FALSE POSITIVES, REAL CONSEQUENCES: RETHINKING AI DETECTION AND TRUST IN HIGHER EDUCATION

Rebat Kumar Dhakal
Kathmandu University, School of Education, Hattiban, Lalitpur, Nepal

Padam Raj Pant
Technology2050, Lalitpur, Nepal

Naba Raj Gautam
Counsel & Counsel Pvt Ltd, Kathmandu, Nepal

Abstract

Purpose – This study explores the growing tension between academic integrity, instructional trust, and the use of AI-detection tools in higher education. It aims to examine how these tools misclassify authentic human writing as AI-generated, overlook sophisticated AI-produced content, and unintentionally undermine the learning environment.

Design/methodology/approach – Using qualitative data obtained from two universities, this study explores the experiences of PhD students and their supervisors with regards to AI-based detection practices. Interviews were conducted to gain insights into their perceptions of accuracy, learning effects, and social-emotional effects of misattribution.

Findings – The study reveals that AI-detection systems produce inconsistent and unreliable results, often generating false positives. Both doctoral students and research supervisors reported frustration, distrust, and emotional distress linked to such inaccuracies. Students felt unfairly accused of misconduct, while supervisors expressed uncertainty about evaluating genuine learning. The findings highlight the paradox of detection technologies—tools meant to uphold integrity instead eroding confidence and transparency in academic relationships.

Research limitations/implications – The study focused on a limited sample of doctoral students and supervisors from two institutions; hence, generalization of findings to other contexts may not be appropriate. Nevertheless, they provide interesting insight into how emerging assessment technologies converge with academic cultures, posing a myriad of questions worthy of investigation.

Practical implications – There is a call for the use of AI-detection software as a resource to be consulted alongside other measures for making integrity judgments rather than as definitive evidence. Human oversight, dialogic feedback, and attention to students' writing processes—such as revision histories and drafting patterns—should guide integrity judgments.

Educators are encouraged to create transparent/ethical AI-use policies and promote environments where students feel secure to disclose their learning practices.

Originality/value – The current research is relevant to an ever-growing number of scholarly works in an emergent academic field known as Educational Ethics of AI, advocating for new detection practices and new approaches in higher education focusing on trust and ethical AI literacy.

Keywords**:** ethical use of AI, AI-detection, academic integrity, student well-being, doctoral supervision

Introduction

This article was started because of the above-discussed documented incident and is a critical analysis of present AI-detection software that is demonstrably unreliable and the unintended consequences it creates when put in practice. Its intentions are not to explain how to circumvent academic integrity rules, but to shed light on systemic issues that put excessive pressure on doctoral candidates and supervisors. This critique should be seen as constructive to enable more reliable, transparent, and equitable evaluation processes. In no way is it in support of claiming AI-generated work as your own. In sum, the central argument is that a narrow conception of errors (or defects) cannot be divorced from the larger, more radical scholarly enterprise: building genuine originality, critical thought and scholarly novelty through the cultivation of mentorship and strong pedagogical process. We propose a shift from a detection-focused, punitive model toward a holistic, process-oriented, and
trust-based framework for assessment in the age of AI—a framework sensitive to the Nepali context.

Scene Setting

During the early phase of this research, I worked with a doctoral student at a Nepali university on her research proposal with regard to AI detection. She expressed concern about the inconsistency between detection reports. By the time I reached the software that was used on the university's official system, I noticed that the level of AI content in her document was 0%, when the university's previous check on the same document showed 3%. The institution was to take such an uncompromising zero-tolerance approach to AI-generated content that the resulting anxiety was beyond her capabilities to endure. "I was falling apart," she admitted unsteadily. "I was consumed by fear. I could not sleep, could not concentrate, I felt totally powerless. How could I even try to prove my innocence when I didn't know where to begin?" she cried out.

Reluctant to give in, the student went on two revisions and re-submitted the same proposal. Each new check, though, produced different results, even with the same detection tool. While my copy of the document included 0% AI content, a second test my colleague ran in the university system discovered 2%. More baffling was that the earlier portions of it classified as "human-written" were also flagged as "AI-generated." The student, confused by these inconsistencies, felt a great *deal of stress and her supervisor also had a hard time giving reassurance in a university's stifling policy framework.*

Anecdote#1, Author

This case highlights a glaring shortcoming of AI-detection software in academic integrity enforcement—which is particularly serious under firm zero-tolerance policies. The doctoral student's experience also illustrates that these tools are not only inconsistent but fundamentally unreliable. Different checks yield variable and contradictory results for the same text. This instability in technology results in a genuine, well-meaning compliance process becoming a source of intense stress and unfair scrutiny. Students and supervisors can only go through their own tests in an arbitrary mode that ends up on their heads. The underlying problem goes beyond the student's distress to a larger ethical and procedural dilemma: institutions may be under threat of punishing the honest work of their members for flawed, non-transparent metrics, which can undermine trust in academic institutions and exert undue strain on scholars trying to navigate an emerging and imperfect technological landscape.

Background

Rapid proliferation of generative artificial intelligence (AI) tools, for example ChatGPT, Perplexity, Copilot, etc., also has created doubts regarding authorship, originality and academic integrity throughout the world in higher education across the globe (Duah & McGivern, 2024; Kotsis, 2025; Pellerin & Ogandaga, 2024). This has prompted universities to respond with a combination of policy changes, staff training and the purchase of AI-detection systems, often overlaid on existing plagiarism-detection tools such as Turnitin's AI detector, GPTZero, and Originality.ai. These tools are reportedly meant to distinguish between writing done by machines and those produced, providing them with an ostensibly neutral context in which misconduct may be detected. Yet there are increasing indications that these technologies are inherently flawed, yielding both false positives (human text flagged as AI) and false negatives (AI text deemed human) (Liang et al., 2023). Scholars have claimed AI detectors to be technically unreliable, misidentify non-native or stylistically dissimilar writers disproportionately and are very easily hidden behind more advanced AI uses (Fraser et al., 2025; Hale, 2025). This technical failure raises a fundamental ethical and pedagogical challenge that shifts assessment from a pedagogic way of learning evaluation to one based on suspicion and monitoring (Dwivedi et al., 2023). An unfounded accusation of AI-assisted misconduct in this realm can have devastating personal and career effects.

Nepali academia encounters unique challenges with its historical fusion of traditional *guru-shishya* (teacher-disciple) mentorship principles, an ever growing but resource-poor digital infrastructure, and a pervasive cultural respectful perception of authority (Sharma, 2019). Concerns over academic dishonesty and research misconduct have therefore increased within the past 10 years prompting bodies like the University Grants Commission and Nepal Health Research Council to elevate research integrity and plagiarism. Universities and other research institutions have already started to require checking for similarities in theses and academic publication, and editorials and commentaries in national outlets showcase the growing risks related to generative AI and "copy paste" cultures in academic writing (Subedi, 2024; Dahal, 2024). One domain, PhD journey, poses as a specific key area of vulnerability. It is a high-stakes domain where original contribution is of vital importance, supervisory relationships are deeply personal and a student's academic and professional future is at stake (Kamler & Thomson, 2014). But so far, the development of policy has primarily emphasized detection and sanction, with little consideration of how AI-detection tools are experienced by students and educators, or how they alter the affective and relational dimensions of supervision and assessment (Das et al., 2025; Kirsanov et al., 2026; Rentier, 2025). Adapting the local context of technological "solutions" developed from the Global North to the local context could create a new layer of complexity in the institutionalisation of AI technologies, reinforcing pre-existing inequalities and undermining the already fragile fabric of academic trust (Regmi, 2022).

International studies have since found that students are losing scholarships, facing visa risks, and under the burden of intense distress to be falsely charged with "AI plagiarism" based on detector marks only (Giray, 2024; Pellerin & Ogandaga, 2024). Advice from universities and professional organizations now frequently emphasizes that such tools are "not accurate or reliable" and should not be used as sole evidence of misconduct (Das et al., 2025; Hadra et al., 2025; Nicholas et al., 2025; Rafiq & Qurat-ul-Ain, 2025). Yet in many institutions, including among some Nepali universities that are going through rapid "digital transformation", AI detection is being institutionalized with limited transparency about error rates, governance, or avenues of appeal. Within this frame, the current study delves into how doctoral students and supervisors at two

Nepali universities engage and respond to AI detection practices in their academic work. Admitting that AI detection tools can provide partial insights into questions of academic integrity, it is nevertheless acknowledged in this study that these tools do not fully account for the complexities of how cases of misconduct in certain areas affect academic integrity. Placing the experiences and perspectives of participants at the centre of the study, it challenges the paradox engrained within these technologies: mechanisms meant to maintain academic integrity which, in practice, may undermine trust, transparency and general well-being within doctoral education.

While the technical limitations of AI detectors are becoming ever more visible, their lived experience and socio-cultural consequences in the Global South such as Nepal have not been thoroughly investigated. Most research has come out of Western and Anglophone countries (Perkins et al., 2023). As such, the particular intersection between these technologies and their hierarchical academic cultures, non-native English language contexts, and resource-constrained environments remains critically unexamined. This study seeks to bridge that gap with empirical, qualitative insights from the ground level of the Nepali doctoral education system. Consequently, this study examines AI-detection tools at Nepali higher education, in particular the doctoral level. It asks: How do false positives generated by AI detection tools affect the trust, well-being, and pedagogical relationship between PhD students and their supervisors in Nepal? What are the implications for academic integrity paradigms when technologically mediated suspicion undermines the very relationships meant to foster originality and ethics?

Academic Integrity, Misconduct and Surveillance

Academic integrity is broadly described as a pledge to honesty, trust, fairness, respect, responsibility, and courage in scholarly work (Bieliauskaitė, 2021; Eaton, 2024). The literature regarding integrity in low- and middle-income countries such as Nepal shows recurrent concerns on plagiarism, ghostwriting, contract cheating and the unethical usage of AI applications in the writing process as well as lack of research ethics training (Dahal & Eaton, 2025). In return, institutions have relied more and more on digital surveillance tech—from learning analytics to similarity checkers—that track students' work.

Critical integrity literature contends that academic integrity is a shared responsibility (Eaton, 2025; Javeed, 2018), as "policing" is often divorced from the real-life contexts of students' learning, consolidates deficit discourses, and risks replacing more educative, dialogic orientations. Generative AI has compounded this challenge, increasing the probability of both ethical violations and modes of algorithmic monitoring. And when integrity is defined primarily in terms of risk management, institutions might begin to rely on technical measures (Mulenga & Shilongo, 2024; Pellerin & Ogandaga, 2024), however inadvertently covering up issues as critical with assessment design or workload, or inequitable access to writing support.

Emerging work documents the psychosocial impact of AI related allegations. According to students, they experience 'world crumbling' situations when there is misconduct notice (Angeles et al., 2024; Chan, 2025) based purely on detector outputs, reporting insomnia, anxiety, depression, and social withdrawal. Cases reported by journalists and advocates show students losing scholarships, facing delayed graduations or being forced to "prove" their innocence against opaque algorithms. Scholars argue that such practices invert due process and violate our educational relationships' very basic principle of honesty foundational to educational relationships.

AI detection tools, false positives and false negatives

AI detectors generally predict the probability of having produced a text by a large language model based on token predictability measures or proprietary classifiers. Empirical evidence is clear of large sample sizes that result in both false positives (human writing is flagged as AI-generated) and false negatives (AI-generated text is considered human) and show higher frequency particularly for short and edited texts and nonstandard varieties of English (Liang et al., 2023). In one Bloomberg test, for example, detectors misclassified 1–2% of pre-ChatGPT human essays as AI written and authors mention that real world error rates may even be higher due to motivation to overflag. OpenAI abandoned its own classifier with "low accuracy," and Liang et al. (2023) found that detectors discriminate in favor of non-native English authors, in that their texts often present reduced linguistic complexity that algorithms misinterpret as AI-like texts. This raises an important problem of bias and fairness of algorithms (Weber-Wulff et al., 2023).

False positives are not distributed evenly. Studies and legal analyses indicate that non-native English speakers, neurodivergent writers and those whose styles have a distinctive "voice" are more likely to be misclassified, exacerbating preexisting inequities in the academy. Meanwhile, machine-aided paraphrasing and prompt engineering can be easily unnoticed, which means the most advanced forms of AI may be not visible but honest students may be more often scrutinized.

The integration of AI into Nepal's universities does appear — as both a path to possibility and also a profound source of tension. While these tools hold the potential to enhance learning, they simultaneously present immediate threats to academic integrity, leading teachers to find it challenging to draw a line between human effort and machine output. This struggle underscores a deeper imperative: the development of clear, ethical paradigms to define AI's rightful place in education (Khatri & Karki, 2023; Ghimire et al., 2024; Adhikari & Pandey, 2025). But Nepali higher education's AI detection landscape is fraught with uncertainty. The existing solutions are still plagued with persistent flaws — false positives and false negatives — that tarnish their credibility. The experience of a doctoral student with two inconsistent outcomes from the identical software mirrors the disquieting contradiction of these tools themselves—less based upon scientific accuracy than on unstable guesswork (Hamdan, 2025). This is not an isolated issue. Several studies have found that detection devices fail to discriminate AI outputs and human-authored text — especially in the case of authors who are not native English speakers or from academic niche fields whose output could cause severe miscategorization (Liang et al., 2023).

The consequences are severe. A false positive— incorrectly branding actual student output as AI-produced—can lead to allegations of misconduct, emotional distress, and a breakdown of trust between the student and supervisor (Perkins, 2023). On the other hand, false negatives can leave AI-written works unrevealed, undermining the principle of academic honesty itself. Collectively, this failure generates a "lose-lose" scenario that could be described as an unsolvable dilemma, putting institutions at risk of perceptions of unfairness and lack of accuracy (Davis, 2023). In such a climate, a punitive, tech-first approach is not working. Nepali higher education's answer is not just to double down on poor detectors but to re-envision integrity: crafting a rigorous set of policies, AI literacy and development of assessments that honor process, critical thinking and students' authentic voice (Cotton et al., 2024).

Research Methodology

The study adopted a qualitative interpretivist design using semi-structured interviews to explore how AI detection is experienced by doctoral students and their supervisors in two Nepali universities – one adheres to 20% AI acceptance whereas the second adopts zero tolerance policy. These institutions were purposefully selected because both had introduced AI-related clauses into their academic integrity policies and either piloted or adopted AI-enhanced detection tools in graduate programmes.

The participants consisted of PhD candidates and education and social science doctoral supervisors. A purposive sampling was used to recruit those who had experienced AI-detection reports directly while taking their studies, on qualifying exams, or in thesis supervision, and were willing to talk about issues such as suspected use of AI, false positives, or disagreements with the results. In the first case, we approached a student that had faced the issue and asked for his or her supervisor to partake. We had to replace one participant (student) since her supervisor denied participating.

The study comprised a purposive yet information-rich sample of eight doctoral candidates and four supervisors, including one supervisor who was intentionally paired with two supervisees as part of the sampling strategy. This design reflects qualitative research traditions that prioritize depth of understanding and contextual insight over statistical generalizability. Pseudonyms were assigned to all participants and institutions, and any identifying information included in quotes was anonymized to maintain confidentiality.

Data were collected with in-depth qualitative semi-structured interviews, which were conducted to understand participants' views on generative artificial intelligence (AI), experience with AI-detection tools, and perceptions of their fairness and accuracy. Participants were also asked to consider the relational implications of AI detection in supervisory conversation, assessment contexts, and academic integrity inquiries, which usually meant sharing personal cases in which these technologies had acted as important tools. The interviews were audio-recorded with consent from participants and transcribed in full, with some AI assistance (Gemini.Google). Given limited resources and the delicate nature of research, the study did not collect institutional misconduct records nor detector logs and instead focused on people's experiences and sense making. Transcripts were thematically analysed in the direction of Braun et al.'s (2023) reflexive thematic analysis as adapted in research on integrity and educational technology. Initial coded segments corresponded to experiences of false positives and false negatives; emotional and psychosocial reactions; supervisory and institutional responses; and participants' preferred alternatives. Codes were iteratively refined into broader themes that incorporated commonalities and differences between students and the supervisors. Attention was given to power relations, the language, and tone used in both narratives of accusation and defence, and how participants related local experience to global conversations around AI and integrity.

Given the potential reputational sensitivities surrounding academic integrity, robust ethical safeguards were integral to the study design. Ethical approval was secured from the first author's university research ethics committee, consistent with established standards for responsible conduct of qualitative research (BERA, 2018; Dhakal, 2016). Participants received detailed information sheets explaining the study's purpose, voluntary nature, and confidentiality measures, including assurances that discussions of AI-detection experiences would be treated in aggregate and anonymized form. They retained the right to withdraw at any stage without consequence, and no detector outputs or institutional case records were accessed. These precautions reflected a commitment to relational ethics, researcher reflexivity, and the protection of participants and institutions from potential harm (Tracy, 2020; Guillemin & Gillam, 2004).

Findings

The research examines the nuances of a generative AI approach along with integrity surveillance in Nepali higher education, among doctoral students and their supervisors. These findings are organized into five thematic chapters, which together constitute a narrative arc: first, from the immediate, disruptive consequences of algorithmic tools and then toward a collective reimagining of ethical governance. The analysis opens with a consideration of how AI detection technologies engender a kind of institutionalized doubt, what we refer to as *living under suspicion,* and a process that underlay the collapse of supervision—as students are positioned as "pre suspected"—to a crucial extent. The mood of distrust this fosters results in high *emotional toll,* the second issue, as the anxiety, shame and loss of trust felt by both sides are chronicled. In turn, the third theme, *negotiating uncertainty*, reflects supervisors' precarious efforts to reconcile rigid institutional demands with their personal professional judgment in an environment of murky and unreliable detection protocols. Instead, there are constructive adaptations, arising from within these tensions. The next—and fourth theme, *process evidence, dialogic feedback, and trust-but-verify*, points to the trend toward relational pedagogies of writing development and continuing dialogue over punitive product checking. A series of themes wraps up this journey culminating in Imagining trust-based, contextually informed AI governance in Nepal, which integrates participant perspectives to offer an institutional frame which is concerned with transparency, localized policy formulation and a restorative dedication to academic integrity.

Together, the following threads map a crucial trajectory from surveillance and suspicion to dialogic and context responsive forms of scrutiny.

Theme 1: Living under suspicion – false positives as default doubt

Students across both universities described a pervasive sense of being "pre‑suspected" whenever they submitted written work, particularly after AI‑detection systems were embedded within written work submission and defence procedures. As a doctoral student remarked,

> *It feels like trust is gone before we even begin. The first reader of my writing is no longer my supervisor but the software.* #Doctoral Student 3, University A

This perception of being under default scrutiny illustrates how algorithms have come to mediate scholarly trust, positioning students within an assumption of guilt that must be algorithmically cleared before human engagement.

Likewise, another student reflected,

> *Even before my supervisor reads my chapter, it is already scanned by the software. It feels like I am guilty first, then maybe innocent if the score is low enough.* # Doctoral Student 1, University A

Students repeatedly described this as a form of academic surveillance that erodes confidence in supervisory relationships and produces anxiety around authenticity. Another student added,

> *After I upload my draft, I wait nervously for the percentage. Even when I know it's my own work, the number decides how honest I look.* # Doctoral Student 8, University B

Such statements reveal an emotional landscape defined by algorithmic doubt rather than dialogic feedback.

Supervisors, however, articulated another dimension to this dynamic, noting that their reliance on detectors sometimes arises from experience rather than mere compliance. A doctoral supervisor acknowledged,

> *A few of my students have submitted chapters that read almost entirely like machine output—fluent but soulless. So I now test each document before reading, just to be sure*. # Doctoral Supervisor 2, University A

Similarly, another doctoral supervisor observed,

> *We cannot always tell at first glance whether a text is helped by AI or produced by it. Running the detector first gives a baseline—though I know it's not perfect.* # Doctoral Supervisor 3, University B

These comments underscore supervisors' pragmatic adoption of detection as a screening practice amid genuine uncertainty about the extent of AI use in doctoral writing.

Yet, supervisors expressed discomfort with the resulting inversion of relational norms. A Supervisor noted,

> *I open the dashboard before opening the document because that's the institutional routine now. It feels like I'm policing rather than mentoring.* # Doctoral Supervisor 3, University B

This sentiment reflects the moral tension between institutional accountability and the traditional mentoring ethos of doctoral education. A supervisor from University A further explained,

> *Sometimes the detector gives 27 percent AI, but I recognize the student's phrasing and logic. The system doesn't see that history.* # Doctoral Supervisor 2, University A

Together, these perspectives suggest a context in which trust has become conditional, mediated by algorithmic reports rather than interpersonal judgment. While supervisors' concerns about genuine AI-generated submissions legitimized the initial need for screening, their reliance on scores also perpetuated environments of routine suspicion. For Nepali doctoral candidates—writing in English as an additional language and navigating unfamiliar digital policies—false positives deepened feelings of marginalization much like what Abdelghaffar and Eid (2025) have experienced. As a doctoral student from University B reflected,

> *When my writing improved after attending academic writing workshops, the detector said it looked robotic. It made me question whether learning itself was allowed.* # Doctoral Student 5, University B

The implications of the previous discussions show that there is a paradigm shift in the ecology of supervision where both the student and the supervisor interact with the notion of credibility through the mediation of algorithms. Plagiarism is increasingly confused with authenticity, while the role of mentoring is complicated by bureaucratic initiatives that emphasize the use of technology over trust. This is what the 'false positives as default doubt' captures.

This thematic exploration is enriched by locating participants' experiences within a broader academic examination of algorithmic governance, digital surveillance and epistemic injustice in higher education (Vučković & Sikimić, 2025). Critical scholarship frames AI detection technologies as mechanisms that extend data-driven audit cultures throughout the pedagogical core of academic writing and that remap the relationship between students and their supervisors to revolve around the principles of monitoring and pre-emptive risk management as opposed to trust and developmental mentorship (Selwyn, 2021). In this context, empirical studies have shown that in reality these tools are technically unreliable, since studies show that GPT detectors are biased and lead to exceptionally high false positive rates for non-native English writing (Liang et al., 2023). This institutional weakness brings up serious questions of equity, as it unfairly accuses and punishes (Perkins et al., 2023) multilingual and international students. Nepali doctoral candidates' narratives—describing finding themselves "guilty until the [detection] score proves otherwise"—clearly provide evidence for these harms documented. Their experiences illustrate how opaque algorithmic systems can implement epistemic injustice, methodically sabotaging the credibility and authority of certain writers while uncritically deferring to the decision of flawed technological arbitration (Rosino, 2019).

Theme 2: Emotional toll – anxiety, shame, and erosion of confidence

Students reported profound emotional distress triggered by AI-detection flags or even the mere anticipation of scrutiny, describing experiences of sleeplessness, self-doubt, and existential fear that years of scholarly effort could unravel due to algorithmic misjudgment. A doctoral student captured this visceral impact:

> *When my supervisor forwarded the AI report with a question mark, I didn't sleep the whole night. I started questioning whether my English was 'too good' for a Nepali student. I felt my own improvement was suspicious.* # Doctoral Student 1, University A

Similarly, another student linked these incidents to entrenched linguistic insecurities:

> *We are told to improve our academic writing, but when we improve, the software says maybe it is AI. So where is the safe place for us to grow as writers?* # Doctoral Student 4, University A

This emotional toll extended beyond immediate anxiety to deeper erosions of confidence and identity. Another student recounted,

> *A 21% AI flag on my literature review made me doubt three years of reading. I wondered if I even belong in a PhD program.* # Doctoral Student 6, University B

Another student from University B described a cascading effect:

> *After the detector flagged my methodology, I rewrote everything simpler, hiding my real voice. Now I write to avoid suspicion, not to think.* # Doctoral Student 8, University B

These accounts reveal how detection systems don't merely assess text—they interrogate the writer's legitimacy, particularly for multilingual scholars navigating English academic norms.

Supervisors experienced parallel emotional strain, positioned uncomfortably as enforcers of institutional policy rather than mentors of scholarly growth. A doctoral research supervisor reflected,

> *I entered academia to mentor scholars, not to act as a police officer with a dashboard. When the system flags something, I feel pressure from the institution to act, even if my instinct says the student is honest*. # Doctoral Supervisor 2, University A

Similarly, another supervisor expressed moral distress:

> *Explaining to a panicked student that their life's work shows '72% AI'—when I see their genuine struggle—feels like institutional betrayal*. # Doctoral Supervisor 3, University B
>
> Another research supervisor added,
>
> *I dread opening those reports. They turn supervision into investigation, and I lose sleep knowing good students suffer*. # Doctoral Supervisor 1, University A

The reciprocally negative emotional costs induced by detection-centric integrity regimes point to the relational harm they induce in higher education. Students often internalize algorithmic suspicion as a reflection of personal inadequacy, experiencing heightened anxiety and diminished confidence in their scholarly identity (Das et al., 2025). At the same time, supervisors undergo ethical unease, forced to apply the judgment of flawed technologies during the conflict between their professional knowledge and pedagogical intuition (Rojas Vistorte et al., 2024).

In the context of Nepali doctoral education—which is already structured by linguistic hierarchies, resource limitations, and uneven access to academic writing support—these ecological dynamics in emotion are particularly destabilizing. Supervision as a dialogic nurturing relationship runs the risk of becoming a terrain of mutual suspicion and eroded trust (Lund et al., 2025). The paradox is a stark one: detection tools developed as means to protect academic integrity simply erode the very psychological underpinnings of doctoral formation that undermine not only student self‑efficacy, but also the mentoring structures crucial to scholarly development.

This finding echoes wider criticisms of AI detection technologies across a number of academic domains (e.g., Liang et al. (2023), which point to unreliable, culturally biased, and the potential for misclassification of non-native written English as machine-generated writing, thus accelerating the cycle of social inequities in the global academic system. The resulting false positives not only harm student reputations but may also jeopardize trust with the supervisor, whereas false negatives may compromise the very credibility these instruments claim to protect (Heaven, 2023). Thus, academics maintain that punitive, technology-driven regimes must give way to moral systems of accountability and institutions must establish policy frameworks for ethical AI literacy and assessments that value process and real student voice (Cotton et al., 2024).

Theme 3: Negotiating uncertainty – Supervisors between policy and judgment

In the two universities, the doctoral research supervisors articulated a profound tension between institutional mandates to "take AI scores seriously" and their professional skepticism regarding the tools' reliability, exacerbated by policy documents that offered scant guidance on error rates, appeal mechanisms, or thresholds for "intermediate suspicion bands." This policy ambiguity compelled reliance on personal judgment and informal practices, fostering ad hoc decision-making that risks inconsistent student treatment (Smit et al., 2025). A supervisor put it this way,

> *The policy says we should check for AI-generated content, but it does not say what percentage is unacceptable or how to differentiate between AI assistance and AI authorship. So we negotiate case by case, often in corridor discussions, which is not transparent.* # Doctoral Supervisor 2, University A

Similarly, another supervisor observed,

> *Institutional emails push us to act on high scores, but no one defines 'high' or what happens if a student appeals. It leaves me guessing between compliance and fairness.* # Doctoral Supervisor 1, University A

This ambivalence mirrors broader patterns in higher education where AI-detection technologies are hastily adopted ahead of robust governance frameworks (Miranda & Arndt, 2025), leading to fragmented implementation and ethical unease among faculty. Another supervisor elaborated,

> *We're told to use the dashboard, but when it flags 25% on a strong student's work, do I fail them or ignore it? Policy silence forces triangulation.* # Doctoral Supervisor 3, University B

Another supervisor added,

> *My department head expects action on flags above 20%, but centrally there's no such rule. We end up with unwritten thresholds that vary by supervisor.* # Doctoral Supervisor 4, University B

To navigate this uncertainty, supervisors developed pragmatic strategies emphasizing human-centered evidence over singular algorithmic outputs. For example, they invite students to orally defend or explain flagged sections, probing for authentic understanding; and scrutinize revision histories, draft trails, and collaborative document metadata to trace authorship evolution (Doctoral Supervisor 1, University A). Another supervisor described one such approach:

> *For a flagged proposal, I asked her to walk me through her lit review verbally—her reasoning was impeccable, unlike rote AI summaries.* # Doctoral Supervisor 2, University A

Such practices align with expert feedback for process-oriented integrity verification, which favors dialogic engagement and authentic (not just probabilistic) assessment (American Association of University Professors, 2025; Kings College London, 2025). This theme reveals an underlying governance failure in AI integration, where rapid technology rollouts outpace policy development, effectively relegating supervisors to policymakers on the ground in an era of institutional uncertainty. This kind of "ad hoc experimentation" undermines governance principles because professors are in uncharted territory without clear decision rights: on one hand, they are liable to overreach, but on the other, they risk producing inequitable student evaluations. Based on Fricker’s (2007) model of epistemic injustice, supervisors’ mandated deference to opaque algorithms diminishes their expertise, and students suffer from unequal judgments—especially stark in resource-limited Nepali settings, where multilingual writing styles magnify such detection-related errors. Algorithmic governance critiques highlight this as a specific type of surveillance creep, where detection tools transform mentorship into a compliance surveillance project that corrodes the relational trust foundational to doctoral pedagogy (Gourlay, 2025). These findings call for co-designed protocols rooted in faculty agency, transparency of thresholds, and hybrid human-AI validation processes, helping to restore supervisory judgment as the foundation of integrity assurance. Another dimension is observed—supervising with AI. This paper does not address this question.

There is yet another dimension – supervising with AI, which this paper does not explore.

Theme 4: Process evidence, dialogic feedback and “trust-but-verify”

Participants from both the student and supervisor cohorts showed a convincing preference for integrity systems that envisioned academic writing as a dynamic *process* rather than a static *product*, focusing on tracing authorship process changes rather than detection scores. Moreover, doctoral students even proposed supervisor-led practices to trace their development stages.

A student articulated this vision:

> *If my supervisor sees my drafts from the first messy version to the final one, they will know I am not just pasting from ChatGPT. The process shows my learning.* # Doctoral Student 8, University B

Echoing this, another student emphasized relational transparency and expressed:

> *We need meetings where I can say, 'I used Grammarly here for commas, but this paragraph is all mine from field notes.' Hiding AI makes us look guilty.* # Doctoral Student 3, University A

Likewise, another student advocated for proactive contracts:

> *An AI agreement upfront—'You can use it for outlines but not core arguments'—would build trust instead of waiting for flags.* # Doctoral Student 5, University B

The PhD Supervisors supported these process-focused strategies, and their importance in teaching authentic from shallow simulation and developing responsive approaches to inappropriate use was emphasized. A supervisor from University A described oral defenses as pivotal:

> *Asking a student to explain their flagged methodology verbally reveals depth—no AI can fake three years of fieldwork passion.# Doctoral Supervisor 1, University A*

Another supervisor from University B championed research diaries:

> *Daily logs of reading, writing struggles, and tool decisions create an audit trail far superior to Turnitin percentages.* # Doctoral Supervisor 4, University B

Quite similarly, another supervisor reflected on iterative feedback:

> *Tracking revisions across drafts shows growth patterns; a sudden 'perfect' chapter without mess stands out more reliably than any algorithm.* # Doctoral Supervisor 2, University A

Also, another supervisor beautifully summarized the ethos as *"Trust but verify—through process, not paranoia"* (Doctoral Supervisor 3, University B).

Engagement with process evidence, and dialogic verification in participants' proposals aligns with sociocultural theories of writing as socially mediated development (Prior, 1998) and global calls for academic integrity to be reframed as a relational virtue, rather than technological policing (Davis, 2023; Eaton, 2024; Miranda & Arndt, 2025). These recommendations, spanning explicit versioning audits to dialogic AI disclosures, constitute a change in paradigm toward genuine assessment designs that embed integrity into learning ecologies (Kritik.io, 2025). By privileging artifacts such as revision histories and co-negotiated AI agreements, these practices reinstate human judgment at the center, alleviating the epistemic perils of opaque detectors (Teaching Communication-Intensive, 2025), which disproportionately penalize multilingual and iterative styles of doctoral work. This "trust-but-verify" heuristic replicates design-driven alternative measures, which see assessment as learning, where integrity flows from scaffolded transparency rather than retrospective accusation. In the Nepal higher education landscape, marked by linguistic diversity and uneven digital literacies, these approaches show potential as agents of equitable governance that turn a potential form of monitoring into collaborative capacity-building. But it depends upon institutional investment in training, platform access, and the development of policies beyond vendor-led detection.

Theme 5: Imagining trust-based, contextually grounded AI governance in Nepal

Lastly, participants from both the student and supervisor categories were able to voice visions for Nepali HEIs, esp. universities, to formulate AI and academic integrity policies that emphasize the need for Nepali perspectives on these matters, thus contrary to adopting standardized models that do not consider the country's regulations on these issues. A majority of the students referred to what they see as "legally ambiguous" areas in regards to AI-generated work.

A student insisted:

> *We need to know if detectors scan our theses, how they decide 'AI,' and who sees the scores before we submit anything.* # Doctoral Student 2, University A

Another student emphasized due process:

> *If flagged, what are my rights? Can I see the raw data, challenge the algorithm, or present my drafts? Right now, it's a black box.* # Doctoral Student 6, University B

Similarly, a student from University B linked transparency to equity:

> *Non-native writers like us get higher false positives—policies must say how they'll protect us from automated bias.* # Doctoral Student 8, University B

The majority of students called for “greater communication regarding the use of AI detection tools on doctoral work,” rights to information regarding “student data and marks,” and processes to “challenge alleged AI misuse.”

The participating supervisors consistently advocated for institutional reinvestment in developmental rather than strictly punitive models. Their suggestions focused on dedicated writing assistance, embedded ethics education, and critical AI literacy programming, suggesting that sustainable academic integrity cannot be achieved without capacity building and pedagogical support (Cotton et al., 2024; Hamid, 2025). The participant-led visions we have in mind align well with emerging policy analysis at the Nepali level, which suggests that the implementation of AI ethics, algorithmic explainability, and participatory governance should also be embedded in national educational strategies; in line with a view of advanced technologies as 'neutral or technical appendages' (Dhakal et al., 2025).

Discussion

This analysis highlights a serious relational crisis within Nepali doctoral supervision, instigated by the uncritical deployment of AI-detection technologies. Algorithmic suspicion, as mapped out across five related themes, systematically displaces dialogic trust, turning pedagogical mentorship into a mechanism for compliance monitoring (McGeehan et al., 2018). The shift in this trajectory—from a climate of pre-emptive suspicion—through the documented affective toll and policy ambiguities—to emergent process-oriented resistances reveals a fundamental contradiction: tools deployed to protect academic integrity actively undermine the psychosocial and relational foundations critical to doctoral formation (Becker et al., 2025; Haley et al., 2025). This dynamic compels an epistemic injustice (Fricker, 2007) in which multilingual Nepali scholars are subjected to testimonial smothering by biased false positives and supervisors are obliged to submit to opaque algorithmic metrics above their own professional judgment (Liang et al., 2023). This paradox is severe, creating an ecology of mutual suspicion in which students may resort to defensive and simplistic writing and supervisors have to negotiate unrecognized institutional thresholds.

These tensions highlight a core governance failure, where technology adoption in resource-constrained Global South contexts exceeds ethical, pedagogical, and infrastructural readiness. Supervisors' ad hoc, relational strategies like verbal defenses, iterative writing audits, or co-constructed AI-use agreements are examples of pragmatic resistance. In this way, these practices resonate with sociocultural understandings of writing as a mediated, iterative process (Prior, 1998), and respond to global calls for authentic assessment to embed integrity within the learning process (Bearman et al., 2023). The visionary critique evoked in Theme 5 brings to light the fundamental unsustainability of relying on imported detection tools. It follows that these systems embed Western linguistic and stylistic prejudices which undermine their integration into Nepal's multilingual academic context and magnify epistemic injustices for scholars that navigate postcolonial academic ideologies (Phyak & Ojha, 2019; Sah, 2024).

Such participant visions are a powerful critique of technological determinism in academic integrity in favor of epistemic governance that is embedded in the context and that incorporates AI ethics, explainability, and participatory policy design—echoing postcolonial critiques of uncritical educational technology transfer (Mühlhoff, 2025). The recognized "legal vacuum" around AI-generated plagiarism is emblematic of Global South higher education's structural vulnerability to proprietary systems of detecting unethical behavior which encase Northern linguistic biases and commercial imperatives (Liang et al., 2023). The significance of dialogic transparency and capacity building for participants fits within relational ethical frames (Raza et al., 2025), reframing integrity as a co-constructed virtue rather than as an algorithmic verdict. For Nepal, which is characterized by linguistic diversity, resource challenges, and non-native English speakers that contribute to unequal detection of AI-informed writing and biases, the situation requires a shift towards hybrid, context-sensitive governance models that reposition an AI project from surveillance as a threat to a potential scholarly partner; the governance process should be critical (Ali, 2024).

The current approach to academic integrity can often focus on compliance and technological monitoring for achieving academic honesty (Bertram Gallant, 2017). This paradigm is increasingly challenged by theorists who emphasize the "positive" or "educational" angle, where integrity is reframed as not an imposed rule but a fundamental academic skill, a sociocultural value worth cultivating (Bretag, 2016). This latter model highlights student agency, mastery of scholarly conventions, and the cultivation of a communal ethical ethos, tying the aspiration for integrity in this model to the more comprehensive educational mandate—the individual and intellectual growth. Conversely, an emphasis on punishment for breaking rules, in this form of detection, leads to a climate characterized by alienation and competition, and does not equip students with the fine-grained ethical thinking needed to handle sophisticated digital environments such as Large Language Models (Eaton, 2021). In response to these limitations, current scholarship calls for more relational and proactive approaches focused on trust, dialogue, and pedagogical design. There are some key takeaways from this research that are both interconnected; redesign assessment of authentic process-based tasks that include feedback that is iterative and multimodal demonstrations of learning, systems based on promoting ethical AI literacy so that learners can critically engage with generative AI to make clear boundaries and to think critically about authorship; and explicit embedding of research integrity and writing pedagogy into curricula. For contexts like Nepal, which has long relied on the retention and retribution of misconduct rather than an upstream development of the capacity necessary for lasting scholarly practice (UNESCO, 2024), the above point carries particular weight. Nepal's recent National AI Policy 2025 is about empowering human resources with the skills to work with AI and fostering research and development in the field (Prasain, 2025). Applying this would require moving beyond vendor-led paradigms toward more nationally responsive policy ecologies that respect indigenous epistemologies but that uphold global norms.

Conclusion

This research illustrates the relational fractures being exacerbated by AI-detection technologies in Nepali doctoral supervision: algorithmic suspicion replacing dialogic trust and changing mentorship practice into a model of probabilistic policing. By five closely related themes—from the pervasiveness of "pre-suspicion," through the emotional toll done by false positives, through the policy-induced ambivalences of supervisors, and then at large, to emerging conceptions of process-based resistance and localized governance—Nepali doctoral students and supervisors demonstrate a radical governance deficit. Instead of preserving authenticity, tools of detection continue to reproduce epistemic injustice against multilingual writers, dissolve confidence in professional judgment, reproduce disparities grounded in colonial language hierarchies and resource scarcity (Liang et al., 2023; Perkins, 2023). The Nepali experience reveals the urgent requirement of hybrid governance arrangements that rethink AI as a scholarly ally, rather than a surveillance threat. These frameworks need to account for transparent detection thresholds with mandatory human override, multilingual writing support, and upstream AI literacies as preventive infrastructure.

Faculty-driven appeals based on process evidence as well as participatory policy co-design relevant to local epistemologies are central to rebuilding trust and fairness (Cotton et al., 2024;

Das et al., 2025). Practical imperatives include integrating version-history audits, oral defenses, and co-constructed AI agreements in doctoral programs, ensuring that it's not permissible to rely solely on detector scores, and advancing integrity education that builds national capacity (Heaven, 2023). At the sectoral level, regulatory bodies such as the University Grants Commission and NAAMII need to uphold vendor transparency, protect data privacy, and align to Nepal's evolving legal landscape regarding copyright and technology governance (Lund et al., 2025).

This work can be further developed in future scholarship by capturing the dynamics of detection across undergraduate, professional, and distance education contexts, through comparative analyses of how AI becomes enmeshed in the legacies of South Asian colonialism, and by pioneering student-led models of emancipatory AI literacy. With a focus on relational ethics and contextual responsiveness, Nepal is situated to not just modify Western ways of detection, but to lead the way to build equitable integrity ecologies—turned disruption of technology into a chance on which to create decolonized, trust-based knowledge production that connects the global South (Adhikari & Pandey, 2025; Sivasubramaniam, 2024; Khatri & Karki, 2023).

References

Abdelghaffar, A., & Eid, L. (2025). A critical look at equity in international doctoral education at a distance: A duo's journey. *British Journal of Educational Technology*, *56*(2), 834-851.

Adhikari, D. P., & Pandey, G. P. (2025). Integrating AI in higher education: Transforming teachers' roles in boosting student agency. *Educational Technology Quarterly, 2*, 151-168.

Ali, S. (2024). Balancing perspectives: Assessing the integration of AI in academic support within higher education (Honors thesis, University of Texas at Arlington). https://mavmatrix.uta.edu/honors_spring2024/37

American Association of University Professors. (2025). *Artificial intelligence and academic professions*. https://www.aaup.org/reports-publications/aaup-policies-reports/topical-reports/artificial-intelligence-and-academic

Angeles, C. N., Samson, B. D., Mama, B. R. Z. I., Luriaga, R. L., Delizo, J. P. D., & Ching, M. R. D. (2024, May). Students' perception of the use of AI Detector System by faculty members in determining the originality of submitted academic requirements. In *Proceedings of the 2024 8th International Conference on E-Commerce, E-Business, and E-Government* (pp. 56-61).

Bearman, M., Dawson, P., Ajjawi, R., Tai, J., & Boud, D. (Eds.). (2023). *Re-imagining university assessment in a digital world*. Springer.

Becker, S., Jacobsen, M., & Friesen, S. (2025). Four supervisory mentoring practices that support online doctoral students' academic writing. *Front. Educ. 10,* 1521452. https://doi.org/10.3389/feduc.2025.1521452

Bertram Gallant, T. (2017). *Academic integrity as a teaching and learning issue: From theory to practice*. Routledge.

Bieliauskaitė, J. (2021). Solidarity in academia and its relationship to academic integrity. *Journal of Academic Ethics, 19*(3), 309-322.

Braun, V., Clarke, V., Hayfield, N., Davey, L., & Jenkinson, E. (2023). Doing reflexive thematic analysis. In *Supporting research in counselling and psychotherapy: Qualitative, quantitative, and mixed methods research* (pp. 19-38). Springer International Publishing.

Bretag, T. (Ed.). (2016). *Handbook of academic integrity*. Springer.

British Educational Research Association. (2018). *Ethical guidelines for educational research* (4th ed.). BERA.

Burr, V. (2022). The supervision of doctoral students in the age of digital surveillance: Trust, anxiety, and the negotiation of academic identity. *Journal of Further and Higher Education, 46(*8), 1105-1118.

Chan, C. K. Y. (2025). Students' perceptions of 'AI-giarism': Investigating changes in understandings of academic misconduct. *Education and Information Technologies*, *30*(6), 8087-8108.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Dahal, B., & Eaton, S. E. (2025). Academic and research integrity in Nepali universities: A comprehensive policy analysis. *Journal of Academic Ethics*, *23*, 2231-2255. https://doi.org/10.1007/s10805-025-09649-5

Dahal, N. (2024). 'Ethics' and 'integrity' in research in the era of generative AI—Are we ready to contribute to scientific inquiries? *GS Spark: Journal of Applied Academic Discourse, 2*(1), 1-6. gsspark.gsmultiplecampus.edu.np/download/ethics-and-integrity-in-research-in-the-era-of-generative-ai-are-we-ready-to-contribute-to-15939.pdf

Das, P., Edgington, W. D., Ghosh, N., & Rahaman, M. S. (2025). Evaluating the effectiveness and ethical implications of AI detection tools in higher education. *Information, 16*(10), 905. https://doi.org/10.3390/info16100905

Davis, A. (2023). Academic integrity in the time of contradictions. *Cogent Education*, *10*(2). https://doi.org/10.1080/2331186X.2023.2289307

Dhakal, C., Adhikari, D., Subba, G., & Kapadi, P. R. (2025). Reimagining Nepal's future: AI for human development and education. *Panauti Journal, 3*, 121-134.

Dhakal, R. K. (2016). Responsible practice of research: Safeguarding research integrity and publication ethics. *Journal of education and research*, *6*(2), 1-11.

Duah, J. E., & McGivern, P. (2024). How generative artificial intelligence has blurred notions of authorial identity and academic norms in higher education, necessitating clear university usage policies. *The International Journal of Information and Learning Technology*, *41*(2), 180-193.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642.

Eaton, S. E. (2021). *Plagiarism in higher education: Tackling tough topics in academic integrity*. ABC-CLIO.

Eaton, S. E. (2025). *A wraparound approach to academic integrity: Centering students in the postplagiarism era.* Available at SSRN. http://dx.doi.org/10.2139/ssrn.5223911

Eaton, S. E. (Ed.). (2024). *Second handbook of academic integrity*. Springer.

Fraser, K. C., Dawkins, H., & Kiritchenko, S. (2025). Detecting AI-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, *82*, 2233-2278.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Ghimire, S. N., Bhattarai, U., & Baral, R. K. (2024). Implications of ChatGPT for higher education institutions: Exploring Nepali university students' perspectives. *Higher Education Research & Development, 43*(8), 1769-1783.

Giray, L. (2024). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. *The Serials Librarian, 85*(5-6), 181-189.

Gourlay, L. (2025). *The university and the algorithmic gaze*. Bloomsbury Academic.

Guillemin, M., & Gillam, L. (2004). Ethics, reflexivity, and "ethically important moments" in research. *Qualitative Inquiry, 10*(2), 261–280.

Hadra, M., Cambridge, K., & Mesbah, M. (2025). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *Research Square,* Pre-print. https://doi.org/10.21203/rs.3.rs-7359956/v1

Hale, R. (2025, Jan 22). She lost her scholarship over an AI allegation — and it impacted her mental health. *USA TODAY.* https://www.usatoday.com/story/life/health-wellness/2025/01/22/college-students-ai-allegations-mental-health/77723194007/

Haley, A., Holmqvist, M., & Johansson, K. (2025). Supervisors' competences from doctoral students' perspectives – a systematic review. *Educational Review*, *77*(6), 1971–1990. https://doi.org/10.1080/00131911.2024.2306938

Hamdan, A. (2025). The double-edged sword of AI-integrated education: an investigation into personalized and inclusive learning in higher education. In Hafinaz, R. Hariharan, & R. Senthil Kumar (Eds.), *Recent research in management, accounting and economics*. Routledge. https://doi.org/10.4324/9781003606642

Hamid, K. (2025, May 16). *From policing to pedagogy: Reimagining academic integrity for the AI age.* https://www.linkedin.com/pulse/from-policing-pedagogyacademic-integrity-ai-era-hamid-khan-fhea-gk3ee/

Heaven, W. (2023). *The unsolvable problem of AI detection*. MIT Technology Review.

Javeed, S. (2018). *Academic advisors as valuable partners for supporting academic integrity*. *1*(1), 22–26. https://doi.org/10.11575/CPAI.V1I1.52759

Kamler, B., & Thomson, P. (2014). *Helping doctoral students write: Pedagogies for supervision*. Routledge.

Khatri, B. B., & Karki, P. D. (2023). Artificial Intelligence (AI) in higher education: Growing academic integrity and ethical concerns. *Nepalese Journal of Development and Rural Studies, 20*(01), 1–7. https://doi.org/10.3126/njdrs.v20i01.64134

Kings College London. (2025). King's guidance on generative AI for teaching, assessment and feedback. https://www.kcl.ac.uk/about/strategy/learning-and-teaching/ai-guidance

Kirsanov, O., Kushwah, L. & Selvaretnam, G. (2026). Beyond detection: How students use—and hide—AI in online assessments and what authentic tasks can do about it. *Journal of Academic Ethics, 24*, Art. 14. https://doi.org/10.1007/s10805-025-09691-3

Kotsis, K. T. (2025). Redefining scientific authorship in the age of AI: Challenges for editors and institutions. *European Journal of Innovative Studies and Sustainability, 1*(5), 23-33. https://doi.org/10.59324/ejiss.2025.1(5).03

Kritik.io. (2025). AI detection vs. AI visibility: The future of academic integrity. *Kritik*. https://www.kritik.io/blog-post/ai-detection-vs-ai-visibility-the-future-of-academic-integrity

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

Lund, B. D., Lee, T. H., Mannuru, N. R., & Arutla, N. (2025). AI and academic integrity: Exploring student perceptions and implications for higher education. *J Acad Ethics, 23*, 1545–1565. https://doi.org/10.1007/s10805-025-09613-3

McGeehan, C., Chambers, S., & Nowakowski, J. (2018). Just because it's digital, doesn't mean it's good: Evaluating digital picture books. *Journal of Digital Learning in Teacher Education*, *34*(2), 58–70. https://doi.org/10.1080/21532974.2017.1399488

Miranda, R., & Arndt, K. (2025, December 30). *How AI is reshaping cloud strategy and governance in higher Education.* EdTech: Focus on Higher Education. https://edtechmagazine.com/higher/article/2025/12/how-ai-reshaping-cloud-strategy-and-governance-higher-education

Mühlhoff, R. (2025). The ethics of AI: Power, critique, responsibility (p. 233). Bristol University Press.

Mulenga, R., & Shilongo, H. (2024). Academic integrity in higher education: Understanding and addressing plagiarism. *Acta Pedagogia Asiana, 3*(1), 30-43.

Nicholas, D., Herman, E., Clark, D., Abrizah, A., Revez, J., Rodríguez‐Bravo, B., ... & Watkinson, A. (2025). Integrity and Misconduct, Where Does Artificial Intelligence Lead? *Learned Publishing*, *38*(3), e2013.

Pellerin, M., & Ogandaga, M. (2024). Rethinking academic integrity and plagiarism for a new AI era. Paper presented at Université du Québec en Outaouais, Canada (21-24 May 2024).

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice, 2,* Art. 07. https://files.eric.ed.gov/fulltext/EJ1382355.pdf

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023). The weakening of academic integrity? A threat beyond ChatGPT. *Journal of University Teaching & Learning Practice*, 20(5).

Phyak, P., & Ojha, L. P. (2019). *Language education policy and inequalities of multilingualism in Nepal*. Routledge.

Prasain, K. (2025, Aug 16). Nepal rolls out ambitious AI policy. *The Kathmandu Post*. https://kathmandupost.com/money/2025/08/16/nepal-rolls-out-ambitious-ai-policy

Prior, P. (1998). *Writing/disciplinarity: A sociohistoric account of literate activity*. Lawrence Erlbaum.

Rafiq, S., & Qurat-ul-Ain, D. A. A. (2025). The role of AI detection tools in upholding academic integrity: An evaluation of their effectiveness. *Contemporary Journal of Social Science Review*, *3*(1), 901-915.

Raza, F. A., Singh, A. D., Kovilpillai, J. J. S., Hamdan, A., & Rajaratnam, V. (2025). Safeguarding Integrity in AI-Enhanced Education: Stakeholder Perspectives on Accuracy, Validity, and Ethics in ASEAN. European Journal of STEM Education, 10(1), 22.

Regmi, K. D. (2022). Digitalization of higher education in the Global South: Challenges of adaptation and neoliberal imperatives. *Policy Futures in Education*, 20(7), 855-872.

Rentier, E.S. (2025). To use or not to use: exploring the ethical implications of using generative AI in academic writing. *AI Ethics, 5*, 3421–3425. https://doi.org/10.1007/s43681-024-00649-6

Rojas Vistorte, A. O., Deroncele-Acosta, A., Martín Ayala, J. L., Barrasa, A., & Martí-González, M. (2024). *Integrating artificial intelligence to assess emotions in learning environments: A systematic literature review*. Frontiers in Psychology, 15, 1387089. https://doi.org/10.3389/fpsyg.2024.1387089

Rosino, M. L. (2019). *Algorithms of oppression: How search engines reinforce racism*. *Social Forces, 97*(4), e1–e3, https://doi.org/10.1093/sf/soz004

Sah, P.K. (2024). Nepal: Language, schooling, and inequalities for ethnic minority children: An ethnography of medium of instruction policy in Nepal's public schools. In R. A. Giri, A. Padwad, & M. M. N. Kabir (Eds.), Equity, social justice, and English medium instruction. Springer. https://doi.org/10.1007/978-981-97-8321-2_5

Selwyn, N. (2021). *Education and technology: Key issues and debates* (3rd ed.). Bloomsbury Academic.

Sharma, G. (2019). *Higher education in Nepal: Policies and perspectives*. Oxford University Press.

Sivasubramaniam, S.D. (2024). Academic integrity in South Asia: Focus on India, Pakistan, and Sri Lanka. In S. E. Eaton (Eds.), *Second handbook of academic integrity*. Springer. https://doi.org/10.1007/978-3-031-54144-5_88

Smit, M., Wagner, R.F., & Bond-Barnard, T. J. (2025). Ambiguous regulations for dealing with AI in higher education can lead to moral hazards among students. *Project Leadership and Society, 6,* 100187. https://doi.org/10.1016/j.plas.2025.100187

Subedi, P. (2024, July). *Generative AI in journalism: International practices and the Nepali context.* https://sl1nk.com/amiOr

Teaching Communication-Intensive. (2025, Nov 04). What we learned about teaching for integrity in the age of AI. https://sl1nk.com/BVZNK

Tracy, S. J. (2020). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact* (2nd ed.). Wiley-Blackwell.

UNESCO. (2024). *Unveiling the intricacies of AI governance in Nepal: Multistakeholder dialogue on artificial intelligence governance*. https://www.unesco.org/en/articles/unveiling-intricacies-ai-governance-nepal-multistakeholder-dialogue-artificial-intelligence

Vučković, A., & Sikimić, V. (2025). Global justice and the use of AI in education: ethical and epistemic aspects. *AI & Society, 40*, 3087–3104. https://doi.org/10.1007/s00146-024-02076-x

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), 1-39.