frontiers in Education

TYPE Systematic Review
PUBLISHED 26 August 2026
DOI 10.3389/feduc.2026.1905889

OPEN ACCESS

EDITED BY
Mohammed Borhandden Musah,
Emirates College for Advanced Education, United Arab Emirates

REVIEWED BY
Sarita Peddi,
SR University, India
Mohammad Bitar,
VIT-AP University, India

*CORRESPONDENCE
Walid Salamah
waleed.salamah@najah.edu

RECEIVED 10 June 2026
REVISED 01 August 2026
ACCEPTED 06 August 2026
PUBLISHED 26 August 2026

CITATION
Salamah W (2026) Ethical implications of artificial intelligence-based academic integrity systems in Palestinian higher education: a systematic review.
Front. Educ. 11:1905889.
doi: 10.3389/feduc.2026.1905889

Ethical implications of artificial intelligence-based academic integrity systems in Palestinian higher education: a systematic review

Walid Salamah*

College of Graduate Studies, An-Najah National University, Nablus, Palestine

An examination of ethical concerns is necessary for analyzing AI's use in detecting and deterring academic dishonesty in higher education. Although AI-powered systems (incorporating natural language processing, stylometry, and machine learning classifiers) could offer solutions to academic integrity concerns that would scale, such systems are vulnerable to algorithmic bias against non-native English speakers, constrained transparency in algorithmic decision-making, and privacy and academic freedom violations from increased surveillance. A systematic literature review was undertaken of English-language peer-reviewed literature over the period 2012–2024, identified across six databases (Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and PhilPapers) and hand-searched reference lists, drawing on theories of AI ethics, procedural justice, epistemic injustice, and critical algorithm studies. Four major themes emerged from the literature review, which included fairness and algorithmic bias, transparency and accountability and due process, student privacy and surveillance and agency, and context-dependent justice in under-resourced settings. No study in the corpus was based in, or focused specifically on, Palestinian higher education; the implications drawn for Palestinian universities are therefore theoretical extrapolations from evidence generated elsewhere, not a synthesis of local empirical findings. All current ethical guidelines were intended for resource-rich western institutions and disregarded the specific context in which universities face resource and digital inequality. Six principles of governance of ethical AI are provided, and an emphasis is placed on how AI is a supplement to human judgment in academia and should therefore be implemented with attention to the specific context.

KEYWORDS

academic integrity, AI ethics, algorithmic bias, contextual justice, epistemic injustice, governance, Palestinian higher education, surveillance

1 Introduction

1.1 Background and rationale

Following academic integrity means being honest and authentic. Recent technological developments, including online resources, cheating platforms, and large language models like ChatGPT, have led to challenges in the conventional approaches. Institutions of higher learning are progressively using AI tools such as machine learning and natural language processing to investigate academic dishonesty (Holmes et al., 2022; Zawacki-Richter et al., 2019), alongside longer-standing applications of learning analytics to student behavior more broadly (Siemens and Baker, 2012). Although these tools can be effective, they raise issues of transparency, privacy, bias, fairness, and other things. Focus of the existing literature is technical performance, and not ethics. This review contends that the adoption of AI-enabled solutions should be guided by justice, accountability, and human dignity. These issues are particularly pertinent in Palestinian higher education, where the universities operate in limitation of resources and infrastructure (Romani, 2009; UNESCO, 2021). No study located by the search below was based in, or focused specifically on, Palestinian institutions (see Sections 3.1 and 3.6); the absence of local studies on AI integrity systems is treated in this review as itself an instance of epistemic injustice, as the knowledge production is exclusive of Palestinian voices.

1.2 Theoretical framework

This review draws on four complementary theoretical traditions, introduced here for their analytical role in structuring the synthesis rather than as conclusions in themselves.

1.2.1 AI ethics: the AI4People framework

Floridi et al. (2018) propose five principles for ethical AI — beneficence, non-maleficence, autonomy, justice, and explicability — intended as an evaluative framework against which a given AI system or deployment can be assessed. In this review, these five principles serve as an evaluative checklist applied to the evidence synthesized in Section 3 and used explicitly in the appraisal in Section 4.1; they are not treated here as an *a priori* verdict on AI integrity systems.

1.2.2 Distributive justice: rawls

Rawls's (1971) theory of justice holds that inequalities in the distribution of benefits and burdens are justifiable only insofar as they work to the advantage of the least well-off. Applied to AI integrity systems, this principle provides an analytical question for the synthesis to address, rather than a predetermined answer: does the evidence show that the burdens of algorithmic error (e.g., false positive rates) fall disproportionately on already disadvantaged student groups, and if so, is this offset by a compensating benefit to those same groups? Section 3.3 evaluates this question against the included evidence.

1.2.3 Epistemic injustice: fricker

Fricker's (2007) epistemic injustice theory identifies injustice done against individuals as knowers, and applies both to the bias dimension and to the Palestinian context. Testimonial injustice occurs when a speaker suffers a reduction of credibility in virtue of some prejudice on the part of the audience due to the identity of the speaker — a concept this review uses to examine, in Section 3.3, whether integrity systems systematically downgrade the academic productions of writers who are not Anglophone. Hermeneutical injustice, the harm that arises when a community lacks the interpretive resources to make sense of its own experience, is used in Section 3.6 to frame the absence of governance models that account for the realities of resource-constrained educational institutions.

1.2.4 Critical algorithm studies

Critical algorithm studies (Crawford, 2021; Eubanks, 2018; Noble, 2018) treat AI systems as artifacts that can reproduce and, in some documented cases, worsen existing social inequalities through their design and training data, rather than as neutral technical tools. This review uses this tradition as an interpretive lens for the bias evidence reviewed in Section 3.3, and returns to it when evaluating the AI4People and Rawlsian criteria together in Section 4.1.

1.3 Research questions

This systematic review is organized around four ethically framed research questions:

RQ1: What ethical risks and harms are associated with AI-based academic integrity systems in higher education, and how are they distributed across student populations?

RQ2: What ethical frameworks and governance principles have been proposed to guide the responsible use of AI in academic integrity?

RQ3: To what extent do existing AI integrity frameworks adequately address the rights and interests of students in resource-constrained institutional contexts?

RQ4 (reframed — Comment 2): What ethical conditions would need to be met before AI-based integrity systems could be justifiably implemented in Palestinian universities specifically? Because no study located by this review examined a Palestinian institution (Sections 3.1, 3.6), RQ4 is not answered empirically by the systematic synthesis in the way RQ1–RQ3 are. It is instead a theoretically derived question, addressed in Section 3.6 and Section 4.5 by applying the frameworks and evidence synthesized under RQ1–RQ3 to the documented conditions of Palestinian higher education. This distinction is stated explicitly wherever RQ4 is addressed below.

2 Methods

2.1 Study design

A SLR was chosen as the most suitable design because the methodology allows for an objective, transparent and replicable synthesis of evidence from a diverse set of studies in a rapidly emerging field. The review was performed following the PRISMA 2020 reporting guideline (Page et al., 2021), and its overall design followed established systematic review methodology for management and social science research (Gough, Oliver, and Thomas, 2017; Tranfield, Denyer, and Smart, 2003). As no primary data was gathered, ethical approval was not necessary.

The search window was bounded at 2012, reflecting the emergence of large-scale learning-analytics and automated-integrity literature (Siemens and Baker, 2012), and closed at December 2024 to allow a stable corpus for full-text screening, appraisal, and synthesis ahead of submission. Because detector capabilities and generative-AI systems continue to change rapidly, this review is explicitly framed as a 2012–2024 evidence snapshot suited to an ethical and theoretical analysis, rather than a real-time technical benchmark of current detection tools; this boundary is revisited in Section 4.6.

2.2 PICOS framework and search strategy

The review was structured using the PICOS framework (Table 1), specifying the Population, Interventions, Comparators, Outcomes, and Study types included in the synthesis.

Six databases were searched: Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and PhilPapers, chosen to cover education, computer science, and philosophy/ethics literatures respectively. Searches were executed in December 2024 and combined three concept blocks with the Boolean AND operator, with synonyms within each block combined by OR: (1) AI methods ("machine learning" OR "natural language processing" OR "learning analytics" OR "text detector*"); (2) academic integrity ("plagiarism" OR "contract cheating" OR "academic misconduct" OR "academic integrity"); and (3) ethics and higher education ("bias" OR "fairness" OR "privacy" OR "surveillance" OR "accountability" OR "governance" OR "justice"), limited to peer-reviewed sources in English, January 2012 to December 2024. An illustrative full string, as run in Scopus, was: ("machine learning" OR "natural language processing" OR "learning analytics") AND ("plagiarism" OR "contract cheating" OR "academic misconduct") AND ("bias" OR "fairness" OR "privacy" OR "surveillance" OR "accountability" OR "governance" OR "justice"). Field tags, subject filters, and syntax were adapted to each database.

2.3 Study selection and data analysis

Inclusion required the studies to be: (1) published in a peer-reviewed journal; (2) published in the English language from January 2012 to December 2024; and (3) included a substantive and ethical discussion on the use of AI and academic integrity issues concerning fairness, bias, privacy, governance or student rights. Materials that were not peer reviewed, that addressed K-12 education only, or that did not substantively engage with ethics were excluded.

Database and hand searching identified 458 records. After 46 duplicates were removed, 412 unique records were screened by title, of which 316 were excluded as clearly irrelevant. The remaining 96 records were screened by abstract against the PICOS criteria; 12 were excluded at this stage, leaving 84 records for full-text retrieval and assessment. At full-text screening, 51 of the 84 records were excluded: not peer-reviewed (18), education level did not match the inclusion criteria (22), and full text not accessible (11). This left 33 studies for inclusion in the synthesis (Figure 1).

Quality appraisal used the Mixed Methods Appraisal Tool (MMAT), chosen because it accommodates the quantitative, qualitative, mixed-methods, and conceptual/theoretical study designs included in this review. MMAT criteria were applied to all 33 included studies; conceptual and theoretical works (e.g., philosophical analyses drawing on Fricker, Rawls, or Floridi et al.) were appraised using the tool's screening questions and a coherence-and-argumentation criterion adapted for non-empirical work, consistent with guidance for ethically focused reviews (Gough et al., 2017). Across the 33 studies, the proportion of applicable MMAT criteria met ranged from 60% to 100%, with a mean of 78%, indicating moderate-to-high overall methodological quality; no study was excluded on quality grounds alone, consistent with a synthesis-informing rather than a synthesis-gating use of appraisal (as applied in comparable qualitative evidence syntheses). Risk of bias at the review level, rather than the individual-study level,

TABLE 1 PICOS framework used to structure the search strategy and eligibility criteria.

Element	Criterion	Application In This Review
Population	Higher education institutions and students	Universities globally, with specific contextual focus on Palestinian higher education institutions
Intervention	AI-based academic integrity tools and systems	Machine learning classifiers, NLP tools, stylometric analyzers, learning analytics platforms, AI text detectors, and generative AI detection systems
Comparators	Not applicable as a defined comparator arm	This synthesis is not a comparative effectiveness review; where relevant, ethical implications are contrasted across AI tool types, student language backgrounds, and institutional resource levels rather than against a no-AI or alternative-tool baseline
Outcomes	Ethical implications, governance, fairness, rights	Algorithmic bias; false positive rates; data privacy; surveillance effects; due process; institutional governance; contextual adequacy; student rights
Study types	Peer-reviewed empirical, conceptual, and review studies (systematic search only — see Comment 3 note in 2.1 for supplementary theoretical/legal sources)	Quantitative, qualitative, mixed-methods, systematic reviews, and conceptual frameworks published in peer-reviewed journals

PICOS, population, interventions, comparators, outcomes, study types. The Comparators element is not operationalized as a comparison condition; it is retained in the table for methodological transparency and completeness of PICOS reporting.

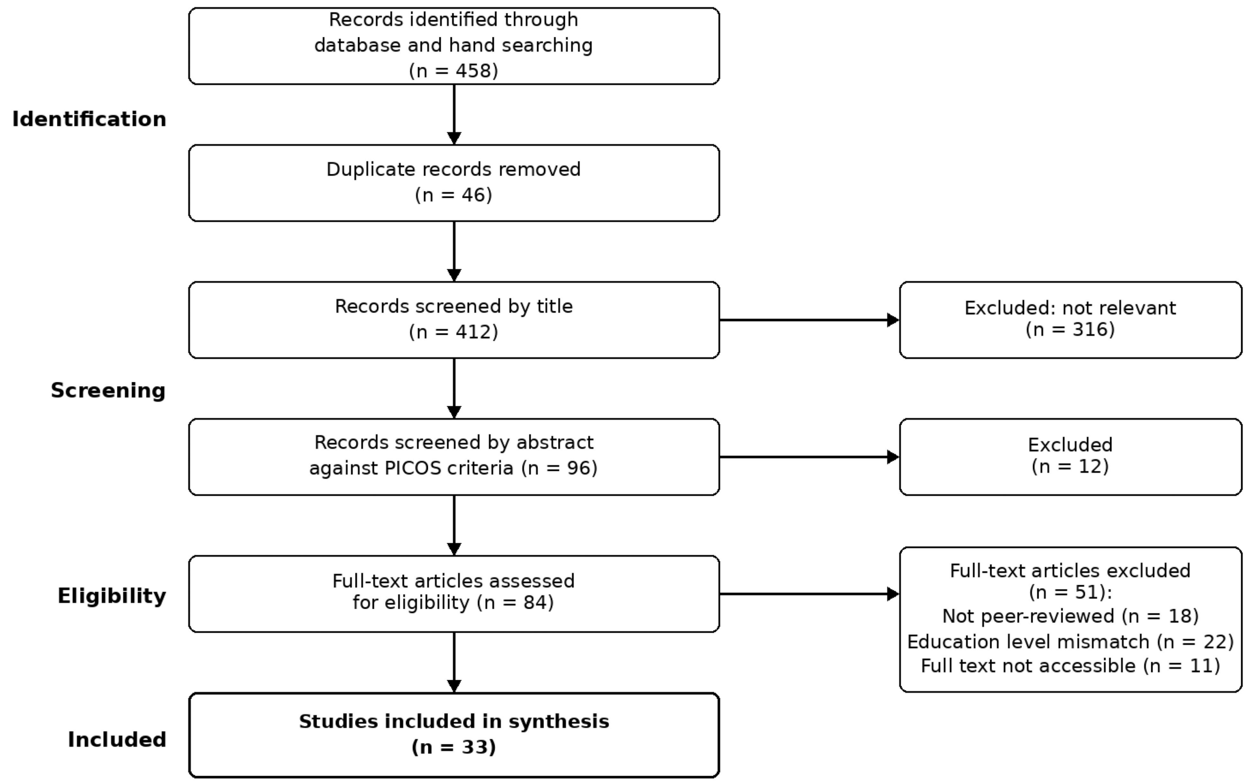

FIGURE 1
PRISMA 2020 flow diagram of study identification, screening, and inclusion.

is addressed in the Limitations (Section 4.6), including the single-reviewer constraint described below.

Using the three-stage process presented by Thomas and Harden (2008), informed by earlier foundational work on thematic analysis in qualitative research (Braun and Clarke, 2006), identified codes were created by undertaking line-by-line coding of the findings of the studies. Codes were grouped to generate descriptive themes. Finally, an interpretive synthesis was undertaken to produce analytical themes that were not present in any particular study. The thematic structure was tested against the four theoretical lenses described in Section 1.2. This review was conducted by a single author, who performed the search, screening, extraction, appraisal, and synthesis; no second reviewer independently screened or appraised the corpus. This constraint, including its implications for inter-rater reliability, is discussed in Section 4.6.

3 Results

3.1 Characteristics of included studies

A total of 33 articles fit the inclusion criteria. 27 of these studies (82%) were published from 2018 through 2024, which indicates both the speedy development of AI in education, and a growth in ethical analysis.

Three of the 33 studies, all based in North America or Europe, made incidental reference to Middle Eastern institutions without focusing on them; no study was based in, or focused primarily on, the Middle East, Africa, or South Asia, and no study considered Palestinian institutions specifically. The omission of such studies represents an ethical dimension of the field, discussed further in Section 3.6.

3.2 Overview of ethical themes

The synthesis identified four main ethical themes: (1) Fairness and algorithmic bias; (2) Transparency, accountability, and due process; (3) Privacy, surveillance, and student autonomy; and (4) Contextual justice in resource-constrained higher education settings. Each theme is elaborated below.

3.3 Theme 1: fairness and algorithmic bias

3.3.1 Discriminatory false positives

The concern most often documented in literature is that integrity systems powered by AI produce false positives that are discriminatory. As found by Liang et al. (2023), commercially available GPT detectors misclassified more than half of TOEFL essays written by non-native English speakers as AI-generated, while classification accuracy on native-English essays was close to perfect. Similarly, Williamson and Eynon (2020) observe that plagiarism detectors more broadly output greater rates of false positives among students whose writing shows features of non-anglophone pedagogies or second language sentence patterns at the sentence level.

This pattern is not a small technical flaw that can be adjusted. According to Noble (2018), the normative assumptions and representational hierarchies behind the design of AI classification systems get encoded within the systems. AI integrity systems that are trained primarily on anglophone academic writing will put in their models a culturally specific concept of legitimate scholarship. As a result, it penalizes students whose writing is different from it.

3.3.2 Authorship verification and stylometric bias

Ethical issues arise from authorship verification systems that create unique writing profiles to spot suspected contract cheating. According to Stamatatos (2018), while these systems may demonstrate considerable accuracy under ideal conditions with many authentic writing samples, they are consistently under-performing on students whose writing styles reflect non-linear development, differ linguistically from the training data, or are still developing their use of language.

It is possible that authentic variation of writing style across a degree could be regarded as suspicious, while on the other hand, natural stylistic variation across a diverse student population may set off false alarms. The most systematic independent evaluation of detector accuracy to date found substantial variation in detector performance and error patterns across tools, with none reliably accurate across all text types (Weber-Wulff et al., 2023). The review revealed that no study offered a validated approach capable of addressing stylometric bias in an educational setting populated by students from diverse linguistic backgrounds. This leads to the conclusion that the uncritical deployment of authorship verification systems in multilingual institutional contexts may be ethically premature.

3.3.3 Generative AI detection: emerging concerns

The rise of powerful LLMs like GPT-3 and GPT-4 brought new integrity challenges and triggered the development of unique detectors. Under controlled conditions, detectors can distinguish AI-generated from human-written text with reasonable accuracy; Gao et al. (2023) found that an AI-output detector and blinded human reviewers were both able to identify most ChatGPT-generated scientific abstracts, although a meaningful minority evaded detection by both. This baseline feasibility, however, does not resolve the fairness problem documented above: Perkins et al. (2023) demonstrated a great deal of disparity between commercially available detectors in terms of accuracy, while highlighting an especially poor detection performance with high false positives on non-native English writers, thereby reproducing the biases evident in more generalized plagiarism detection systems. As Kasneci et al. (2023) have indicated, for automatic detectors to maintain a sufficient level of reliability in the detection of AI-written text, such text would need to increase its structural resemblance to human writing. The researchers mention a sort of "arms race" between detectors and generative models which continuously improves on how it avoids detection after detection systems have been updated.

3.4 Theme 2: transparency, accountability, and Due process

3.4.1 Opacity and the accountability deficit

Another ethical issue relates to how little is known about how AI makes decisions and how that affects accountability. According to Floridi et al. (2018), explicability refers to an AI system's ability to explain how and why it has produced a given output. It is essential for ethical AI. There are many systems of integrity that provide similarity scores or classification outputs but do not specify which features were given importance, how these were compared with training data, or what other interpretations were explored (Crawford, 2021; Holmes et al., 2022). This lack of clarity causes what Crawford (2021) calls a structural accountability deficit; institutions cannot comply with their duty to answer to those whose conduct is judged and affected by technology. It is not possible for students to meaningfully challenge allegations whose evidential basis is concealed. Similarly, it is not possible for instructors to apply critical professional judgement to algorithmic outputs. Finally, external observers cannot verify that due process requirements are being met.

3.4.2 Procedural justice and student rights

Rawls (1971) defines two types of procedural justice: "perfect" where there is no possibility of the procedure yielding an unjust result, and "imperfect" where there is a desire to use the "best possible procedure" even though the procedure may yield unjust results. Academic integrity procedures cannot fulfill perfect procedural justice, as the full facts of a student's actions are likely unknowable, yet they can fulfill imperfect procedural justice and must therefore include the best possible procedure along with fair opportunities to participate, and reasoned decision-making. However, when AI outputs are used as dispositive rather than merely prima facie evidence in disciplinary procedures, it does not meet the criteria for imperfect procedural justice.

3.5 Theme 3: privacy, surveillance, and student autonomy

3.5.1 Data privacy and the ethics of collection

To uphold AI integrity systems, extensive student information must be gathered and stored for a long period of time. This includes writing samples, submission metadata, engagement logs, behavior patterns, and biometric data in remote proctoring deployments, i.e., face and environmental data. Remote proctoring in particular raises distinct ethical concerns beyond text-based integrity tools: Coghlan, Miller, and Paterson (2021) argue that online exam supervision technologies combine surveillance, algorithmic judgment, and high-stakes disciplinary consequences in ways that existing informed-consent practices rarely address, while Swauger (2020) documents how proctoring algorithms encode assumptions about "normal" bodies and behavior that disadvantage disabled, neurodivergent, and non-white students.

These concerns compound the data privacy issues addressed below and are especially salient in Theme 4, given the extent of remote and hybrid assessment in resource-constrained settings. According to Zuboff (2019), having the ability to collect the behavioral data that we generate (often in private) to predict our future actions is a major change in power relations between the institutions that collect the data and the people who the data is about. In the education sector, this power asymmetry is structurally embedded: students cannot decline to be monitored without risking an academic penalty, such as being marked absent or losing access to assessment. Williamson and Eynon (2020) and Floridi et al. (2018) both view data minimization as a central ethical requirement, a principle developed early in the learning analytics literature (Slade and Prinsloo, 2013), that institutions only collect data that is necessary to a specified, legitimate purpose. In places with strong data protection laws, such as the EU General Data Protection Regulation (GDPR), a right not to be subject to fully automated decisions with significant impacts is guaranteed in Article 22. The requirement is not sufficiently complied with even in those jurisdictions. In contexts where the data protection frameworks are weaker, or governance capacity is lacking, for example, many Palestinian universities, there are increased student privacy risks.

3.5.2 Surveillance culture and epistemic harm

Second-order risk that this ubiquitous AI surveillance engenders, according to Holmes et al. (2022) and Williamson and Eynon (2020) is the conditioning of students' behavior through the knowledge of that continuous surveillance. Zuboff (2019) argues that surveillance architectures do not only 'witness' and shape the subjects who produce the data. Students who are aware that their every utterance is subject to continuous AI scrutiny may be afraid to take intellectual risks-that is, the unconventional argument, non-standard form, or experimental writing style which might set off a false alarm. This chilling effect constitutes a harm to students in their capacity as knowers — the core injury Fricker (2007) identifies through her concepts of testimonial and hermeneutical injustice. This chilling effect undermines the academic pursuit itself, of which the promotion of the free and critical use of intelligence is a cardinal concern; universities do not promote that pursuit by creating learning environments in which the risk of a negative and punitive institutional response outweighs the incentives to write and reason freely and critically-especially when such surveillance systems are employed by institutions lacking student consent, oversight, and transparency.

4 Discussion

4.1 The primacy of ethics: why implementation must follow analysis

This review argues that AI-based academic integrity systems must not be implemented without prior ethical analysis. The literature consistently documents a pattern in which institutions adopt AI tools on the basis of their detection capabilities, with ethical concerns emerging reactively, after deployment has created institutional dependency and student vulnerability (Holmes et al., 2022; Zawacki-Richter et al., 2019). Assessed against Floridi et al.'s (2018) AI4People framework introduced in Section 1.2.1, current systems as documented in Section 3 fail across multiple foundational requirements: they generate documented discriminatory harms for identifiable student groups (Section 3.3), violating non-maleficence; they replace professional educator discretion with algorithmic judgement (Section 3.4), failing autonomy; they assign the consequences of error disproportionately across student populations (Section 3.3.1), distorting justice in the Rawlsian sense set out in Section 1.2.2; and they deploy opaque systems in consequential disciplinary processes (Section 3.4.1), failing explicability. Technologies that do not meet these fundamental criteria must not be adopted in their current forms.

4.2 Counterarguments and responses

A complete ethical analysis requires engagement with counterarguments. Three objections warrant serious consideration. First, proponents of AI integrity systems argue that the risks of insufficient integrity oversight—credential devaluation, disadvantage to honest students, and reputational damage to higher education—are themselves ethical harms that must be weighed against the risks of AI deployment (McCabe et al., 2012; Newton, 2018). This objection has genuine force. The appropriate response is not to reject AI integrity systems categorically but to insist that their deployment be subject to the preconditions outlined in the governance framework proposed below. The choice is not between flawed AI and no oversight, but between responsible and irresponsible adoption.

Second, some researchers hold that algorithmic bias, while real, is technically improvable, such that restricting AI use based on current limitations prevents access to future systems that might be substantially fairer (Kasneci et al., 2023). This argument is methodologically sound but temporally misapplied: it justifies the future use of improved systems, not the continued deployment of currently biased ones. In contexts such as Palestinian universities, where no validated, contextually appropriate systems exist and where the consequences of false accusation are particularly severe, the ethical case for restraint is especially strong.

Third, some argue that an imperfect AI integrity tool is preferable to no integrity infrastructure at all, particularly in resource-constrained settings. Dawson (2020) offers the most developed version of this position, treating assessment security as a foundational design requirement that institutions cannot responsibly defer. Rawls (1971) provides the appropriate response: that the least advantaged members of an institution must not be made to bear the costs of arrangements designed primarily for institutional convenience. The ethical path for Palestinian universities is not uncritical importation of externally developed tools, but active participation in the development and validation of contextually appropriate governance frameworks.

4.3 From detection to educational ethics: a paradigm shift

A significant and encouraging theme in the reviewed literature is the growing academic consensus—though not yet fully realized in practice—that punitive, detection-centered academic integrity models are both ethically problematic and less effective than educational alternatives. Bretag (2019), Eaton (2020), and Bertram Gallant (2017) argue that highly automated, punitive models that position students primarily as suspected transgressors generate distrustful learning communities without meaningfully reducing dishonest behavior. McCabe et al.'s (2012) landmark study of student behavior demonstrated that educational and relational models of integrity—grounded in authentic community values and genuine connection to the academic enterprise—consistently outperformed surveillance and punishment approaches.

From an ethical perspective, educational approaches are not only more effective but more defensible. Kant's (1785/1996) categorical imperative holds that people must be treated as ends in themselves, never merely as means. Integrity systems that treat students primarily as threats to be identified, rather than learners to be supported, cannot be said to uphold this principle. This ethical foundation also supports the constructive integration of AI writing assistants rather than their prohibition; treating AI as a pedagogical resource and developing student capacity for its appropriate use honors the educational function of universities rather than substituting surveillance for teaching (Chan, 2023).

4.4 Proposed ethical governance framework

Drawing on the reviewed literature and the four theoretical frameworks, this review proposes an ethical governance framework for AI in academic integrity consisting of six principles (Table 2), consistent in spirit with the ENAI Recommendations on the ethical use of AI in education (Foltynek et al., 2023) but organized around named ethical theory rather than practical guidance alone. This framework differs from existing proposals in three important respects: each principle is grounded in identified ethical theory rather than offered as practical guidance alone; contextual justice is treated as a core principle rather than an optional refinement; and the framework is intended as a precondition for deployment rather than as an aspirational standard to be approached retrospectively.

4.5 Contextual application: Palestinian higher education

Palestinian universities constitute a rich case study for contextual ethical analysis. Palestinian higher education comprises a mix of universities, university colleges, and community colleges; official enrollment statistics across these institution types are published by the Palestinian Ministry of Education and Higher Education in cooperation with the Palestinian Central Bureau of Statistics (PCBS and Ministry of Education and Higher Education, 2024). This review focuses on universities offering undergraduate and graduate degrees, which operate under conditions that include intermittent physical disruption and access limitations; uneven and unreliable internet connectivity; constrained access to licensed commercial software; limited institutional governance capacity; and a student population managing significant personal adversity alongside their academic studies (Fasheh, 2015; Romani, 2009; UNESCO, 2021). Each of the ethical concerns identified in Themes 1–3 is substantially amplified in this context.

Algorithmic bias challenges are intensified for Arabic-speaking writers using English as a second or third language—a writing profile vastly underrepresented in the training data of current integrity systems. Applying Fricker's (2007) concept of testimonial injustice, these students' academic authenticity is systematically accorded lower credibility by systems trained to recognize a normative form of academic writing that differs from their own. This is not a technical problem but a structural one that further entrenches educational disadvantage. The risk of false-positive outcomes is elevated in contexts where genuine institutional capacity for independent human review is constrained and where students may lack the social, legal, or institutional resources to challenge automated findings. Surveillance concerns are heightened in a population already subject to extensive external monitoring; AI academic surveillance implemented without explicit consent, clear data protection, and meaningful appeal rights represent an additional and disproportionate infringement of student autonomy (Zuboff, 2019). The complete absence of empirical studies examining AI integrity systems in Palestinian universities—no single study was identified in this review—is itself an instance of epistemic injustice: institutions that adopt these technologies do so without any local evidence base, compelled to apply frameworks developed under fundamentally different conditions.

The ethical argument against AI integrity systems in Palestinian universities is not categorical. In contexts where human reviewer capacity is limited, AI systems capable of identifying patterns across large student cohorts and directing human attention to cases warranting review could provide genuine preventative and remedial benefits (Baker and Inventado, 2014). The ethical question is not whether AI should be used, but under what conditions its use would be permissible within a contextual justice framework. These conditions include: validation of any AI system on a corpus representative of second-language English learners and Arabic speakers, or explicit documentation of known biases with compensatory human review protocols; rigorous data minimization and genuine informed consent; demonstrated institutional capacity for meaningful human review before AI outputs enter disciplinary processes; transparent student notification and substantive mechanisms for challenging AI findings; and active engagement of Palestinian scholars and educational leaders in the governance frameworks applied to their own institutional context.

For Palestinian universities, the governance framework proposed in Section 4.4 is both constraining and generative. Its constraints follow directly from the analysis above: the elevated risk of bias against Arabic-speaking writers; the surveillance

TABLE 2 Proposed ethical governance framework for AI-based academic integrity systems, organized by ethical principle.

Principle	Requirement	Status	Operational Indicator	Responsible Actor	Implementation Mechanism	Theoretical/ Legal Basis
1. Transparency & Explicability	AI tools must be explainable in nonspecialist terms; black-box outputs must not serve as disciplinary evidence	Ethical recommendation; legal requirement only within EU/EEA jurisdiction (EU AI Act Art. 13)	% of flagged cases with a documented, human-readable explanation on file	IT/academic integrity office	Vendor contract clause requiring explainability documentation; annual audit	Floridi et al. (2018); EU AI Act Art. 13
2. Human Oversight & Accountability	AI outputs as prompts for inquiry only; clear human accountability for all integrity decisions; no automated sanctions	Ethical recommendation	% of sanctions with a named, accountable human decision-maker on record	Academic integrity committee chair	Mandatory human sign-off workflow before any sanction is applied	Kant (1785/1996); Floridi et al. (2018)
3. Student Rights & Contestation	Meaningful mechanisms to challenge AI outputs; institution bears burden of proof	Ethical recommendation; legal requirement only within EU/EEA jurisdiction (GDPR Art. 22)	Median time to resolve a student appeal; appeal upheld rate	Student ombudsperson/ appeals board	Published appeals procedure with defined timeline	Rawls (1971); GDPR Art. 22
4. Data Minimization & Consent	Collect only necessary data; genuine informed consent; clear retention and sharing policies	Ethical recommendation; good practice regardless of jurisdiction	Data retention period; % of collected fields justified against a stated purpose	Data protection officer/registrar	Data protection impact assessment prior to tool adoption	Floridi et al. (2018); Zuboff (2019); Slade and Prinsloo (2013)
5. Equity Auditing	Regular bias audits disaggregated by language background, nationality, disability; discontinue biased tools	Ethical recommendation	False positive rate by student language background, reported annually	Institutional research office	Annual disaggregated audit with a public summary report	Eubanks (2018); Noble (2018); Rawls (1971)
6. Contextual Justice	Frameworks validated for specific institutional contexts before deployment; local communities as co-designers	Ethical recommendation	Evidence of local validation study and local stakeholder co-design prior to procurement	Provost/VP academic affairs	Local pilot and validation study required before institution-wide rollout	Fricker (2007); Tlostanova and Mignolo (2012); UNESCO (2021)

EU AI Act, European Union Artificial Intelligence Act (2024); GDPR, General Data Protection Regulation. Both instruments bind institutions operating within the EU/EEA; they are included here as normative reference points, not as requirements applicable to Palestinian universities, which fall outside their jurisdiction (see Section 4.5).

concerns in an already extensively monitored social environment; the complete absence of a local evidence base; and the limited institutional governance capacity to establish genuine human oversight. Adopting commercially available tools designed for resource-rich, anglophone institutional contexts without contextual adaptation would violate all six governance principles and constitute the kind of hermeneutical injustice Fricker (2007) describes.

The generative dimension follows from Palestinian institutions' established capacity for innovation under constraint and the potential for carefully implemented, educationally oriented AI tools to support limited institutional capacity in appropriate ways. Fasheh (2015), writing on Palestinian educational philosophy, has identified a distinctive capacity to develop locally grounded educational practices under conditions of adversity—a capacity that holds important lessons for constructing context-informed AI governance rather than merely importing externally constructed frameworks. The ethical path for Palestinian universities involves deliberate, evidence-based adoption processes that include institutional capacity-building in AI literacy and integrity pedagogy; the development of context-specific policies aligned with the six governance principles; collaborative research partnerships aimed at developing and evaluating Arabic-language validation corpora; emphasis on preventative and supportive rather than detection-and-punishment interventions; and active advocacy to ensure that Palestinian higher education is represented in global research agendas for AI in education. This path reflects UNESCO's

(2021) recommendations for the responsible integration of technology in educationally challenged settings and the epistemic decolonization efforts described by Tlostanova and Mignolo (2012) and Mbembe (2016).

4.6 Limitations

Several limitations of this review warrant acknowledgment. The restriction of the search to English-language publications is a significant limitation, and one in some tension with this review's own thesis: because the argument is precisely that anglophone framings crowd out other knowledge, an English-only search method risks reproducing in its methodology the injustice the review diagnoses in its object of study. This review was not able to conduct a supplementary search of Arabic-language regional indexes (e.g., e-Marefa, AskZad, Shamaa, Al Manhal) before submission. This is flagged here as an unresolved limitation rather than a resolved one: a documented absence of Arabic-language evidence, established through an actual supplementary search and reported even if the result is null, would be a substantially stronger empirical basis for this review's argument than the current assumed gap, and should be treated as a priority for any revision or follow-up study. The absence of prospective protocol registration means that some degree of *post-hoc* analytical decision-making cannot be ruled out as a potential source of bias. The search closed in December 2024 while this manuscript was finalized in mid-2026; detector tools and generative models have continued to change over that period, and findings regarding specific detection tools should be read as characteristic of the 2012–2024 evidence base rather than as a real-time technical assessment (see Section 2.1). Thematic synthesis involves interpretive judgements that cannot be fully standardized, even with systematic procedures. This review was conducted by a single author, who performed the search, screening, extraction, quality appraisal, and synthesis without a second independent reviewer; inter-rater reliability statistics could therefore not be calculated for either the screening or the appraisal stage, and the consistency of these judgements could not be externally verified. This is a genuine constraint on a systematic review of this kind and should be weighed accordingly by readers. Most significantly, no study included in this review directly examined Palestinian universities; the Palestinian material in Sections 3.6 and 4.5 is theoretical extrapolation from evidence and frameworks developed elsewhere, explicitly flagged as such at each point above, and should not be read as a systematic review finding about Palestinian institutions specifically. Future research should prioritize empirical investigation in Palestinian and comparable institutional contexts.

5 Conclusion

Artificial intelligence can genuinely support academic integrity. The systematic evidence reviewed here demonstrates that plagiarism detection, stylometric analysis, learning analytics, and AI writing tools have produced real advances that could, under appropriate conditions, help institutions maintain meaningful standards of academic honesty at scale. These capabilities are significant and should not be discounted.

The central conclusion of this review, however, is that those capabilities can only be ethically exercised within governance frameworks that treat fairness, transparency, accountability, privacy, and contextual justice as preconditions for deployment rather than aspirational goals. The ethical risks documented in this review—algorithmic bias against non-anglophone writers; opacity-driven procedural injustice; surveillance architectures that degrade the epistemological conditions of learning; and the contextual injustice of applying externally designed frameworks to institutions for which they were never validated—are not incidental features of current AI integrity systems. They are structural features that require structural responses grounded in the ethical theories that explain why they are wrong: the AI ethics principles of Floridi et al. (2018), the difference principle of Rawls (1971), the epistemic injustice framework of Fricker (2007), and the critical algorithm studies of Noble (2018) and Eubanks (2018).

For Palestinian universities, these vulnerabilities are specific and compounding, though—as stated throughout Sections 3.6 and 4.5—the specificity is inferred from evidence gathered elsewhere rather than demonstrated with local data. The complete absence of scholarship examining AI integrity in this institutional context is not merely an academic lacuna but an ethical failure—an instance of the hermeneutical injustice Fricker describes, in which a community is denied the interpretive resources needed to make sense of a challenge that directly affects its students' academic trajectories and its institutions' credibility.

This review makes three calls: to the research community, to establish an empirical research agenda focused on AI integrity systems in Palestinian and comparable resource-constrained institutional contexts as a matter of both academic and ethical priority; to institutions, to treat the governance framework proposed here as a precondition for AI adoption rather than a *post-hoc* aspiration; and to the global AI ethics community, to recognize that epistemic frameworks suited to well-resourced institutional environments are not universally applicable, and that genuine ethical governance of AI in education requires the meaningful participation of diverse institutions and communities across the full range of global higher education.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: No original dataset was generated for this study. The analysis was based on previously published literature identified through the review process. All sources analyzed are cited in the manuscript's reference list. Therefore, no repository link or accession number is applicable.

Author contributions

WS: Methodology, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baker, R. S., and Inventado, P. S. (2014). "Educational data mining and learning analytics," in *Learning Analytics*, eds. J. Larusson, and B. White (New York, NY: Springer), 61–75. doi: 10.1007/978-1-4614-3305-7_4

Bertram Gallant, T. (2017). Academic integrity as a teaching and learning issue: from theory to practice. *Theory. Pract.* 56 (2), 88–94. doi: 10.1080/00405841.2017.1308173

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3 (2), 77–101. doi: 10.1191/1478088706qp063oa

Bretag, T. (2019). "Academic integrity," in *Oxford Research Encyclopedia of Education*, (Oxford, UK: Oxford University Press), 910. doi: 10.1093/acrefore/9780190264093.013.910

Chan, R. Y. (2023). Generative AI and academic integrity: emerging challenges and institutional responses. *J. Univ. Teach. Learn. Pract.* 20 (2), 7. doi: 10.53761/1.20.02.07

Coghlan, S., Miller, T., and Paterson, J. (2021). Good proctor or "big brother"? ethics of online exam supervision technologies. *Philos. Technol.* 34 (4), 1581–1606. doi: 10.1007/s13347-021-00476-1

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Dawson, P. (2020). *Defending Assessment Security in a Digital World: Preventing E-cheating and Supporting Academic Integrity in Higher Education*. Routledge.

Eaton, S. E. (2020). Academic integrity in 2020: year in review. *Can. Perspect. Acad. Integr.* 3 (2). doi: 10.11575/cpai.v3i2.71636

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

European Parliament and Council (2024). *Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act)*. Brussels: Official Journal of the European Union.

Fasheh, M. (2015). Over 68 years of living and learning with mathematics. *For Learn. Math.* 35 (2), 2–7.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People: an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* 28 (4), 689–707. doi: 10.1007/s11023-018-9482-5

Foltynek, T., Bjelobaba, S., Glendinning, I., Khan, Z. R., Santos, R., Pavletic, P., et al. (2023). ENAI Recommendations on the ethical use of artificial intelligence in education. *Int. J. Educ. Integr.* 19, 12. doi: 10.1007/s40979-023-00133-4

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198237907.001.0001

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., et al. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digit. Med.* 6, 75. doi: 10.1038/s41746-023-00819-6

Gough, D., Oliver, S., and Thomas, J. (2017). *An introduction to Systematic Reviews*. 2nd edn. London: SAGE.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., et al. (2022). Ethics of AI in education: towards a community-wide framework. *Int. J. Artif. Intell. Educ.* 32 (3), 504–526. doi: 10.1007/s40593-021-00239-1

Kant, I. (1996). *Groundwork of the Metaphysics of Morals (M. Gregor, Trans. and Ed.)*. Cambridge, MA: Harvard University Press. (Original work published 1785).

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103, 102274. doi: 10.1016/j.lindif.2023.102274

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT Detectors are biased against non-native English writers. *Patterns* 4 (7), 100779. doi: 10.1016/j.patter.2023.100779

Mbembe, A. (2016). Decolonizing the university: new directions. *Arts Humanit. High. Educ.* 15 (1), 29–45. doi: 10.1177/1474022215618513

McCabe, D. L., Butterfield, K. D., and Treviño, L. K. (2012). *Cheating in College: Why Students do it and What Educators can do About it*. Baltimore, MD: Johns Hopkins University Press.

Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Front. Educ.* 3, 67. doi: 10.3389/feduc.2018.00067

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press. doi: 10.18574/nyu/9781479833641.001.0001

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Br. Med. J.* 372, n71. doi: 10.1136/bmj.n71

Palestinian Central Bureau of Statistics [PCBS] and Ministry of Education and Higher Education (2024). *Higher Education Statistical Yearbook, Scholastic Year 2022/2023*. PCBS.

Perkins, M., Roe, J., Postma, D., McGaughran, J., and Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: combining academic judgement and software to identify generative AI tool misuse. *J. Acad. Ethics* 22 (1), 89–113. doi: 10.1007/s10805-023-09492-6

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Romani, V. (2009). The politics of higher education in the Middle East: problems and prospects. *Middle East Brief* 36, 1–8.

Siemens, G., and Baker, R. S. (2012). "Learning analytics and educational data mining: towards communication and collaboration," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* eds. S. B. Shum, D. Gašević, and R. Ferguson (New York, NY: ACM), 252–254. doi: 10.1145/2330601.2330661

Slade, S., and Prinsloo, P. (2013). Learning analytics: ethical issues and dilemmas. *Am. Behav. Sci.* 57 (10), 1510–1529. doi: 10.1177/0002764213479366

Stamatatos, E. (2018). Masking topic-related information to enhance authorship attribution. *J. Assoc. Inf. Sci. Technol.* 69 (3), 461–473. doi: 10.1002/asi.23968

Swauger, S. (2020). *Our Bodies Encoded: Algorithmic Test Proctoring in Higher Education*. Hybrid Pedagogy.

Thomas, J., and Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC. Med. Res. Methodol.* 8, 45. doi: 10.1186/1471-2288-8-45

Tlostanova, M., and Mignolo, W. (2012). *Learning to Unlearn: Decolonial Reflections from Eurasia and the Americas*. Columbus, OH: Ohio State University Press.

Tranfield, D., Denyer, D., and Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *Br. J. Manag.* 14 (3), 207–222. doi: 10.1111/1467-8551.00375

UNESCO (2021). *Education in Emergencies: Palestine Overview*. Paris: UNESCO Publishing.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., et al. (2023). Testing of detection tools for AI-generated text. *Int. J. Educ. Integr.* 19 (1), 26. doi: 10.1007/s40979-023-00146-z

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

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: where are the educators? *Int. J. Educ. Technol. High. Educ.* 16, 39. doi: 10.1186/s41239-019-0171-0

Zuboff, S. (2019). *The age of Surveillance Capitalism: The Fight for a Human Future at the new Frontier of Power*. New York, NY: PublicAffairs.