JTEM : Journal of Technology in Educational Management
Lontara DIGITECH INDONESIA
E-ISSN : XXXX-XXXX
P-ISSN : XXXX-XXXX

# Pemanfaatan Kecerdasan Buatan (AI) dalam Asesmen Pendidikan: Implikasi Etika, Bias, dan Tantangan Implementasi

Achmad Shabir*
*Universitas Negeri Makassar*

*Email: achmadshabir@unm.ac.id*

***ARTICLE INFO***

**Kata Kunci:** AI Ethics, Algorithmic Bias, Artificial Intelligence, Automated Scoring, Educational Assessment

***ABSTRACT***

*This study examines the utilization of artificial intelligence (AI) in educational assessment, focusing on the ethical implications, biases, and implementation challenges. A total of 376 articles were identified between 1986 and 2025, with 302 articles found to be relevant. After filtering for publications from 2021 to 2025, the number of relevant articles decreased to 264. Of these, 263 articles were not retracted, 262 articles were not duplicates, and 176 articles were available as open access full-text. After further screening based on inclusion criteria (primary empirical research, excluding systematic reviews, scoping reviews, literature reviews, meta-analyses, narrative reviews, framework papers, qualitative studies, or conceptual papers), only 28 articles met the inclusion criteria. Key findings indicate that while AI can enhance the quality and accuracy of assessments, its effectiveness heavily depends on the implementation context and the quality of methods used. Major challenges related to AI implementation in assessment include algorithmic bias, data privacy issues, and lack of transparency in decision-making processes. The study also discusses potential mitigations for these issues through bias audits, fairness-aware algorithms, and human oversight in AI-based assessment systems. Overall, AI has the potential to improve fairness and accessibility in educational assessment systems, provided it is implemented with inclusive design and careful attention to ethical and bias-related concerns.*

**ARTICLE INFO**

**Kata Kunci:** Asesmen Pendidikan, Bias Algoritmik, Etika AI, Kecerdasan Artificial, Penilaian Otomatis

**ABSTRACT**

Studi ini mengkaji pemanfaatan kecerdasan buatan (AI) dalam asesmen pendidikan, dengan fokus pada implikasi etika, bias, dan tantangan implementasi. Dari 376 artikel yang ditemukan dengan rentang tahun 1986 hingga 2025, 302 artikel teridentifikasi sebagai artikel yang relevan. Setelah memfilter berdasarkan tahun publikasi antara 2021 hingga 2025, jumlah artikel yang relevan berkurang menjadi 264. Dari jumlah tersebut, 263 artikel tidak retracted, 262 artikel bukan duplikat, dan 176 artikel tersedia dalam open access full-text. Setelah dilakukan penyaringan lebih lanjut berdasarkan kriteria inklusi (penelitian empiris primer, bukan systematic review, scoping review, literature review, meta-analisis, narrative review, framework paper, studi kualitatif, atau studi konseptual), hanya 28 artikel yang memenuhi kriteria inklusi. Temuan utama menunjukkan bahwa meskipun AI dapat meningkatkan kualitas dan akurasi penilaian, efektivitasnya sangat bergantung pada konteks implementasi dan kualitas metode yang digunakan. Tantangan utama terkait implementasi AI dalam asesmen mencakup bias algoritmik, masalah privasi data, dan kurangnya transparansi dalam pengambilan keputusan. Penelitian ini juga membahas potensi mitigasi masalah tersebut melalui audit bias, algoritma yang lebih sadar keadilan, serta pengawasan manusia dalam sistem asesmen berbasis AI. Secara keseluruhan, AI berpotensi meningkatkan keadilan dan aksesibilitas dalam sistem asesmen pendidikan, asalkan diterapkan dengan desain yang inklusif dan perhatian terhadap isu etika dan bias.

## 1. PENDAHULUAN

Penggunaan kecerdasan buatan (AI) dalam asesmen pendidikan telah mengalami perkembangan pesat, dengan penerapan teknologi seperti Machine Learning (ML), Natural Language Processing (NLP), dan Generative AI. Teknologi-teknologi ini menawarkan potensi untuk mengubah cara penilaian pendidikan dilakukan, dengan memberikan penilaian otomatis, pemberian umpan balik adaptif, dan pengujian berbasis komputer yang lebih efisien dan akurat (1,2). Namun, meskipun AI dapat memperbaiki kualitas dan akurasi penilaian, berbagai tantangan etika dan bias algoritmik muncul dalam penerapannya.

Dalam banyak studi yang telah dilakukan, AI terbukti meningkatkan efektivitas asesmen, terutama dalam penilaian tugas-tugas terstruktur seperti penilaian esai dan penilaian MCQ (*Multiple Choice Questions*), yang memiliki kriteria penilaian yang jelas dan terdefinisi dengan baik (3,4). Misalnya, sistem AI berbasis deep learning seperti BERT dan CNN menunjukkan tingkat akurasi yang sangat tinggi dalam penilaian esai (5). Dalam konteks yang lebih spesifik, AI juga digunakan untuk memberikan umpan balik adaptif yang disesuaikan dengan tingkat pemahaman siswa, yang dapat meningkatkan engagement dan motivasi dalam proses pembelajaran (6).

Namun, meskipun AI dapat meningkatkan kualitas penilaian, masalah besar muncul terkait bias algoritmik. Beberapa penelitian menunjukkan bahwa sistem AI dapat memperburuk ketidakadilan sosial dan disparitas pendidikan jika data yang digunakan untuk melatih model AI tidak representatif atau tidak inklusif (3,7). Bias algoritmik ini dapat terjadi dalam berbagai bentuk, seperti bias demografis, bias budaya, atau bias sosial-ekonomi, yang sering kali merugikan kelompok siswa tertentu, seperti siswa dari latar belakang sosial-ekonomi rendah atau siswa minoritas (1,8).

Privasi data juga menjadi isu penting dalam penerapan AI dalam asesmen pendidikan. Sistem AI memerlukan akses ke data pribadi siswa, yang menimbulkan kekhawatiran mengenai keamanan data dan penggunaan data yang tidak sah (9). Selain itu, kurangnya transparansi dalam cara sistem AI membuat keputusan juga menimbulkan masalah terkait akuntabilitas dan kepercayaan terhadap teknologi ini (10,11).

Sistem AI dalam asesmen pendidikan juga menimbulkan pertanyaan tentang keadilan dalam penilaian. Meskipun AI dapat memberikan penilaian yang lebih konsisten dan objektif dalam penilaian tugas terstruktur, tantangan muncul dalam penilaian tugas kognitif yang lebih kompleks yang membutuhkan pemahaman kontekstual yang lebih dalam (11). Hal ini menunjukkan bahwa meskipun AI unggul dalam tugas-tugas terdefinisi dan terstruktur, sistem AI masih memiliki keterbatasan dalam mengenali nuansa atau konteks dalam penilaian yang lebih kompleks, seperti penilaian berbasis kreatifitas atau penilaian keterampilan klinis (12).

Di sisi lain, penerapan AI dalam penilaian adaptif dan personalisasi menunjukkan hasil yang menjanjikan. AI memungkinkan penyesuaian konten ujian sesuai dengan kemampuan dan kemajuan individu siswa, yang dapat meningkatkan hasil belajar dan menurunkan beban kognitif siswa dalam proses penilaian (6). Namun, penelitian menunjukkan adanya paradoks beban kognitif- kualitas, di mana pengurangan beban kognitif yang disebabkan oleh penilaian berbasis AI dapat merugikan kualitas penalaran dan analisis siswa (11).

Sementara itu, tantangan dalam implementasi AI dalam asesmen mencakup hambatan teknis seperti kualitas data, integrasi sistem, dan skala sistem AI, serta hambatan organisasi seperti resistensi terhadap perubahan, kekurangan pelatihan pendidik, dan keterbatasan infrastruktur digital (13,14). Selain itu, kesadaran dan persiapan pendidik untuk menggunakan AI dalam penilaian juga memengaruhi keberhasilan implementasi sistem ini (2,15).

Secara keseluruhan, meskipun AI dalam asesmen pendidikan menawarkan banyak keuntungan, penerapannya memerlukan perhatian terhadap etika, keadilan, privasi, dan kualitas

data yang digunakan untuk melatih sistem. Untuk memaksimalkan potensi AI dalam penilaian pendidikan, desain sistem yang inklusif, pengawasan manusia, dan pengembangan kebijakan yang hati-hati sangat diperlukan. AI harus digunakan untuk melengkapi penilaian manusia, bukan untuk menggantikannya, dan implementasinya harus mempertimbangkan faktor kontekstual serta tantangan etika yang muncul.

## 1. METODE PENELITIAN

Penelitian ini menggunakan pendekatan Systematic Literature Review (SLR) untuk mengeksplorasi pemanfaatan kecerdasan buatan (AI) dalam asesmen pendidikan, serta implikasi etika dan bias yang terkait. Proses SLR dilakukan dalam beberapa tahap, dimulai dengan pencarian artikel, penyaringan berdasarkan kriteria yang telah ditentukan, ekstraksi data, dan sintesis hasil.

### Pencarian dan Penyaringan Artikel

Pencarian dimulai dengan menggunakan kata kunci yang relevan untuk topik AI dalam asesmen pendidikan. Pencarian dilakukan di berbagai database ilmiah terkemuka, termasuk PubMed, Crossref, Google Scholar, Scopus, Web of Science, Semantic Scholar, dan Preprint servers. Total artikel yang ditemukan sebanyak 376 artikel, yang mencakup publikasi dari tahun 1986 hingga 2025.

Setelah dilakukan penyaringan, hanya 302 artikel yang teridentifikasi sebagai artikel yang relevan dengan topik. Artikel-artikel ini kemudian difilter lebih lanjut berdasarkan tahun publikasi (antara 2021 dan 2025), yang menghasilkan 264 artikel. Dari jumlah tersebut, satu diantaranya retracted dan ditemukan dua duplikat, sehingga tersisa 262 artikel. Namun hanya terdapat 176 artikel tersedia dalam open access full-text.

Setelah dilakukan screening manual terhadap abstrak, 28 artikel memenuhi kriteria inklusi. Kriteria inklusi yang diterapkan adalah sebagai berikut:

1. Desain penelitian empiris (penelitian eksperimental, kuasi-eksperimental, observasional, studi kasus, atau mixed-methods).
2. Studi primer (bukan systematic review, scoping review, literature review, bibliometric review, meta-analisis, narrative review, framework paper, studi kualitatif, atau studi konseptual).
3. Artikel yang membahas aplikasi AI dalam asesmen pendidikan, dengan fokus pada penilaian otomatis, pemberian umpan balik adaptif, atau pengujian berbasis komputer.

### Ekstraksi Data

Setelah artikel yang memenuhi kriteria inklusi dipilih, langkah selanjutnya adalah ekstraksi data dari artikel yang relevan. Data yang diekstraksi meliputi:

1. Detail Teknologi AI: Jenis AI yang digunakan (misalnya, machine learning, natural language processing (NLP), deep learning, dan generative models seperti GPT), algoritma atau model yang diterapkan, dan sumber data yang digunakan untuk melatih sistem AI.
2. Aplikasi dalam Asesmen: Bagaimana AI diterapkan dalam asesmen pendidikan, mencakup jenis asesmen (formative, summative, diagnostic, adaptive, personalized), format asesmen (MCQ, esai, berbasis kinerja), dan fungsi asesmen (penilaian, generasi soal, deteksi plagiarisme, umpan balik otomatis).
3. Konteks Pendidikan: Tingkat pendidikan (K-12, pendidikan tinggi, pengembangan profesional), jenis institusi (publik, swasta, online), serta lokasi geografis dan konteks budaya tempat penelitian dilakukan.
4. Evaluasi Efektivitas: Metode yang digunakan untuk mengukur efektivitas sistem AI dalam asesmen, termasuk metodologi evaluasi (desain eksperimental, kelompok pembanding, dll.), metrik yang digunakan (akurasi, reliabilitas, validitas, kinerja siswa, kepuasan), serta hasil kuantitatif dengan ukuran efek dan signifikansi statistik.
5. Masalah Etika yang Ditemukan: Kekhawatiran etika, bias yang teridentifikasi (algoritmik, demografis, budaya), masalah privasi dan perlindungan data, serta transparansi dan masalah explainability.
6. Strategi Mitigasi: Pendekatan yang digunakan untuk mengatasi kekhawatiran etika dan bias, termasuk solusi teknis (algoritma fairness-aware, AI yang dapat dijelaskan), perlindungan prosedural (pengawasan manusia, proses validasi), kebijakan atau langkah-

langkah tata kelola, serta pelatihan atau intervensi pendidikan untuk pengguna.
7. Tantangan Implementasi: Hambatan teknis (kualitas data, integrasi sistem, skalabilitas), hambatan organisasi (resistensi terhadap perubahan, keterbatasan sumber daya, kebutuhan pelatihan), serta masalah adopsi dan kepercayaan pengguna.
8. Dampak terhadap Pemangku Kepentingan: Pengaruh terhadap hasil siswa, dampak pada pendidik, efek pada institusi, serta implikasi yang lebih luas terhadap sistem pendidikan.

**Sintesis dan Analisis**

Setelah data diekstraksi, langkah berikutnya adalah sintesis tematik dan analisis terhadap temuan utama dari artikel-artikel yang disaring. Sintesis ini mencakup pengelompokan temuan berdasarkan tema utama, seperti teknologi AI yang digunakan, keefektifan dalam penilaian, implikasi etika, bias dalam algoritma, dan tantangan implementasi. Analisis dilakukan untuk memahami hubungan antara teknologi AI dan berbagai variabel yang mempengaruhi hasil asesmen pendidikan.

.

## 2. HASIL DAN DISKUSI

Tinjauan sistematis ini mencakup 28 studi empiris primer yang meneliti pemanfaatan Kecerdasan Buatan (AI) dalam asesmen pendidikan dari berbagai belahan dunia. Geografinya mencakup Timur Tengah (Yordania, Arab Saudi, Pakistan), Eropa (Slovenia, Denmark, Jerman, Portugal, Turki), Asia (Indonesia, Malaysia, Kazakhstan), Australia, dan Amerika Serikat.

Sebagian besar studi (18 dari 28) dilakukan di lingkungan pendidikan tinggi, mencakup bidang seperti pendidikan medis, kimia, ilmu komputer, dan linguistik. Sejumlah penelitian juga berfokus pada konteks K-12 (sekolah dasar dan menengah) serta pengembangan profesional (1,2,5,14). Dari segi metodologi, terdapat keragaman yang mencakup studi eksperimental, survei, pengembangan model, studi kasus, dan metode campuran (*mixed-methods*). Teknologi AI yang diteliti sangat beragam, mulai dari model generatif seperti ChatGPT dan GPT-4, arsitektur *deep learning* (BERT, CNN, *transformer*), hingga pendekatan *machine learning* tradisional seperti *Bayesian Networks* dan model *Item Response Theory* (IRT).

**Hasil Analisis Tematik**

**1) Pemanfaatan Teknologi AI dan Aplikasi dalam Asesmen Pendidikan**

Hasil analisis menunjukkan ekosistem teknologi AI yang kaya dan beragam yang diaplikasikan untuk fungsi asesmen yang luas. Aplikasi dapat dikategorikan berdasarkan tingkat otomasi dan kompleksitas tugas.

**a) Model Generatif AI untuk Pembuatan dan Umpan Balik Konten**

Model generatif, khususnya varian ChatGPT dan GPT-4, telah banyak diadopsi untuk fungsi pembuatan soal (item generation) dan pemberian umpan balik otomatis. Studi oleh Sabqat et al (8) mengeksplorasi penggunaan ChatGPT untuk mendeteksi dan mengoreksi kelemahan pada soal pilihan ganda (MCQ) dalam pendidikan medis. Sementara itu, Karademir dan Altan (2) memanfaatkan GPT-4 untuk mengevaluasi rencana pelajaran dalam pedagogi desain teknik, membandingkannya dengan evaluasi pakar. Aplikasi serupa untuk generasi soal dan umpan balik dilaporkan dalam studi-studi survei yang mengukur persepsi pengguna (16–18).

**b) Deep Learning dan NLP untuk Penilaian Kompleks dan Otomatis**

Arsitektur *deep learning* dan Pemrosesan Bahasa Alami (NLP) menunjukkan efektivitas tinggi dalam menilai tugas-tugas kompleks yang sebelumnya hanya dapat dilakukan oleh manusia. Penilaian esai otomatis merupakan area yang matang, dengan model seperti DualBERT-Trans-CNN (5) dan varian BERT (10) mencapai akurasi dan kesepakatan antar-penilai (Cohen's κ) yang setara bahkan melampaui kesepakatan antar-manusia. Aplikasi meluas ke bidang lain seperti penilaian keterampilan klinis menggunakan CNN (12), evaluasi kreativitas figural dengan *vision transformers* (19), dan deteksi emosi melalui analisis pose (*PoseNet*) untuk asesmen afektif (1).

**c) Machine Learning Tradisional untuk Diagnostik dan Personalisasi**

Pendekatan *machine learning* yang lebih tradisional dan dapat diinterpretasi tetap relevan, terutama untuk asesmen diagnostik dan pembelajaran adaptif. Guzmán dan Millan (20)

membandingkan model IRT berbasis *copula* dengan *Bayesian Networks* untuk estimasi pengetahuan, menunjukkan korelasi tinggi dengan penilaian pakar. Sistem seperti RiPPLE memanfaatkan model inferensi probabilistik untuk meningkatkan akurasi penilaian sejawat dan umpan balik individual (15). Demikian pula, platform *deep learning* untuk jalur pembelajaran personal dilaporkan mampu meningkatkan nilai siswa secara signifikan (6,13).

**d) Level Otomasi dan Integrasi**

Tingkat otomasi bervariasi dari sistem terotomasi penuh (seperti penilaian esai dan *debugging* kode) hingga sistem semi-otomatis dengan manusia dalam loop (*human-in-the-loop*). Studi-studi menekankan bahwa untuk tugas kompleks dan berisiko tinggi, integrasi dengan penilaian manusia dan pengawasan ahli tetap krusial (10,12). Sistem pendukung keputusan yang lebih sederhana, seperti metode *Simple Additive Weighting* untuk penilaian karakter, juga mengadopsi pendekatan semi-otomatis di mana guru memberikan input akhir (14).

Keragaman aplikasi ini mencerminkan evolusi AI dari alat otomasi tugas sederhana menjadi mitra kolaboratif dalam asesmen formatif dan sumatif. Namun, temuan ini juga menyoroti paradoks efisiensi-kualitas. Sementara AI sangat efisien untuk tugas terstruktur, penerapannya pada tugas kognitif tinggi memerlukan desain yang cermat untuk memastikan kedalaman pemahaman dan penalaran tetap terjaga, sebagaimana diperingatkan oleh studi Stadler et al., (11) tentang penurunan kualitas justifikasi saat menggunakan LLM.

**2) Dampak AI pada Kualitas, Akurasi, dan Hasil Belajar**

Kinerja sistem AI dalam asesmen menunjukkan heterogenitas yang substansial, yang sangat dipengaruhi oleh kompleksitas tugas, kualitas desain sistem, dan ranah pengetahuan.

**a) Kinerja pada Tugas Terstruktur vs. Kompleks**

Bukti konsisten menunjukkan keunggulan AI pada tugas terstruktur dengan rubrik jelas. Ganne dan Leddo (3) melaporkan keakuratan AI sebesar 95% dalam *debugging* kode, mengungguli manusia (89%). Martin et al., (10) mencapai akurasi 87% ($\kappa$=0.86) dalam penilaian argumentasi kimia. Sebaliknya, pada tugas kompleks yang membutuhkan pemahaman kontekstual mendalam, kinerja AI lebih bervariasi dan sering kali di bawah manusia. Sallam dan Al-Salahat (4) menemukan bahwa ChatGPT 3.5 menjawab 80% MCQ mikrobiologi medis dengan benar, tetapi nilainya (80.5/100) secara signifikan di bawah rata-rata mahasiswa (86.21/100). Demikian pula, pelatihan khusus pada ChatGPT untuk mendeteksi kelemahan MCQ tidak menghasilkan peningkatan signifikan dalam kemampuannya (8).

**b) Pengaruh Rekayasa *Prompt* dan Desain Sistem**

Faktor implementasi, khususnya rekayasa *prompt* (*prompt engineering*), merupakan moderator kritis efektivitas AI. Karademir dan Altan (2) menunjukkan perbedaan dramatis: *prompt* terstruktur menghasilkan kesepakatan tinggi dengan evaluasi pakar (ICC = 0.708), sementara prompt tidak terstruktur menghasilkan keluaran yang tidak konsisten dan tidak dapat diandalkan (ICC = 0.076). Temuan ini menegaskan bahwa kualitas input dan desain tugas secara langsung memengaruhi validitas dan reliabilitas asesmen berbasis AI.

**c) Dampak pada Hasil Belajar dan Keterampilan Kognitif**

Dampak AI terhadap hasil belajar tidak selalu linier dan positif. Beberapa studi melaporkan peningkatan signifikan. Fawad Naseer et al. (2024) melaporkan peningkatan nilai sebesar 25% dengan platform pembelajaran personal berbasis *deep learning*. Masoud Rahimi et al. (2024) menemukan bahwa pembelajar Bahasa Inggris sebagai Bahasa Asing (EFL) yang menggunakan umpan balik korektif otomatis (AWCF) seperti Grammarly mengungguli kelompok kontrol dalam prestasi tugas dan akurasi tata bahasa.

Namun, temuan kritis dari Stadler et al., (11) mengungkap "paradoks beban kognitif-kualitas". Meskipun LLM mengurangi beban kognitif siswa selama pengumpulan informasi, mereka menghasilkan penalaran dan justifikasi dengan kualitas yang lebih rendah dibandingkan dengan siswa yang menggunakan mesin pencari tradisional. Ini menunjukkan bahwa kemudahan kognitif yang diberikan AI mungkin mengorbankan proses berpikir mendalam yang penting untuk pembelajaran bermakna.

**d) Bukti Reliabilitas dan Validitas**

Secara psikometrik, sistem AI menunjukkan potensi yang kuat namun dengan catatan. Sistem penilaian esai mencapai kesepakatan antar-penilai yang tinggi (5,10). *Cebeci Test of Creativity* yang dikomputerisasi menunjukkan reliabilitas yang kuat (ω = 0.833 dan 0.872) dan invariansi pengukuran antar tingkat kelas (19). Namun, dalam bidang klinis, reliabilitas absolut penilaian berbasis AI (0.59) dilaporkan lebih rendah dibandingkan penilaian berbasis ahli (0.65), mengindikasikan bahwa reliabilitas sangat bergantung pada domain dan kompleksitas tugas (12).

Heterogenitas kinerja ini menunjukkan bahwa AI bukanlah solusi generik untuk asesmen pendidikan. Efektivitasnya sangat kontekstual. "Paradoks beban kognitif-kualitas" merupakan peringatan penting bahwa integrasi AI harus dirancang untuk meningkatkan, bukan menggantikan, keterlibatan kognitif siswa. Asesmen formatif yang memanfaatkan AI perlu memasukkan "titik gesekan" yang disengaja untuk mendorong pemrosesan yang lebih dalam. Temuan ini mendukung model di mana AI digunakan untuk tugas-tugas yang secara mekanis dapat diotomasi dengan baik, sementara manusia berfokus pada penilaian tingkat tinggi yang membutuhkan interpretasi, empati, dan pemahaman kontekstual.

**3) Implikasi Etika, Bias Algoritmik, dan Keadilan**

Tinjauan ini mengungkap kekhawatiran etika yang mendalam dan meluas terkait penggunaan AI dalam asesmen, dengan bias algoritmik sebagai isu sentral.

**a) Ancaman terhadap Integritas Akademik**

Kekhawatiran paling langsung yang diidentifikasi adalah ancaman terhadap integritas akademik. Pendidik di berbagai studi menyatakan kecemasan signifikan mengenai penggunaan konten yang dihasilkan AI oleh siswa untuk menyesatkan penilaian (17,18,21). Kemudahan akses ke alat seperti ChatGPT telah menimbulkan dilema antara manfaat pedagogis potensial dan risiko terhadap keaslian karya siswa.

**b) Manifestasi Bias Algoritmik dan Prinsip Anna Karenina**

Bias algoritmik diidentifikasi dalam beberapa bentuk: pertama, bias demografis dan budaya. Utamanya, terkait dengan ketidakrepresentatifan data pelatih. AI yang dilatih terutama pada data berbahasa Inggris atau dari konteks tertentu mungkin tidak berlaku adil untuk siswa dari latar belakang linguistik atau budaya berbeda (10).

Kedua, bias sosioekonomi. Siswa dari distrik yang lebih kaya mungkin memiliki keunggulan dalam asesmen terkomputerisasi karena akses dan literasi digital yang lebih baik, berpotensi memperparah kesenjangan yang ada.

Ketiga, "prinsip Anna Karenina" dalam umpan balik otomatis. Temuan penting dari Schleifer et al., (7) menunjukkan pola bias sistematis. Embedding dari LLM menangkap respons berkualitas tinggi dengan baik, tetapi gagal membedakan antara berbagai jenis respons yang salah. Saat jawaban siswa semakin menyimpang dari pemahaman yang benar, jawaban-jawaban tersebut menjadi semakin beragam, sehingga kurang mirip satu sama lain dalam ruang *embedding*. Akibatnya, sistem umpan balik otomatis memberikan informasi diagnostik yang semakin generik atau tidak akurat kepada pelajar yang sedang berjuang, tepat ketika mereka paling membutuhkan umpan balik yang berkualitas dan spesifik. Ini menciptakan "Efek Matius" dalam asesmen formatif, di mana yang kaya (dalam hal pengetahuan) semakin kaya (mendapat umpan balik baik), dan yang miskin semakin miskin.

**c) Kekhawatiran atas Transparansi, Privasi, dan Agen Manusia**

Isu transparansi dan keterjelasan (*explainability*) muncul dalam beberapa studi. Analisis kompleks AI tidak selalu dapat memberikan penjelasan yang dapat diinterpretasi untuk prediksinya, yang menantang akuntabilitas dan kepercayaan (8,10). Kekhawatiran privasi data juga disebutkan, mengingat sistem AI sering memproses data sensitif siswa. Selain itu, terdapat kekhawatiran tentang erosi agensi manusia dan keahlian profesional guru jika terjadi ketergantungan berlebihan pada sistem otomatis.

Implikasi etika dari temuan ini sangat dalam. Bias bukanlah *bug* teknis semata, melainkan fitur yang dapat tertanam dalam sistem jika data dan desainnya tidak dikritisi. "Prinsip Anna Karenina" secara khusus mengkhawatirkan karena mengungkap bagaimana bias dapat secara halus dan sistematis merugikan populasi yang sudah rentan, bertentangan langsung dengan

tujuan keadilan dalam pendidikan. Hal ini menuntut pendekatan yang lebih canggih daripada sekadar mengejar akurasi agregat. Upaya mitigasi harus secara eksplisit menguji dan merancang untuk kinerja yang adil di seluruh spektrum kemampuan siswa.

**4) Strategi Mitigasi dan Pengamanan yang Diusulkan**

Studi-studi yang ditinjau mengusulkan serangkaian strategi mitigasi yang saling melengkapi, menekankan bahwa pendekatan teknis saja tidak cukup. Ada tiga strategi mitigasi yang diusulkan, yakni:

**a) Solusi Teknis:** mencakup penerapan algoritma sadar keadilan (*fairness-aware*) yang dirancang untuk mendeteksi dan mengurangi bias (7,10), penggunaan ai yang dapat dijelaskan (XAI) dengan teknik seperti SHAP untuk membuat keputusan model lebih dapat dipahami (10), serta proses pelatihan yang dikendalikan dan audit bias secara berkala (10,22).

**b) Pengamanan Prosedural dan Pengawasan Manusia:** melibatkan peran sentral manusia dalam *loop* (*human-in-the-loop*) untuk pengawasan dan keputusan akhir dalam asesmen berisiko tinggi (8,12), pendekatan "baca ganda" (*double read*) yang menggabungkan penilaian AI dengan ahli (12), dan penggunaan umpan balik pemilkiran berantai (*chain-of-thought*) dalam perancangan *prompt* untuk meningkatkan kualitas umpan balik (7).

**c) Kerangka Kebijakan, Tata Kelola, dan Pengembangan Profesional:** meliputi pengembangan kebijakan dan pedoman institusional yang kuat (17,22), pengembangan profesional yang ditargetkan untuk pelatihan pendidik dalam rekayasa *prompt* dan kompetensi digital (2,14,16), serta keterlibatan pemangku kepentingan dalam proses desain dan implementasi (14,15).

Strategi-strategi mitigasi yang diusulkan tersebut, menggambarkan pergeseran dari melihat etika sebagai "*add-on*" menuju integrasi etika-*by-design*. Pendekatan berlapis—teknis, prosedural, kebijakan, dan pedagogis—diakui sebagai kebutuhan. Temuan ini menunjukkan bahwa kepercayaan pada AI dalam asesmen tidak diperoleh melalui performa teknis semata, tetapi melalui ekosistem tanggung jawab yang transparan, dapat diaudit, dan selalu melibatkan keahlian manusia.

**5) Tantangan dalam Implementasi**

Implementasi sistem asesmen berbasis AI menghadapi banyak rintangan di berbagai level, diantaranya:

**a) Tantangan Teknis:** mencakup isu kualitas dan kuantitas data, dimana kinerja AI bergantung pada data pelatihan yang besar dan berkualitas tinggi (10,23), kesulitan dalam integrasi sistem dan skalabilitas dengan infrastruktur yang ada (8), serta keterbatasan pemahaman kontekstual yang dapat menyebabkan saran yang salah dari alat umpan balik otomatis (23).

**b) Tantangan Organisasional dan Sumber Daya:** meliputi resistensi terhadap perubahan dari institusi dan pendidik (17), keterbatasan pelatihan dan sumber daya seperti waktu dan anggaran (9,14), dan risiko memperburuk kesenjangan digital di antara siswa (13).

**c) Tantangan Penerimaan Pengguna dan Kepercayaan:** muncul dari kekhawatiran atas keandalan dan akurasi sistem AI (8,23) serta kecemasan akan integritas akademik (4,18).

Tantangan-tantangan ini saling terkait dan menunjukkan bahwa keberhasilan teknis tidak menjamin adopsi yang efektif. Faktor manusia dan organisasi sering kali menjadi penghalang yang lebih signifikan daripada keterbatasan algoritma itu sendiri. Implementasi yang sukses membutuhkan rencana perubahan yang holistik yang mengatasi kebutuhan pelatihan, menyediakan dukungan teknis berkelanjutan, membangun kepercayaan melalui transparansi, dan secara proaktif mengatasi masalah kesenjangan digital.

**6) Dampak terhadap Pemangku Kepentingan**

AI dalam asesmen memiliki dampak berbeda-beda pada siswa, pendidik, dan institusi, misalnya:

**a) Dampak pada Siswa:** memiliki potensi positif berupa pengalaman belajar yang dipersonalisasi dan umpan balik segera (6,15,24), namun juga membawa risiko dan dampak negatif seperti "paradoks beban kognitif" (11), bias dalam umpan balik (7), serta potensi atrofi keterampilan menulis dan berpikir kritis.

**b) Dampak pada Pendidik (Guru/Dosen):** meliputi pengurangan beban kerja administratif

melalui otomatisasi (23,25), tetapi juga menuntut perubahan peran dan kebutuhan pengembangan profesional yang signifikan seiring pergeseran peran guru menjadi fasilitator (2,14). Kesiapan dan persepsi guru sangat mempengaruhi adopsi AI (9).

**c) Dampak pada Institusi Pendidikan:** mencakup potensi peningkatan efisiensi dan jaminan kualitas (26,27), tetapi juga membawa tantangan kebijakan dan tata kelola untuk mengembangkan kebijakan etis dan memastikan kesetaraan akses (17,22).

Dampak pada pemangku kepentingan bersifat paradoksal. AI berjanji memberdayakan guru dan mempersonalisasi pembelajaran, tetapi juga berisiko mendisrupsi praktik pedagogis tradisional dan memperburuk ketidaksetaraan jika tidak dikelola dengan hati-hati. Transformasi yang sukses bergantung pada apakah institusi dapat memanfaatkan efisiensi AI untuk membebaskan sumber daya manusia yang kemudian diinvestasikan kembali dalam hubungan pedagogis yang lebih dalam dan dukungan yang ditargetkan, alih-alih melihatnya sebagai pengganti tenaga kerja semata.

**7) Peran AI dalam Asesmen Adaptif dan Personalisasi**

AI memungkinkan terobosan signifikan dalam asesmen adaptif dan personalisasi dengan menganalisis data kinerja siswa secara *real-time* dan menyesuaikan kesulitan konten serta jalur pembelajaran. Fitur-fitur tersebut termasuk pembuatan jalur pembelajaran personal menggunakan *deep learning* dan NLP untuk merekomendasikan konten yang disesuaikan (6,13), pemberian umpan balik dan motivasi berkelanjutan yang dimediasi oleh persepsi kegunaan dan kepuasan belajar (24), serta ketersediaan 24/7 dan adaptasi real-time untuk pelatihan dan umpan balik personal (15,28).

Personalisasi berbasis AI menawarkan jalan keluar dari model "satu ukuran untuk semua". Namun, temuan review mengingatkan bahwa personalisasi tidak boleh hanya didorong oleh algoritma efisiensi. Personalisasi harus tetap berpusat pada pedagogi, memastikan bahwa adaptasi konten memang mendorong pemahaman yang lebih dalam dan tidak sekadar mempermudah jalan menuju penyelesaian tugas. Selain itu, sistem adaptif harus dirancang dengan hati-hati untuk menghindari "*filter bubble*" pendidikan, di mana siswa hanya terpapar pada konten di tingkat mereka saat ini tanpa dorongan yang cukup untuk mencapai pemahaman yang lebih tinggi.

## 3. KESIMPULAN DAN SARAN

Sintesis dari 28 studi ini memberikan jawaban yang bernuansa terhadap pertanyaan penelitian. AI dimanfaatkan secara luas dan semakin canggih dalam asesmen pendidikan, menunjukkan efektivitas tinggi pada tugas terstruktur namun dengan kinerja yang bervariasi dan risiko etika yang signifikan pada tugas kompleks.

Namun, beberapa keterbatasan dalam literatur yang ada patut dicatat. Banyak studi memiliki sampel yang relatif kecil dan homogen, membatasi generalisasi temuan –mis., (10,23). Sejumlah penelitian bersifat eksploratori atau proof-of-concept, dan kurangnya studi longitudinal membatasi pemahaman kita tentang dampak jangka panjang AI pada hasil belajar dan praktik mengajar. Selain itu, sebagian besar penelitian berfokus pada pendidikan tinggi, menunjukkan kebutuhan akan lebih banyak penelitian di konteks K-12.

Beberapa implikasi terhadap penelitian di masa mendatang. Bagi penelitian longitudinal diperlukan studi yang melacak dampak penggunaan AI dalam asesmen selama periode waktu yang lebih panjang terhadap hasil belajar, motivasi, dan keterampilan meta-kognitif siswa. Sementara penelitian yang berfokus pada keadilan dan bias, harus secara eksplisit menyelidiki dan mengembangkan metode untuk mengurangi bias seperti "prinsip Anna Karenina", dengan menggunakan dataset yang lebih beragam dan representatif.

Pada bidang studi desain dan pedagogi, penelitian tentang bagaimana merancang prompt, rubrik, dan aktivitas pembelajaran sangat diperlukan. Khususnya, bagaimana mengintegrasikan AI secara optimal untuk meningkatkan (bukan mengurangi) keterlibatan kognitif mendalam. Adapun untuk penelitian kebijakan dan implementasi, lebih dibutuhkan eksplorasi tentang model tata kelola, kerangka etika, dan strategi pengembangan profesional yang paling efektif untuk mendukung implementasi yang adil dan bertanggung jawab.

Di sisi lain, praktik dan kebijakan pendidikan juga terdampak. Kebijakan institusional harus menekankan bahwa AI adalah alat untuk meningkatkan keahlian profesional pendidik, bukan

menggantikannya. Pengawasan manusia harus tetap menjadi batu penjuru untuk penilaian berisiko tinggi dan untuk populasi siswa yang rentan. Alokasi sumber daya juga harus diimbangi antara pembelian teknologi dan pelatihan guru yang komprehensif dalam literasi AI, rekayasa prompt, dan evaluasi kritis terhadap keluaran AI. Sementara pengembang sistem dan institusi harus menerapkan standar untuk transparansi algoritma dan audit bias rutin, memastikan bahwa sistem dapat dipertanggungjawabkan.

Desain sistem asesmen AI harus memprioritaskan kesetaraan sejak awal, dengan pengujian ketat pada berbagai subkelompok siswa untuk mengidentifikasi dan memperbaiki bias sebelum penerapan skala penuh. Berdasarkan bukti, AI paling tepat untuk (a) penilaian tugas terstruktur dengan rubrik jelas, (b) latihan dan umpan balik formatif adaptif dalam domain terdefinisi baik, dan (c) tugas berorientasi efisiensi. AI harus digunakan dengan kehati-hatian dan pengawasan manusia yang substansial untuk (a) keputusan sumatif berisiko tinggi, (b) penilaian penalaran kompleks, dan (c) pemberian umpan balik kepada pelajar yang mengalami kesulitan.

Integrasi AI dalam asesmen pendidikan berada pada persimpangan jalan yang penuh dengan janji dan bahaya. Realisasi potensinya untuk memberdayakan guru dan mempersonalisasi pembelajaran bergantung pada kapasitas kita untuk mengarahkan perkembangannya dengan kesadaran etis yang mendalam, komitmen pada keadilan, dan pengakuan yang tidak tergoyahkan bahwa pendidikan pada intinya adalah usaha manusia. Masa depan bukanlah tentang guru versus mesin, tetapi tentang guru yang dilengkapi dengan mesin, bekerja sama untuk menumbuhkan pemikiran kritis, kreativitas, dan potensi setiap siswa.

**REFERENSI**

1. Artanto H, Arifin F. Emotions and gesture recognition using affective computing assessment with deep learning. IJ-AI. 2023 Sept 1;12(3):1419.
2. Karademir T, Altan EB. Comparative Analysis of AI and Expert Evaluations in Engineering Design Pedagogy. Plos One. 2025;
3. Ganne Y, Leddo J. A Comparison of the Relative Effectiveness of AI vs Humans in Debugging Computer Code. International Journal of Social Science and Economic Research. 2024;
4. Sallam M, Al-Salahat K. Below Average ChatGPT Performance in Medical Microbiology Exam Compared to University Students. Frontiers in Education. 2023;
5. Cho M, Huang J, Kwon O. Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading. ETRI Journal. 2024 Feb;46(1):82–95.
6. Naseer F, Khan MN, Tahir M, Addas A, Aejaz SMH. Integrating deep learning techniques for personalized learning pathways in higher education. Heliyon [Internet]. 2024 June 1;10. Available from: https://www.sciencedirect.com/science/article/pii/S2405844024086596
7. Gurin Schleifer A, Beigman Klebanov B, Alexandron G. Uncovering Measurement Biases in LLM Embedding Spaces: The Anna Karenina Principle and Its Implications for Automated Feedback. Int J Artif Intell Educ [Internet]. 2025 May 30; Available from: http://dx.doi.org/10.1007/s40593-025-00485-7
8. Sabqat M, Khan RA, Jawaid M, Sajjad M. Artificial Intelligence Meets Item Analysis (AI meets IA): A Study of Chatbot Training and Performance in detecting and correcting MCQ Flaws. Pak J Med Sci. 2025 Feb 20;41(3):652–6.
9. Ibrahim AW, Taura AA, Iliyasu A, Shogbesan YO, Lukman SA. Artificial Intelligence (AI): Perception and Utilization of AI Technologies in Educational Assessment in Nigerian Universities. Edukasiana Jurnal Inovasi Pendidikan. 2024;
10. Martin PP, Kranz D, Wulff P, Graulich N. Exploring New Depths: Applying Machine Learning for the Analysis of Student Argumentation in Chemistry. Journal of Research in Science Teaching. 2023;
11. Stadler M, Bannert M, Sailer M. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. Comput Hum Behav. 2024 Nov 1;160:108386.
12. Johnsson V, Søndergaard MB, Kulasegaram K, Sundberg K, Tiblad E, Herling L, et al. Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. Medical Education. 2023 Aug 24;58(1):105–17.
13. Anuyahong B, Rattanapong C, Patcha I. Analyzing the Impact of Artificial Intelligence in Personalized Learning and Adaptive Assessment in Higher Education. International Journal of Research and Scientific Innovation. 2023;
14. Wardani S. Analysis of Decision Support System for Character Assessment of Elementary

School Students to Improve Teacher Assessment. International Journal of Information and Education Technology. 2024;
15. Darvishi A, Khosravi H, Sadiq S, Gašević D. Incorporating Ai and Learning Analytics to Build Trustworthy Peer Assessment Systems. British Journal of Educational Technology. 2022;
16. Allehyani SH, Algamdi MA. Digital Competences: Early Childhood Teachers' Beliefs and Perceptions of ChatGPT Application in Teaching English as a Second Language (ESL). International Journal of Learning Teaching and Educational Research. 2023;
17. Khlaif ZN, Ayyoub A, Hamamra B, Bensalem E, Mitwally MAA, Ayyoub A, et al. University Teachers' Views on the Adoption and Integration of Generative AI Tools for Student Assessment in Higher Education. Education Sciences. 2024;14(10):1090.
18. Titko J, Steinbergs K, Achieng M, Užule K. Artificial Intelligence for Education and Research: Pilot Study on Perception of Academic Staff. Virtual Economics. 2023;
19. Cebeci SM, Acar S. Development and Validation of the Cebeci Test of Creativity: A Computerized Test of Figural Creativity. Journal of Creative Behavior [Internet]. 2025 June;59(2). Available from: http://dx.doi.org/10.1002/jocb.70030
20. Guzmán E, Millan ER. Evaluating the Performance of Copula-Based Item Response Theory Models for Interpretable Assessment. IEEE International Conference on Consumer Electronics. 2024;
21. Kerneža M, Zemljak D. Science Teachers' Approach to Contemporary Assessment With a Reading Literacy Emphasis. Journal of Baltic Science Education. 2023;
22. Lim T, Gottipati S, Cheong MLF. What Students Really Think: Unpacking AI Ethics in Educational Assessments Through a Triadic Framework. International Journal of Educational Technology in Higher Education. 2025;
23. Rahimi M, Fathi J, Zou D. Exploring the impact of automated written corrective feedback on the academic writing skills of EFL learners: An activity theory perspective. Educ Inf Technol. 2024 July 30;30:2691–735.
24. Ji H, Suo L, Hua C. AI Performance Assessment in Blended Learning: Mechanisms and Effects on Students' Continuous Learning Motivation. Frontiers in Psychology. 2024;
25. Sá P, Bryda G, Parton N, Saúde S, Barros JP, Almeida I, et al. Impacts of Generative Artificial Intelligence in Higher Education: Research Trends and Students' Perceptions. 2024;
26. Cingillioglu I, Gal U, Prokhorov A. AI-experiments in education: An AI-driven randomized controlled trial for higher education research. Educ Inf Technol. 2024 Mar 26;29:19649–77.
27. Khan ABF, Raja AS. Evaluating Online Learning Adaptability in Students Using Machine Learning-Based Techniques: A Novel Analytical Approach. Education Science and Management. 2024;
28. Tapalova O, Zhiyenbayeva N. Artificial Intelligence in Education: AIEd for Personalised Learning Pathways. The Electronic Journal of E-Learning. 2022;