidades técnicas (Kizilcec, 2024). En estudios de percepción, las actitudes favorables se asocian con formación específica, soporte institucional y fiabilidad de la herramienta, mientras que la falta de explicabilidad, la incertidumbre ética y la escasez de garantías de privacidad disminuyen la aceptación (Yim y Wegerif, 2024). En términos normativos, las orientaciones internacionales insisten en transparencia, explicabilidad, minimización de datos, participación estudiantil y supervisión humana significativa como condiciones que habilitan la confianza y legitiman usos pedagógicos, no exclusivamente disciplinarios de la IA.

Discusión

Los hallazgos convergen en que el valor educativo de la IA no reside únicamente en su capacidad de detección, sino en su articulación con ecosistemas preventivos que integren prácticas pedagógicas, alfabetización digital crítica y protocolos de convivencia con control humano significativo. En esta lógica, los modelos de PLN y analítica de aprendizaje deben operar como apoyos a procesos ya existentes —observación docente, mediación, rutas de denuncia— y no como sustitutos de la deliberación profesional. La literatura sobre confianza y aceptación docente subraya que la utilidad percibida depende de la alineación con finalidades pedagógicas y de la transparencia respecto de lo que el sistema puede o no puede hacer; cuando se comunica la lógica de decisión y se garantiza supervisión

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

humana, la disposición a usar IA en *K-12* aumenta (Kizilcec, 2024; Viberg et al., 2024).

En paralelo, el enfoque de derechos planteado por organismos internacionales exige explicabilidad, minimización de datos y evaluaciones de impacto antes del despliegue, con especial cuidado al tratar información de menores. Además, la operacionalización del control humano y la comunicación transparente puede apoyarse en guías de interacción humano-IA que recomiendan informar al usuario, permitir corrección y explicar decisiones en el momento oportuno (Amershi et al., 2019). En la misma línea, la literatura de IA explicable advierte que las explicaciones deben ser fieles, útiles y comprensibles para evitar falsas sensaciones de certeza en decisiones sensibles (Barredo et al., 2020).

Para sistemas escolares públicos, la adopción responsable se materializa en diseños centrados en el estudiante y en la comunidad: conjuntos de datos representativos del habla escolar y de las variedades del español; métricas de rendimiento que prioricen la equidad y la calibración; y flujos de *triage* que unan la señal algorítmica con la verificación humana, la comunicación con familias y la activación de apoyos graduados. La evidencia sobre detección de ciberacoso muestra mejoras con transformadores y arquitecturas conversacionales, pero también advierte sobre sensibilidad al dominio, desbalance de clases y riesgos de sobreconfianza si no hay validación externa y análisis de errores (Elsafoury et al., 2021; Hasan et al., 2023; Paul y Saha, 2022).

Por ello, las recomendaciones incluyen validación temporal y entre centros, auditorías de sesgo y reportes estandarizados que expliciten límites y condiciones de transferencia, así como formación docente para interpretar explicaciones locales y tomar decisiones pedagógicas proporcionales. En particular, las auditorías de sesgo se benefician de marcos académicos sobre *bias* y *fairness* que distinguen fuentes de sesgo en datos, etiquetas y modelos, y orientan métricas y estrategias de mitigación (Mehrabi et al., 2021). Asimismo, la tradición de analítica de aprendizaje ha enfatizado dilemas como consentimiento, obligación de actuar y límites del monitoreo, útiles para anticipar riesgos en contextos escolares (Slade y Prinsloo, 2013).

Finalmente, los componentes tecnológicos deben insertarse en una estrategia institucional más amplia: currículo de ciudadanía y bienestar digital; dispositivos de reporte accesibles y no estigmatizantes, incluidos chatbots con salvaguardas; y espacios de co-diseño con estudiantes y docentes que incrementen la pertinencia y la confianza (Gabrielli et al., 2020; Lafrance St-Martin y Villeneuve,

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

2024). Este enfoque desplaza el énfasis desde "detectar para sancionar" hacia "comprender para prevenir", con métricas de éxito que trasciendan la precisión técnica y midan resultados educativos y de convivencia —tiempos hasta la intervención, reincidencia, percepción de clima escolar—, de acuerdo con un marco de gobernanza que combine transparencia, rendición de cuentas y mejora continua.

Modelo conceptual para prevención asistida por IA

Se propone un modelo en cinco capas: (1) Gobernanza ética (políticas, roles, evaluación de impacto); (2) Datos responsables (minimización, anonimización, seguridad); (3) Algoritmos explicables (métricas, auditoría, umbrales con intervención humana); (4) Prácticas pedagógicas (psicoeducación, denuncia, mediación, seguimiento); y (5) Evaluación y mejora (indicadores de convivencia, satisfacción de la comunidad, costo-efectividad). La Tabla 1 resume el modelo conceptual propuesto.

Tabla 1. *Modelo conceptual por capas para prevención asistida por IA en escuelas públicas*

Capa	Componentes	Indicadores sugeridos
Gobernanza ética	Política IA escolar; comité; DPIA.	Actas; evaluación de impacto documentada.
Datos responsables	Mínimos necesarios; anonimización; seguridad.	Tasa de incidentes; tiempos de borrado.
Algoritmos explicables	Métricas; auditoría; umbrales.	AUC/F1; % falsas alarmas; revisiones humanas.
Prácticas pedagógicas	Currículo; rutas de denuncia; mediación.	Reportes; asistencia; reincidencia.
Evaluación y mejora	Monitoreo y ajuste continuo.	Tendencias de clima escolar; encuestas.

Nota. Elaboración propia, informada por orientaciones internacionales. DPIA = Evaluación de Impacto en Protección de Datos (*Data Protection Impact Assessment*); AUC = área bajo la curva; F1 = medida F1.

Implicaciones para la práctica y la investigación

Para la práctica, los hallazgos recomiendan operacionalizar la adopción de IA en tres frentes.

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

Primero, dotar al profesorado de manuales de uso con ejemplos situados —casos de aula, guías de interpretación de métricas, pautas de actuación proporcional— y módulos de ciudadanía digital que articulen prevención, reporte y cuidado entre pares. Segundo, institucionalizar protocolos de *triage* con tiempos máximos de respuesta, escalamiento a equipos humanos y registros trazables, acompañados de formación específica en explicabilidad (lectura de señales locales, límites del modelo) y en sesgos/errores frecuentes para evitar decisiones automáticas indeseadas.

Tercero, profesionalizar la relación con proveedores mediante acuerdos de servicio y cláusulas de protección de datos: minimización y seudonimización, evaluación de impacto previa al despliegue, pruebas de equidad, validación externa y planes de continuidad; estas condiciones elevan la confianza y alinean la tecnología con las finalidades pedagógicas del centro (Kizilcec, 2024; Viberg et al., 2024). Cuando se utilicen chatbots, su valor aumenta si se integran a rutas institucionales y contienen salvaguardas comunicadas con claridad (Gabrielli et al., 2020). Estas prácticas se alinean con pautas de interacción humano-IA y con enfoques de explicabilidad orientados a usuarios no expertos, que enfatizan retroalimentación, capacidad de corrección y claridad sobre incertidumbre (Amershi et al., 2019; Barredo et al., 2020).

Para la investigación, se identifican cinco prioridades. (a) Desarrollo y apertura de corpus en español —con variedad dialectal y registros juveniles— que permitan entrenar y evaluar modelos con validez ecológica; la literatura muestra que las brechas idiomáticas limitan la portabilidad de sistemas entrenados en otros contextos (Hasan et al., 2023). (b) Evaluaciones longitudinales y diseños cuasi-experimentales que midan no sólo desempeño técnico sino resultados educativos y de convivencia (tiempos a la intervención, reincidencia, clima de aula). (c) Estudios en contextos rurales y de baja conectividad, con dispositivos ligeros y metodologías de participación estudiantil y docente para aumentar la pertinencia local. (d) Auditorías participativas de sesgo y transparencia —con docentes, familias y estudiantes— que exploren efectos no intencionados y diseñen estrategias de mitigación. (e) Reportes estandarizados que incluyan calibración, análisis de error y validación entre centros, favoreciendo la reproducibilidad y la toma de decisiones informada.

Limitaciones

Esta revisión es de corte narrativo y, por la heterogeneidad de enfoques, métricas y poblaciones,

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

no realiza síntesis cuantitativa. La estrategia de búsqueda, aunque amplia, pudo omitir literatura gris o estudios de baja visibilidad y concentra principalmente publicaciones en español e inglés. Dado el rápido ritmo de cambio tecnológico y regulatorio en IA educativa, los hallazgos deben actualizarse periódicamente para reflejar nuevas evidencias, estándares y marcos de gobernanza.

Conclusiones

La evidencia revisada converge en que la inteligencia artificial puede aportar valor real a la prevención del acoso escolar cuando se implementa como apoyo a la gestión pedagógica y socioemocional, no como un mecanismo automático de sanción. En particular, los enfoques más prometedores combinan: (i) analítica y detección temprana para priorizar riesgos, (ii) intervención educativa (convivencia, habilidades socioemocionales y ciudadanía digital) y (iii) rutas institucionales claras con acompañamiento de orientación escolar y participación familiar.

En términos metodológicos, la literatura muestra un predominio de estudios de validación técnica (precisión, recall, F1) y una menor proporción de evaluaciones de efectividad en contextos reales (impacto en clima escolar, reducción de incidentes, percepción de seguridad, bienestar y participación estudiantil). Por ello, la pertinencia de la IA en educación pública debe juzgarse con indicadores mixtos (técnicos, pedagógicos y psicosociales) y con diseños robustos (cuasi-experimentales o experimentales cuando sea viable) que contemplen sesgos, deriva de datos y adaptaciones culturales.

En el plano ético y de gobernanza, el mayor riesgo identificado no es únicamente el error del modelo, sino el uso indebido de datos estudiantiles, la vigilancia excesiva y la estigmatización. Se recomienda priorizar: minimización y anonimización de datos, consentimiento informado cuando aplique, transparencia sobre fines y límites del sistema, auditorías de sesgo y mecanismos de apelación y revisión humana. La escuela debe conservar la decisión final y documentar criterios de actuación para garantizar debido proceso y enfoque restaurativo.

A partir del modelo conceptual por capas presentado, se proponen cuatro recomendaciones operativas para instituciones educativas: 1) fortalecer la alfabetización digital y la cultura de reporte seguro; 2) articular la IA a los comités y protocolos existentes (convivencia, orientación, protección); 3) capacitar docentes en interpretación de alertas y en intervención pedagógica; y 4) evaluar continuamente

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

la implementación con métricas de convivencia, satisfacción y aprendizaje. Como líneas futuras, se sugiere investigar la eficacia comparada de intervenciones asistidas por IA frente a intervenciones tradicionales, así como el desempeño de modelos explicables en poblaciones diversas y en contextos rurales o con conectividad limitada.

Referencias

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Su, H., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., y Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), 1-13. https://doi.org/10.1145/3290605.3300233

Barredo Arrieta, A., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., y Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Council of Europe. (2022). Guidelines on children's data protection in education settings. Council of Europe. https://www.coe.int/en/web/data-protection/guidelines-education

Crompton, H., Jones, M. V., y Burke, D. (2022). Affordances and challenges of artificial intelligence in K–12 education: A systematic review. Journal of Research on Technology in Education, 1–21. https://doi.org/10.1080/15391523.2022.2121344

Elsafoury, A., Palade, V., y Ellahham, S. (2021). A comprehensive survey on automated detection of cyberbullying: Progress and challenges. IEEE Access, 9, 106247–106270. https://doi.org/10.1109/ACCESS.2021.3100428

Gabrielli, S., Rizzi, S., Carbone, S., y Donisi, V. (2020). A chatbot-based coaching intervention for adolescents to promote life skills: Pilot randomized controlled trial. JMIR Human Factors, 7(1), e16762. https://doi.org/10.2196/16762

Hasan, M., Parde, N., y Yu, D. (2023). A comprehensive survey of automatic cyberbullying detection. Future Internet, 15(5), 179. https://doi.org/10.3390/fi15050179

Kizilcec, R. F. (2024). To advance AI use in education, focus on understanding educators. International Journal of Artificial Intelligence in Education. *Advance online publication*. https://doi.org/10.1007/s40593-023-00351-4

Lafrance St-Martin, A., y Villeneuve, F. (2024). Agents conversationnels en contexte scolaire: Défis, opportunités et recommandations. Observatoire international sur les impacts sociétaux de l'IA et du numérique (OBVIA). https://observatoire-ia.ulaval.ca/publications/agents-conversationnels-contexte-scolaire

Menesini, E., y Salmivalli, C. (2017). Bullying in schools: The state of knowledge and effective interventions. Psychology, Health y Medicine, 22(S1), 240–253.

ISSN L 3072-9815
Vol.2. NÚM 1.
Abril 2026 a marzo 2027
pp. 397-415

https://doi.org/10.1080/13548506.2017.1279740

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., y Galstyan, A. (2021). A survey on *bias* and *fairness* in machine learning. ACM Computing Surveys, 54(6), 1-35. https://doi.org/10.1145/3457607

OECD. (2021). OECD digital education outlook 2021: Pushing the frontiers with AI, blockchain and robots. OECD Publishing. https://doi.org/10.1787/589b283f-en

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., y Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Paul, S., y Saha, S. (2022). CyberBERT: BERT for cyberbullying identification. Multimedia Systems, 28(6), 1897–1904. https://doi.org/10.1007/s00530-020-00710-4

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Paulino, P., Simão, A. M. V., y Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333–345. https://doi.org/10.1016/j.chb.2018.12.021

Slade, S., y Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57(10), 1509-1528. https://doi.org/10.1177/0002764213479366

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

UNESCO. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000385723

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., y Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PLOS ONE, 13(10), e0203794. https://doi.org/10.1371/journal.pone.0203794

Viberg, O., Hatakka, M., Samuelsen, J., Mavroudi, A., y Gerlič, I. (2024). What explains teachers' trust in AI in education across six countries? International Journal of Artificial Intelligence in Education. *Advance online publication*. https://doi.org/10.1007/s40593-023-00376-9

Yi, P., y Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. Online Social Networks and Media, 36, 100250. https://doi.org/10.1016/j.osnem.2023.100250

Yim, I. H. Y., y Wegerif, R. (2024). Teachers' perceptions, attitudes, and acceptance of artificial intelligence (AI) educational learning tools: An exploratory study on AI literacy for young students. Futures of Education Research, 3(1), e65. https://doi.org/10.1002/fer3.65