journal of ethics in higher education

Rethinking Sustainable Integrity Practices: Ethical AI Use in Higher Education through Panellists' Lived Experiences

Micheal M. van Wyk

Published June 2026

Keywords

Artificial Intelligence, AI guardrails, ethics, panellist, panel discussion methods, qualitative approach, single case study design

Abstract

This study explores panellists' understanding of AI guardrails and the challenges they identify, to inform better strategies for ethical and responsible AI use in academia. The research adopts Sociotechnical Systems Theory as its primary approach, these frameworks help analyse interactions between people and technology, as well as the processes by which ethical decisions are made. The study uses a qualitative single-case design, focusing on a panel discussion to examine key issues related to AI safeguards and the importance of ethical AI use. NVivo was used to analyse the data and identify themes. The findings show that panellists view AI guardrails as tools for protecting systems, securing information, monitoring unethical behaviour, and ensuring that institutions follow rules. Based on these findings, the study suggests implications for higher education to support robust ethical governance and responsible AI use.

Corresponding Author: Prof. Micheal M. van Wyk, Department of Curriculum and Instructional Studies, College of Education, University of South Africa, Email: vanwykm4@gmail.com.

To quote this article: Van Wyk, Micheal. 2026. "Rethinking Sustainable Integrity Practices: Ethical AI Use in Higher Education through Panellists' Lived Experiences", *Journal of Ethics in Higher Education* 8.1(2026): 89–119. DOI: https://doi.org/10.26034/fr.jehe.2026.9765 Visit: https://jehe.globethics.net

1. Introduction

This discussion focuses on a key question: Why is ethics so important to humanity? Ethics shapes who we are, guides how we interact, and supports fairness and respect in our relationships. It helps us distinguish right from wrong, encourages peaceful living, and protects human dignity.

As Langlois et al. (2025, p.2) note: *"Ethics is a complex human concept anchored in cultural norms, values and social beliefs."* Therefore, the importance of ethics lies in ensuring AI is developed and used responsibly, protecting privacy, fairness, transparency, and accountability, upholding human rights, and preventing harm or bias.

Bringing this sentiment to the ethical foundation of Artificial Intelligence (AI) is both necessary and urgent. As AI and machine learning systems play a greater role in decisions that affect people's lives, their impact on society grows. These technologies are no longer neutral tools; they influence social, economic, and educational outcomes. Therefore, ethics must guide the development and use of AI. Ethics supports fairness, respect, and accountability; helps prevent AI from exacerbating existing biases and inequalities; and promotes transparency, accountability, and human rights. By putting ethics and humanity at the centre of AI, we can ensure technology advances in line with human values and helps build a fairer society.

Studies have reported that the exponential rise of generative artificial intelligence (GenAI) tools in higher education has deeply altered teaching, learning, student support, and assessment practices (Farrokhnia et al, 2025; Kaouni et al., 2025). While these technologies deliver considerable benefits, they also entail serious risks of academic dishonesty and unethical behaviour. To reduce these risks, some HEIs have implemented AI safety measures, such as detection tools, proctoring systems, and plagiarism checks, to protect the integrity of online assessments.

Semantically, a guardrail could be understood as a set of protective boundaries to prevent harm, shape human behaviour and enforce a system within permissible risk limits. An AI guardrail is a tool that provides guidance, safety and support for sustainable integrity by helping students and staff to use AI responsibly. On the other hand, gatekeeping may create

resistance, inequality or even push misuse underground. Gatekeeping is the exclusion of others from participating in an event or from controlling access to the event or to information in an enterprise data system. This paper focuses on AI guardrails. Empirically, Mishra (2025, p.2) conceptualised that:

> *"AI guardrails are structured mechanisms—spanning technical, ethical, and regulatory dimensions—that keep artificial intelligence systems aligned with defined standards and societal values."*

Studies also describe AI guardrails as essential policies that include rules, regulations, technical limits, and ethical guidelines to protect data, privacy, and security (Clark et al., 2025; Dev et al., 2025; Kumar, 2026; Mulahuwaish et al., 2025). AI guardrails focus on safety and fairness, ensuring accountability, supporting transparency, and maintaining security. In this sense, an AI guardrail is a safety and cybersecurity measure built on policies, rules, and guidelines to detect and protect personal data from unethical actions. AI guardrails are suggested as ways to encourage responsible use by adding ethical, technical, and procedural protections. However, there is still little research on how key stakeholders, especially expert panellists, understand and view these guardrails.

Recent evidence indicates no common or clear understanding of what AI guardrails are or how they help maintain academic integrity (Mishra, 2025; Mulahuwaish et al., 2025). This uncertainty makes it harder to implement AI guardrails due to challenges across institutions, technologies, and teaching methods, potentially limiting their effectiveness in supporting ethical and responsible academic practices. The lack of clear and consistent approaches creates a gap between what AI guardrails are meant to do and how they are actually used in higher education.

An article in The Conversation by Rebecca Davis (2025) reported a major concern about a cheating crisis involving AI technologies in South African universities. Davis (2025) reported:

> *"Now, they face an insidious new challenge: across the country, students are using ChatGPT and other increasingly sophisticated large language models (LLMs) to generate*

> *essays, solve assignments, and cheat their way through their degrees. The institutions are scrambling to find ways to either stop them or harness the use of AI as a positive academic tool" (p.2).*

Specifically, studies have revealed that:

- Students increasingly use AI to generate assignments, reducing authentic student work (Davis, 2025; Jonas, 2024; Nasution & Fransiska, 2026).
- Detection of AI-produced content is becoming unreliable as technologies advance (Boutadjine et al., 2025; Elkhatat et al., 2023).
- Universities are struggling to respond effectively, bringing about inconsistent policies, unclear guidelines, and uneven enforcement (Zhai, 2024).
- Excessive dependence on flawed AI-detection tools causes concerns about fairness and credibility (Giray, 2024; Jonas, 2024; Zhai, 2024).

Recently, the South African Minister of Communications and Digital Technologies gazetted the Draft National AI Policy for public comment, but three days later retracted it amid an outcry over "fabricated citations" that misrepresented the policy (Tembo, 2026). The draft National AI Policy represented a constructive step toward addressing the unethical use of large language models. However, the inclusion of "fake sources" undermined its ethical integrity, procedural validity, and technical credibility. As a result, the Minister withdrew the document, referred it to the Parliament Portfolio Committee Communications for revision, and indicated that it would be republished for public comment (Slimi & Chichti, 2026).

Although many people see these challenges, higher education institutions (HEIs) respond differently to protecting academic integrity as AI use increases. Some are proactive, building AI guardrails into their policies, training staff, and planning for ethical risks. Others only react after problems occur, leading to inconsistent, sometimes confusing policies for staff and

students. Many institutions are still not fully ready to protect academic integrity in environments where AI is common. Even though research points out both the benefits and risks of AI in education, there is still no comprehensive, institution-wide framework for AI governance. Most South African institutions do not treat academic integrity as a shared responsibility, and there is little coordination between policies, teaching, and technical controls (Davis, 2025; Giray et al (2025). As a result, research and actions are fragmented and do not cover all the challenges to academic integrity in the age of AI. There is still a significant gap between the widespread use of AI in higher education and the lack of integrated frameworks for managing its ethical use. (Maleki, 2026)

Within the context of this study, the College of Education at the University of South Africa (UNISA) launched the Digitalisation Project as part of the university's strategic focus on AI and digitalisation. As a project coordinator and member, I planned, hosted and participated in several webinars on Navigating Ethical Challenges: The Role of AI in Academic Integrity and Plagiarism Prevention (May 12, 2025). This project is part of the College Strategic Objectives to advance one of UNISA's catalytic niche areas, AI and Digitalisation.

This study explores panellists' understanding and perceptions of AI guardrails, as well as the challenges they identify, to inform better strategies for promoting ethical and responsible AI use in academic environments. These panellists are from Southern African universities who agreed to participate in the webinar. Based on this aim, the following specific research questions (SRQs) are formulated:

- SRQ1: How do panellists conceptualise and interpret the function of AI guardrails in preventing unethical practices?
- SRQ2: How do panellists perceive the role of AI guardrails in providing ethical, technical, and procedural safeguards to protect academic integrity?
- SRQ3: How do panellists describe the challenges associated with implementing AI guardrails to ensure responsible and compliant academic practices?

The subsequent sections present the study's theoretical framework, detail the mechanisms of AI guardrails, describe the empirical research design, report the findings, and conclude with specific implications for AI ethics.

2. Literature Review

Theoretical framework for the study

This study draws on several theories to examine how AI guardrails are understood, used, and evaluated in higher education. These theories are Sociotechnical Systems Theory, Responsible AI and Ethical Governance frameworks, Institutional Theory, and Academic Integrity Theory, which underpin this study. First, Sociotechnical Systems Theory (STS) provides the principal analytical foundation for this study. STS conceptualises organisations as systems in which social elements (people, practices, values) and technical elements (tools, infrastructures, algorithms) are interdependent and must be jointly optimised for effective functioning (Baxter & Sommerville, 2011; Memarian & Doleck, 2025). This perspective is particularly relevant to AI guardrails, which involve algorithmic detection tools, institutional policies, educators' judgment, and student behaviour. Applying STS enables this study to conceptualise AI guardrails as institutional systems rather than discrete technologies. For example, AI detection tools do not operate in isolation; their effectiveness depends on how lecturers interpret reports, how students understand expectations around integrity, and how institutions regulate their use. Thus, STS supports a critical analysis of documented failures of AI guardrails, such as false positives, biased outcomes, and staff resistance, as manifestations of misalignment between technical systems and social practices rather than purely technological deficiencies (Memarian & Doleck, 2025; Williamson & Eynon, 2020).

While STS explains how AI guardrails function, Responsible AI and Ethical Governance frameworks provide the normative justification for their adoption. These frameworks emphasise principles such as fairness, transparency, accountability, explainability, privacy, and human monitoring

(Carney, 2022; Floridi et al., 2018; Floridi, 2019; OECD, 2019; UNESCO, 2021). In higher education, these principles are increasingly invoked to guide the ethical use of AI in assessment, monitoring, and decision-making. In this study, Responsible AI frameworks inform the analysis of ethical and governance guardrails, including institutional AI policies, ethical review mechanisms, and compliance structures. They also underpin the evaluation of technical guardrails, particularly AI detection and proctoring tools, by highlighting concerns about algorithmic bias, opacity, and disproportionate surveillance of marginalised students (García-Peñalvo et al., 2021; Swauger, 2020). The frameworks enable the study to examine whether AI guardrails foster ethical legitimacy and trust or merely serve as compliance-driven risk-management tools.

To explain variations in the adoption and enactment of AI guardrails across institutions, the study draws on Institutional Theory (DiMaggio & Powell, 1983; Meyer & Rowan, 1977). Institutional Theory argues that organisations often adopt policies and practices to maintain legitimacy and conformity rather than to improve effectiveness. Within this lens, AI guardrails may be adopted in response to outside pressures, such as public concern about AI-enabled cheating, accreditation requirements, or sector-wide policy trends, without being meaningfully embedded in teaching and assessment practices. Institutional Theory, therefore, supports a careful review of the observed discrepancy between policy rhetoric and operational reality, explaining why AI policies and detection tools often exist without consistent staff training, evaluation mechanisms, or pedagogical alignment (Selwyn, 2023).

Given the study's explicit attention to academic dishonesty, Academic Integrity Theory gives a crucial foundation for pedagogy. Academic integrity scholarship conceptualises integrity not simply as rule enforcement, but as a shared educational value embedded within institutional culture and evaluation design (Bretag, 2016; Macfarlane et al., 2014). Contemporary approaches advocate a shift away from punitive surveillance towards educative, preventive, and developmental strategies. Within the outlined framework, pedagogical and procedural guardrails, including authentic assessment, transparent guidelines for AI use, and AI literacy development, are provided. These are viewed as essential complements to technical

detection tools (Eaton, 2023). Academic Integrity Theory thus strengthens the study's critique of excessive reliance on AI surveillance and positions pedagogical guardrails as central to sustainable integrity practices in the AI era.

A synopsis of the mechanisms of AI guardrails

As mentioned earlier, studies describe AI guardrails as technical, ethical, and governance tools that help AI and machine learning systems operate safely, fairly, and responsibly. In higher education and society, these guardrails act as preventive and corrective measures to guide how people and AI interact, reduce harm, and protect institutional integrity (Floridi et al., 2018; Eaton, 2023). AI guardrails are becoming increasingly important due to risks such as misinformation, academic dishonesty, and bias. Institutions are beginning to implement policies, detection tools, and guidelines to manage the use of AI in teaching, learning, and assessment (UNESCO, 2023; OECD, 2021). Nikolinakos (2023) also suggests measures based on fairness, accountability, transparency, and ethics (FATE). However, the use of AI guardrails remains uneven, with some institutions only acting after security problems arise. Some take proactive steps, while others react, leading to inconsistent practices and policy gaps. (Dabis & Csáki, 2024) Studies show that effective AI guardrails need a holistic approach that combines technology, ethical awareness, staff training, and ongoing policy review (Chan, 2023; Evangelista, 2025; Zlotnikova et al., 2025). Investing in AI guardrails should not be just about restrictions but about creating frameworks that support innovation while protecting human values and academic integrity (Harjika, 2026; Molina-Carmona & García-Peñalvo, 2025).

AI guardrails can be better understood by unpacking their technical, ethical, and governance mechanisms, each operating at different yet interconnected levels of control and accountability. First, technical mechanisms are embedded directly within AI systems to prevent harmful or unreliable outputs. This view is supported by Atmakuri (2025, p. 11), who states that "embedded AI's democratizing potential depends on sustained commitment to inclusive design principles, diverse stakeholder engagement, and the development of technologies that serve the broader interests of global society

rather than perpetuating existing inequalities. Scholars have proposed several technical mechanisms, including algorithmic auditing, bias-detection and mitigation models, content filtering, and explainability tools, to enhance transparency and make decision-making processes interpretable (Clark et al., 2025; Dev et al., 2025). Furthermore, Sharma (2025) proposed a PPO-based RLHF framework integrating adversarial testing and automated monitoring to align outputs with human values. In the context of ML, these safeguards aim to improve robustness, fairness, and accuracy while reducing risks like hallucinations and discriminatory outputs (Floridi et al., 2018; Xui et al., 2023).

As alluded to, ethics lies at the heart of humanity; therefore, ethical mechanisms focus on normative principles that guide the responsible use of AI. Rooted in Ethics, these include fairness, accountability, transparency, and respect for human autonomy (Gianni et al., 2022; Ryan & Stahl, 2021). Studies have shown that ethical guardrails require developers and users to critically evaluate issues such as bias, consent, privacy, and the potential societal impact of AI systems (Clark et al., 2025; Dev et al., 2025; Kumar, 2026; Mulahuwaish et al., 2025). In addition, AI ethics frameworks proposed by organisations such as UNESCO (2023), OECD (2019), and UNESCO (2021) emphasise human-centred AI, ensuring that systems augment rather than undermine human dignity and agency.

In the context of this study, AI guardrails and governance are a system of specific rules, processes, and oversight to ensure that AI systems are designed, deployed, and used ethically, safely, and in line with policy imperatives and institutional values. Therefore, HEIS should ensure that governance mechanisms are implemented and operate at institutional and regulatory levels. These governance mechanisms are policies, compliance frameworks, auditing standards, and oversight bodies that regulate AI deployment. As an excellent example, Agarwal and Nene (2025) propose a five-layer framework for AI governance that could be used for integrating regulation, standards, and certification. In higher education, governance may involve academic integrity policies, guidelines for AI use, and data protection regulations that align with established frameworks (Mahrishi et al., 2025). Effective governance ensures effective decision-making and accountability

through clear roles, monitoring systems, enforcement procedures, and the enforcement of those constraints (Ademeso et al., 2025).

These mechanisms work together in layers: technical systems provide real-time protection, ethical principles guide responsible design and use, and governance structures ensure accountability and long-term success.

3. Methodology

This study used an exploratory qualitative approach, with a panel discussion as the main research method. The panel drew on panellists' knowledge and experiences, encouraging them to share evidence-based insights and engage with one another (Brannick et al., 2010; Marchant et al., 2001; Nasiri & Khojasteh, 2024). The single case study centred on a critical discussion of AI guardrails. The panellists and moderator were from the University of Ghana, the University of Mauritius, the Namibia University of Science and Technology, and the University of South Africa. Specific codes (PM1 to PM5) were used to protect the identity and confidentiality of the panellists and the moderators for the panel discussion. Five panellists from Southern African universities participated, and all consented to the online session. The 50-minute panel discussion allowed participants to share their views on issues related to student misconduct and the misuse of AI technologies. Panellists prepared in advance to answer four research questions, each speaking for up to 10 minutes. They discussed topics such as safety measures in their courses (RQ1), using AI detection to identify misuse by students and staff (RQ2), and managing AI-generated content, academic integrity courses, and tools like Turnitin and the Invigilator app to prevent cheating or plagiarism (RQ3). After the discussion, there was a 10-minute Q&A session during which the audience could ask questions directly or via the MS Teams chat. All sessions were recorded, and transcripts and chat responses were downloaded and checked for accuracy. For data analysis, NVivo 14.0 was used to identify themes through a step-by-step coding process.

- Step 1: Data Importation and Preparation – The transcripts and video recordings from MS Teams were imported into the NVivo

computerised qualitative software, titled 'Digitalisation Project'. Each case for this project was created using codes PM1-PM5.

- Step 2: Initial coding of the dataset – The transcripts were read line by line, and codes were assigned to each case using an inductive approach (direct verbatim extracts for PMs). Codes were assigned to each case: AI detection, AI provenance integration, AI surveillance, and Challenges.
- Step 3: NVivo-Assisted Pattern Recognition – After the coding process is complete, NVivo "runs" and generates frequently coded concepts or nodes, linking them to co-occurrences of codes.
- Step 4: Node Aggregation of Categories – AI-detection, Turnitin, Invigilator App, unethical behaviour, flagging non-compliance, and AI-generated content.
- Step 5: Theme Development (interpretative synthesis) -Themes were generated by synthesising categories linked to the research questions.
- Step 6: Validation and Refinement – themes were refined throughout the process through constant comparison of transcripts, revisiting disconfirming evidence, and cross-checking against theory and literature.

After the NVivo analysis was completed, the transcripts and recordings were sent to the panellists for accuracy checks. Each panellist confirmed by email that the transcripts accurately reflected the panel discussion.

4. Findings

The first RQ1 explores panellists' perceptions of AI guardrails regarding unethical behaviour, particularly academic dishonesty. This research question examines issues related to measuring information safeguarding (private data), protecting sensitive information, and the ethical use of AI guardrails to ensure academic integrity.

Safety and cybersecurity strategies for guarding, protecting and detecting unethical behaviour in the academic context

AI guardrails are conceptualised as protective mechanisms that safeguard information, uphold ethical conduct, and maintain responsible use of AI systems. They function by securing data through controlled access, ensuring that sensitive information is not exposed or misused. Guardrails also monitor user behaviour and AI-generated outputs to detect unethical practices such as manipulation, misinformation, or academic dishonesty.

Another panellist (PM3) stated that the concept is:

AI safeguards are designed to guide, protect, and restrict or monitor unethical behaviour. It is for protecting information and data, and for monitoring the unethical behaviour of users of AI technologies.

This panellist (PM1) provided a synthesis of what he understood an AI guardrail is about. He stated that it:

> *secures data by controlling access to sensitive information, thereby preventing exposure or misuse. It is to detect unethical practices, misinformation, or academic dishonesty.*

Participants see AI guardrails as tools or measures to protect, detect, restrict, and monitor data and sensitive information, and to identify and track unethical behaviour and practices in academia.

In response to RQ2, panellists discussed their perceptions of AI guardrails as ethical, technical, and procedural safeguards to uphold academic integrity in higher education. Several themes were identified and reported with verbatim quotes from participants.

Cybersecurity measures for identifying ethical risk and flagging non-compliance in the academic context

Panellists viewed the AI policy and guidelines as drivers of academic integrity and as tools to prevent non-compliance. A major issue raised by some panellists was the detection of AI-generated content. Panellists supported AI guardrails [Turnitin AI detection or Invigilator App] that

continuously scan outputs for ethical risks and policy violations. Some also indicated that these guardrails analyse linguistic patterns, metadata, and behavioural cues to identify anomalies or synthetic text, helping institutions flag potential misuse in teaching and assessment. Ethical risk indicators, such as biased language, privacy breaches, or unsafe reasoning, trigger automatic alerts to ensure the responsible use of AI. When non-compliance is detected, the system triggers enforcement actions aligned with academic integrity and governance policies. This integrated approach enhances transparency, upholds ethical standards, and fosters the adoption of trustworthy AI in higher education. Most participants expressed support for policy guidelines related to upholding academic integrity and preventing non-compliance.

Another panellist (PM2) mentioned that she had implemented preventive measures to determine which AI technologies are allowed and which are not for her module assignments.

> *Students were informed that, for one of my assignments, no AI technology is allowed. I could easily identify that the AI-generated content in the assignment was manipulated and non-compliant.*

Several panellists agreed that teaching students about AI detection and raising awareness are key strategies for dealing with the misuse of AI technologies. Panellist (PM3) responded to non-tolerance in his course:

> *In my postgraduate course, students were introduced to and trained on the use of AI technologies. During my timed examination, students must use the AI detection App to complete it. There is no place for ignorance. I believe that the more you are exposed to AI-generated content, the more quickly you will be able to identify ethical risks and safety flags in content. We are trained to increase compliance by monitoring potential risks, but we need to enforce policies to protect our qualifications.*

Following PM3's comments, higher education institutions are seeing the value of combining provenance systems with required academic integrity

courses for students. The academic integrity course was well received, and students understood its importance for responsible AI use, ethical decision-making, and dealing with misuse. Academic integrity courses, AI guardrails, better provenance systems, and structured training all help build a culture of accountability, honesty, and trustworthy academic practice.

Another panellist (PM1) echoed sentiments about AI-generated content. This is what he alluded to:

> *AI detection is now much easier for us, enabling better source traceability of generated AI content. Which specific AI tools, like ChatGPT, Jenni.ai, or Copilot, were used? For me, the transparency and explainability of AI-generated content are potential indicators of academic dishonesty.*

Furthermore, most of the panellists believed that it is in the best interests of all at the faculty to adopt ethical governance to support academic practices. Panellist (PM1) said:

> *I am sure that provenance and structured training in AI detection tools, along with an ethical governance management strategy to foster accountability, honesty, transparency, and trustworthy academic practices, are essential.*

With reference to RQ3, panellists faced challenges in implementing AI guardrails to ensure responsible and compliant academic practices. The following issues emerged from the critical reflections.

Policy-governance misalignment and ineffective enforcement of AI guardrails in the academic context

Most often, panellists viewed the absence or weak governance of policy enforcement and ineffective institutional systems as creating uncertainty and misalignment in the implementation of AI safety measures. This is a major concern in the quest for academic integrity. Panellists noted that the absence of clear, consistent policies complicates efforts to regulate AI use and uphold academic standards. Additionally, widespread ignorance, misinformation, and limited guidance on AI safety contribute to confusion and inconsistent

practices across departments. These gaps heighten concerns about how to address unethical academic behaviours, including dishonesty and irresponsible use of AI. Consequently, ensuring responsible engagement with AI technologies and promoting compliant academic practices remain ongoing challenges, requiring stronger governance, clearer communication, and more robust institutional support mechanisms.

Several panellists (PM2 & PM3) raised concerns about the lack of awareness, inconsistent policy directives and irresponsible use of AI. A panellist highlighted:

> *that some of our policies are differently interpreted by lecturers. Some lecturers and students ignored policy guidelines.*

Panellist (PM5) opines:

> *For all of us, institutional AI policies, oversight committees, inclusion in codes of conduct, and academic integrity policies are crucial measures to combat the misuse of AI technologies.*

A limited understanding of how an AI guardrail functions, combined with low levels of AI literacy, makes it harder for lecturers to interpret system results, enforce ethical standards, and guide students to use AI responsibly. This knowledge gap weakens efforts to maintain academic integrity and creates uncertainty about how to apply institutional AI policies, making it challenging for lecturers. Moreover, the inconsistent detection accuracy, integration difficulties, and frequent system constraints increase complexity and reduce trust in these tools. Such challenges hinder the smooth implementation of AI guardrails as reliable mechanisms for supporting ethical compliance and academic integrity. (White, 2023) Another serious concern was whether panel members were compromising security. This panellist (PM1) stressed his views during the critical conversation:

> *I noticed one of our lecturers sent the username and password to the support staff in our department. How can that be allowed, and is this not a breach of ICT security or a compromise of the very essence of data protection?*

Most panellists identified several challenges in implementing AI guardrails. These included weak policy enforcement, inconsistent guidance from institutions, and uncertainty about when to apply sanctions, all of which caused confusion and made it harder for lecturers to interpret guardrail results or spot unethical student behaviour. Technical complexity, system limitations, and unreliable detection accuracy also reduced people's trust in AI tools.

5. Discussion

The findings from the SRQ1 reveal that panellists conceptualise AI guardrails primarily as a cybersecurity-oriented protective mechanism designed to safeguard information, regulate user conduct, and detect unethical practices within AI-mediated environments. Across the narratives, AI guardrails are consistently framed not merely as technical controls but as integrated socio-technical systems that combine data protection, behavioural monitoring, and ethical oversight. Panellists emphasised the protective and preventative function of AI guardrails. As noted in the verbatim account by PM3, *"AI safeguards are designed to guide, protect, and restrict or monitor unethical behaviour. It is for protecting information, data, and monitoring the unethical behaviour of users of AI technologies."* This articulation reflects a dual-layered understanding: first, guardrails act as data security mechanisms (Molina-Carmona & García-Peñalvo, 2025); second, they serve as tools for behavioural regulation (Atmakuri, 2025). This duality aligns with contemporary perspectives in cybersecurity and responsible AI, in which safeguarding systems are expected to address both technical vulnerabilities and human misuse (Makanto & Eze, 2022). Similarly, PM1 highlighted the role of AI guardrails in controlling access to sensitive information, stating that they *"secure data by controlling access to sensitive information, thereby preventing exposure or misuse… [and] detect unethical practices, misinformation, or academic dishonesty."* This view reinforces the notion that guardrails operate through access control and surveillance mechanisms, ensuring that only authorised users can access sensitive data while also identifying deviations from ethical norms (AbdelRahman et al., 2024). The emphasis on detecting misinformation and academic dishonesty is

particularly significant in educational contexts, where AI tools can be misused to compromise academic integrity. Critically, the findings suggest that participants do not perceive AI guardrails as passive or static controls, but rather as active monitoring systems capable of identifying and responding to unethical behaviour in real time. This reflects an understanding of AI governance that extends beyond compliance to include continuous oversight and adaptive intervention **(**Oluoha et al., 2022). However, while this perspective highlights the strengths of AI guardrails, it also raises important concerns that excessive monitoring may lead to a culture of control, potentially undermining academic freedom (Hermanowicz, 2025). It is argued that AI guardrails are not only tools for protecting data but also mechanisms for upholding ethical standards in AI use. In academic settings, this is particularly relevant, as the misuse of AI technologies can manifest as plagiarism, fabrication, and information manipulation.

Second, SRQ2 indicate that panellists strongly position AI policies, guidelines, and detection mechanisms as central drivers in upholding academic integrity and preventing non-compliance in higher education. AI guardrails are not viewed in isolation, but as part of a broader governance ecosystem that integrates policy enforcement, detection technologies, provenance systems, and student training (Floridi et al., 2018; Sharma, 2025; Xui et al., 2023). A dominant issue raised by panellists concerns the detection of AI-generated content, which is perceived as both a growing challenge and an evolving capability. Panellists expressed support for AI guardrails such as Turnitin AI detection and the Invigilator App, which *"continuously scan outputs for ethical risks and policy violations."* These tools were described as leveraging linguistic pattern analysis, metadata tracking, and behavioural cues to identify anomalies indicative of synthetic or manipulated text (Garib & Coffelt, 2024; Haroon-Sulyman et al., 2024). Such capabilities reflect a shift toward algorithmic surveillance and automated compliance systems, where ethical risks, such as biased language, privacy violations, or unsafe reasoning, trigger alerts and potential enforcement actions (Boutadjine et al., 2025; Ryan & Stahl, 2021).

Critically, this reliance on detection technologies signals a growing institutional emphasis on accountability and transparency, yet it also raises questions about the accuracy and reliability of AI detection tools (Atmakuri,

2025; Kaouni et al., 2025). While panellists expressed confidence in these systems, the literature cautions that AI detection is not always definitive and may yield false positives or fail to detect sophisticated AI-assisted writing. Panellists further emphasised the role of explicit policy enforcement and preventative measures. For example, PM2 described implementing clear restrictions: "for one of my assignments, no AI technology is allowed." This illustrates a compliance-driven approach, with boundaries clearly defined to minimise ambiguity in AI use (Gianni et al., 2022; Mahrishi et al., 2025). However, the ability to "easily identify… manipulated and non-compliant" content also suggests reliance on educator expertise alongside technological tools.

This reinforces the importance of pedagogical awareness and professional judgement in identifying misuse. In addition, PM3 emphasised a zero-tolerance stance, supported by training and controlled assessment environments, stating that students were required to use the Invigilator App during timed examinations. His assertion that "there is no place for ignorance" reflects a strong belief in exposure and training as mechanisms for compliance. The idea that increased familiarity with AI-generated content enhances the ability to detect ethical risks highlights the importance of AI literacy among both students and lecturers (Chan, 2023; Langlois et al., 2023). However, such strict enforcement approaches may risk creating a compliance culture driven by surveillance, rather than fostering intrinsic ethical responsibility (Chan, 2023; Evangelista, 2025).

A significant contribution of this theme is its emphasis on AI provenance as a complementary mechanism to detection. Provenance systems, which track the origin, lineage, and authenticity of AI-generated content, were seen as strengthening transparency and aligning outputs with ethical standards (Boutadjine et al., 2025; Clark et al., 2025). As PM1 noted, "AI detection is now much easier for us for source traceability… Which specific AI tools… were used?" This reflects a growing expectation that AI systems should provide traceable and explainable outputs, enabling lecturers to verify authenticity and identify potential misconduct. Importantly, panellists linked these technological measures to educational interventions, particularly academic integrity courses (Bretag, 2018; Memarian & Doleck, 2023). These courses were well received by students and helped them understand ethical

decision-making, the responsible use of AI, and the consequences of misuse (Zlotnikova et al., 2025). This suggests that effective governance requires not only enforcement mechanisms but also capacity-building initiatives that cultivate a culture of integrity (Bretag, 2018; Mabhoko, 2025). Measures such as structured training, provenance systems, and AI guardrails collectively promote accountability, honesty, and transparency (Harjika, 2026; Mishra, 2025).

Finally, panellists emphasised the importance of ethical governance frameworks at the faculty level. As articulated by PM1, the integration of provenance, training, and detection tools forms *"an ethical governance management strategy to foster accountability, honesty, transparency and trustworthy academic practices."* This highlights a shift from fragmented interventions toward a holistic governance approach that aligns policy, technology, and pedagogy to support responsible AI use.

The findings from SRQ3 reveal that ineffective institutional systems significantly undermine the implementation of AI safety measures, creating uncertainty, inconsistency, and misalignment in efforts to uphold academic integrity. Panellists consistently identified the absence of clear, coherent, and consistently applied policies as a central challenge. Without unified institutional direction, lecturers and students interpret policies differently, leading to fragmented practices and weakened enforcement (Gupta et al., 2025; Hermanowicz, 2025). As one panellist noted, *"some of our policies are differently interpreted by lecturers. Some lecturers and students ignored policy guidelines."* This highlights a critical gap between policy formulation and policy enactment, where ambiguity fosters non-compliance rather than accountability.

A major concern emerging from the data is the lack of awareness and widespread misinformation regarding AI safety and ethical use. Panellists (PM2 & PM3) pointed to limited guidance and inconsistent communication as factors contributing to cross-departmental confusion. This environment of uncertainty complicates efforts to regulate AI use effectively and to address unethical academic behaviours, such as dishonesty and the misuse of AI tools (Davis, 2025; Eaton, 2023). The findings suggest that in the absence of clear institutional messaging, individuals rely on personal interpretations that may not align with broader academic integrity frameworks. Panellists also

emphasised the importance of institutional governance structures, such as oversight committees and the integration of AI policies into codes of conduct. As PM5 opined, *"institutional AI policies, oversight committees, inclusion in codes of conduct, and academic integrity policies are crucial measures to combat the misuse of AI technologies."* This reflects an understanding that formal governance mechanisms are necessary to standardise practices, ensure accountability, and provide a foundation for consistent enforcement (Carney, 2022; Floridi et al., 2018; Floridi, 2019; Nikolinakos, 2023). However, the mere existence of such structures is insufficient without effective implementation, monitoring, and communication. Another critical issue identified is the limited AI literacy among lecturers, particularly regarding how AI guardrails function (Emiri et sl, 2024; Ngoveni, 2025).

Therefore, a lack of understanding of AI guardrail systems and their outputs constrains lecturers' ability to interpret results, uphold ethical standards, and guide students responsibly. This knowledge gap weakens the effectiveness of AI interventions and contributes to uncertainty in the application of institutional policies. Importantly, this finding underscores that technological solutions cannot compensate for insufficient human capacity; rather, they require informed users who can critically engage with AI outputs. The technical complexity and limitations of AI guardrails further exacerbate these challenges (Agarwal & Nene, 2025; Mishra, 2025). Panellists reported issues such as inconsistent detection accuracy, integration difficulties with existing systems, and frequent technical constraints. These limitations undermine trust in AI tools and hinder their adoption as reliable mechanisms for ensuring ethical compliance. The perception that detection systems may be unreliable or difficult to use creates reluctance among lecturers, ultimately weakening the intended impact of these technologies on academic integrity.

A particularly concerning issue raised relates to basic cybersecurity practices and potential security breaches. PM1 highlighted a case in which login credentials were shared with support staff, asking: *"How can that be allowed, and is this not a breach of ICT security?"* This example illustrates that, beyond sophisticated AI guardrails, fundamental security practices remain vulnerable, potentially undermining the integrity of the entire system (Gupta et al., 2025; Hermanowicz, 2025). It also reflects a broader issue of institutional culture and awareness, where lapses in basic digital ethics can

compromise more advanced security measures (Gupta et al., 2025; Mishra, 2025). Furthermore, panellists pointed to weak policy enforcement and uncertainty around sanctions as significant barriers. The lack of clarity regarding when and how to apply consequences for non-compliance creates hesitation and inconsistency among lecturers. This not only undermines the credibility of institutional policies but also signals to students that enforcement may be negotiable or uneven. Such inconsistencies weaken the overall goal of fostering a culture of integrity and responsible AI use.

Implications

To use AI guardrails effectively in higher education, strong policies must be complemented by practical steps. Policies should be clear, consistent, and flexible, and should include ethical principles such as transparency, fairness, accountability, and respect for user autonomy. They should also match academic integrity and data protection standards, balancing oversight with privacy. Regular policy reviews and stakeholder involvement are important for keeping policies relevant and building trust in the institution.

In practice, effective use of AI guardrails depends on strong institutional support. Universities should focus on improving AI literacy for students, staff, and faculty, and offer professional development to encourage responsible AI use in teaching and assessment. Investing in secure technologies, like provenance and detection systems, is also important. Creating a culture of ethical responsibility through open communication and shared accountability is more effective than relying solely on punishment.

6. Conclusion

Although panellists held differing views on AI guardrails, they agreed that these tools are essential for protecting systems, securing information, monitoring unethical behaviour, and ensuring institutions adhere to their own policies. AI guardrails are seen as integrated systems that safeguard sensitive data, guide user behaviour, and detect unethical actions, combining cybersecurity with ethical governance. The findings underscore the need for better governance, clearer communication, thorough training, and strong institutional support to reduce cyber risks. If these areas are not addressed, AI

guardrails may perform less effectively or consistently, potentially undermining ethical compliance and academic integrity (Alsharefeen & Sayari, 2025). Drawing on the study and existing research, a systematic literature review will be conducted to develop a comprehensive AI ethics framework for higher education that strengthens ethical governance and responsible AI use. The study also highlights the value of combining qualitative and quantitative research to better understand and evaluate AI guardrails, with long-term, participatory approaches helping to develop effective governance frameworks.

7. Bibliography

Abdel Rahman, A. A., Abbas, H. H., Alhamadani, B. T. R, Ahmed, R. K. A, Majeed, Y. (2024). In Iryna Savelieva "Decoding Personal Security–Strategies to Safeguard Humans in the Era of Intelligent Machines," 2024 36th Conference of Open Innovations Association (FRUCT), Lappeenranta, Finland, 2024, 3–12.

Ademeso, T., Uloma, E., Yusuf, L., & Abdulaziz, I. (2025). The role of good governance in enhancing effective public management decision-making: Challenges and prospects. African Journal of Politics and Administrative Studies,18(1), 425-448.

Agarwal, A., & Nene, M. J. (2025). A five-layer framework for AI governance: integrating regulation, standards, and certification. Transforming Government: People, Process and Policy,19(3), 535-555. https://doi.org/10.1108/TG-03-2025-0065

Atmakuri, N. B. (2025). Democratizing Advanced Technology: How Edge AI in Embedded Systems Transforms Everyday Life. Journal of Multidisciplinary, 5(8), 253-263.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. Interacting with computers, 23(1), 4-17. doi: 10.1016/j.intcom.2010.07.003

Boutadjine, A., Harrag, F., & Shaalan, K. (2025). Human vs machine: A comparative study on the detection of AI-generated content. ACM Transactions on Asian and Low-Resource Language Information Processing, 24(2), 1-26. https://doi.org/10.1145/3708889

Brannick, M. T., Chan, D., Conway, J. M., Lance, C. E., & Spector, P. E. (2010). What is method variance, and how can we cope with it? A panel discussion. Organisational research methods,13(3), 407-420. https://doi.org/10.1177/1094428109360993

Bretag, T. (2018). Academic integrity. In Oxford Research Encyclopedia of Business and Management. https://doi.org/10.1093/acrefore/9780190224851.013.147

Carney, R. (2022). Reimagining our futures together: a new social contract for education, Comparative Education, 58:4, 568-569. https://doi.org/10.1080/03050068.2022.2102326

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International journal of educational technology in higher education, 20(1), 1-38. https://doi.org/10.1186/s41239-023-00408-3

Clark, H. B., Benton, L., Searle, E., Dowland, M., Gregory, M., Gayne, W., & Roberts, J. (2025, July). Building effective safety guardrails in AI education tools. In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED. AIED 2025. Springer, Cham. https://doi.org/10.1007/978-3-031-99261-2_12

Davis, R. (2025). CheatGPT crisis – SA universities faced with a burgeoning degree of AI-written academic assignments. Daily Maverick (5 April 2025). Johannesburg. https://www.dailymaverick.co.za/article/2025-04-05-cheatgpt-crisis-sa-universities-faced-with-a-burgeoning-degree-of-ai-written-academic-assignments

DiMaggio, P. J., & Powell, W. W. (2000). The iron cage revisited: institutional isomorphism and collective rationality in organisational fields. In JAC Baum & F Dobbin (Eds.). Economics meets sociology in strategic management (pp. 143-166). Emerald Group Publishing Limited. https://doi.org/10.1016/S0742-3322(00)17011-1

Dev, J., Akhuseyinoglu, N. B., Kayas, G., Rashidi, B., & Garg, V. (2025). Building guardrails in AI systems with threat modelling. Digital Government: Research and Practice,6(1), 1-18. https://doi.org/10.1145/3674845

Eaton, S. E. (Ed.). (2023). Handbook of academic integrity. Second Edition. Springer Nature.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity,19(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Emiri, O. T., Ijiekhuamhen, O. P., & Nwankwo, W. (2024). Digital Literacy Among Lecturers in the Age of Artificial Intelligence: A Case Study. Delta Journal of Computing, Communications & Media Technologies, 1. 76-90. https://doi. org/10. xxx.

Evangelista, E. D. L. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. Contemporary Educational Technology,17(1), 1-19. https://doi.org/10.30935/cedtech/15775

Farrokhnia, M., Soleimani, S., & Noroozi, O. (2025). Generative AI in higher education: Transformative tools for research, teaching, and assessment. In Navigating Generative AI in Higher Education (pp. 33-53). Edward Elgar Publishing. https://doi.org/10.4337/9781035337873.00007

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for

a good AI society: Opportunities, risks, principles, and recommendations. Minds and machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. Philosophy & Technology,32(2), 185-193. https://doi.org/10.1007/s13347-019-00354-x

Garib, A., & Coffelt, T. A. (2024). DETECTing the anomalies: Exploring implications of qualitative research in identifying AI-generated text for AI-assisted composition instruction. Computers and Composition,73, 102869. https://doi.org/10.1016/j.compcom.2024.102869

Gianni, R., Lehtinen, S., & Nieminen, M. (2022). Governance of responsible AI: From ethical guidelines to cooperative policies. Frontiers in Computer Science, 4, 873437. https://doi.org/10.3389/fcomp.2022.873437

Giray, L., Sevnarayan, K., & Ranjbaran Madiseh, F. (2025). Beyond policing: AI writing detection tools, trust, academic integrity, and their implications for college writing. *Internet Reference Services Quarterly*, *29*(1), 83-116.https://doi.org/10.1080/10875301.2024.2437174

Giray, L. (2024). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. The Serials Librarian, 85(5-6), 181-189. https://doi.org/10.1080/0361526X.2024.2433256

Gupta, R., Song, Q., Wagner, M., Engström, E., Söderberg, E., Borg, M., & Runeson, P. (2025, November). AI alignment for ethical compliance and risk mitigation in industrial applications. In International Conference on Product-Focused Software Process Improvement (pp. 20-35). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-12089-2_2

Harjika, R. (2026). Guardrails for Innovation: Governance, Ethics, and Responsible AI. In " Architecting Enterprise AI Strategies: From

Vision to Scalable Execution (pp. 203-253). Berkeley, CA: Apress. https://doi.org/10.1007/979-8-8688-2219-3_7

Haroon-Sulyman, S. O., Kamaruddin, S. S., & Ahmad, F. K. (2024). Text Anomaly Detection Advancements, Challenges and Pathways: A Systematic Literature Review. Journal of Logistics, Informatics and Service Science,11(6), 198-218. https://doi.org/10.33168/JLISS.2024.0612

Hermanowicz, J. C. (2025). Internal Threats to Academic Freedom: Problems of Professional Control. European Review. 2025;33(S1):44-55. https://doi.org/10.1017/S1062798725000171

Jonas, M. (2024). Mitigating Use of Artificial Intelligence in Student Assignments. Journal of Computing Sciences in Colleges,39(8), 173-181

Kaouni, M., Lakrami, F., Laouidya, O., & Baddi, Y. (2025). Higher education and generative artificial intelligence: Applications, challenges, opportunities, and ethics. In Baddi, Y., Maleh, Y., Alsmadi, I., & Lahby, M. (Eds.). Generative AI for Cybersecurity and Privacy (1st ed.). (pp.187-204). CRC Press. https://doi.org/10.1201/9781003597476

Kumar, A. (2026). Designing Guardrails: Ensuring Responsible AI Behaviour. In: Singh, S.,et al.GenAI and LLMs for Beyond 5G Networks (pp.107-144). Springer, Cham. https://doi.org/10.1007/978-3-032-06418-9_5 Cham: Springer Nature Switzerland.

Langlois, L., Dilhac, M. A., Dratwa, J., Ménissier, T., Ganascia, J. G., Weinstock, D. & Marchildon, A. (2023). Ethics at the Heart of AI.Obvia: Montreal, QC, Canada. https://www.obvia.ca/sites/obvia.ca/files/ressources/202310-OBV-Pub-EthiqueCoeurIA-EN_0.pdf

Mabhoko, M. (2025). Governing Intelligent Systems: Global Frameworks, Implementation Gaps, and Pathways to Protect Humanity.

Implementation Gaps, and Pathways to Protect Humanity (November 28, 2025). https://dx.doi.org/10. 2139/ssrn.5975274

Macfarlane, B., Zhang, J., & Pun, A. (2014). Academic integrity: A review of the literature. Studies in Higher Education, 39(2), 339–358. https://doi.org/10.1080/03075079.2012.709495

Mahrishi, M., Abbas, A., & Siddiqui, M. K. (2025). Global initiatives towards regulatory frameworks for artificial intelligence (AI) in higher education. Digital Government: Research and Practice,6(2), 1-9. https://doi.org/10.1145/3672462

Makanto, P. K., & Eze, J. S. (2022). Mitigating Human Vulnerabilities in Cybersecurity: Understanding Human Flaws and Implementing Effective Countermeasures. Dept. Comput. Informatics, Bournemouth Univ., Bournemouth, UK.

Marchant, T., Gavian, S., & Eriksson, J. (2001). Panel Discussion: Quantitative and Qualitative. In Feinstein, O.N. (2001). Evaluation and Poverty Reduction (1st ed.). (pp. 89-109. Routledge. https://doi.org/10.4324/9781351325325

Memarian, B., & Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence,5, 10 0152. https://doi.org/10.1016/j.caeai.2023.100152

Memarian, B., & Doleck, T. (2025). Education with a systems theory and control perspective. Education and Information Technologies, 1-18. https://doi.org/10.1007/s10639-024-13223-8

Meyer, J. W., & Rowan, B. (1977). Institutionalised organisations: Formal structure as myth and ceremony. American Journal of Sociology, 83(2), 340-363. https://doi.org/10.1086/226550

Mishra, A. (2025). Understanding AI guardrails: Concepts, models, and methods. International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences,13, 1-7.

Mogoale, P. D., Pretorius, A., Mogase, R. C., & Segooa, M. A. (2025). Integrating artificial intelligence within South African higher learning institutions. South African Journal of Information Management, 27(1), 1939. https://hdl.handle.net/10520/ejc-info_v27_n1_a1939

Molina-Carmona, R., & García-Peñalvo, F.J. (2025). Safeguarding Knowledge: Ethical Artificial Intelligence Governance in the University Digital Transformation. In: Vendrell Vidal, E., Cukierman, U.R., Auer, M.E. (eds) Advanced Technologies and the University of the Future. (pp. 201-220). Lecture Notes in Networks and Systems, vol 1140. Springer, Cham. https://doi.org/10.1007/978-3-031-71530-3_14

Mulahuwaish, A., El-Khoury, M., Qolomany, B., Bou Abdo, J., & Zeadally, S. (2025). Does AI need guardrails?. International Journal of Pervasive Computing and Communications, 21(2), 177-186. https://doi.org/10.1108/IJPCC-07-2024-0224

Nasiri, E., & Khojasteh, L. (2024). Evaluating panel discussions in ESP classes: an exploration of international medical students' and ESP instructors' perspectives through qualitative research. BMC Medical Education,24(1), 925. https://doi.org/10.1186/s12909-024-05911-3

Nasution, P. T., & Fransiska, W. (2026). AI-Generated Assignments: Lecturers' Perspectives on Linguistic Integrity and Pedagogical Strategies in Higher Education. Utamax: Journal of Ultimate Research and Trends in Education,8(1), 14-29. https://doi.org/10.31849/h1yr9h06

Ngoveni, M. (2025). Bridging the AI knowledge gap: The urgent need for AI literacy and institutional support. The International Journal of Technologies in Learning,32(2), 83. https://doi.org/10.18848/2327-0144/CGP/v32i02/83-100

Nikolinakos, N.T. (2023). Ethical Principles for Trustworthy AI. In: NT Nikolinakos. EU Policy and Legal Framework for Artificial

Intelligence, Robotics and Related Technologies - The AI Act. Law, Governance and Technology Series, 53, (pp. 101-166). Springer, Cham. https://doi.org/10.1007/978-3-031-27953-9_3

OECD (2019). An OECD Learning Framework 2030. In: Bast, G., Carayannis, E.G., Campbell, D.F.J. (eds) The Future of Education and Labour. Arts, Research, Innovation and Society. Springer, Cham. https://doi.org/10.1007/978-3-030-26068-2_3

Oluoha, O. M., Odeshina, A., Reis, O., Okpeke, F., Attipoe, V., & Orieno, O. H. (2022). Artificial intelligence integration in regulatory compliance: A strategic model for cybersecurity enhancement. Journal of Frontiers in Multidisciplinary Research,3(1), 35-46.

Ryan, M., & Stahl, B. C. (2021). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. Journal of Information, Communication and Ethics in Society,19(1), 61-86. https://doi.org/10.1108/JICES-12-2019-0138

Selwyn, N. (2023). Digitalisation of Education in the Era of Climate Collapse and Planetary Breakdown. In B Williamson, J Komljenovic & K Gulsom. World Yearbook of Education 2024(pp. 261-275). Routledge.

Sharma, A. (2025). PPO-based Reinforcement Learning with Human Feedback with Hybrid Oversight and Predictive Reward Evaluation for AGI. Journal of Future Artificial Intelligence and Technologies,2(3), 493-503. https://doi.org/10.62411/faith.3048-3719-276

Slimi, H., & Chichti, J. (2026). The Nexus Between Cybercrime and Irregular Migration in the Age of Cyberspace: Implications for Geographic Flexibility and Mobility Management. Global Journal of Flexible Systems Management, 1-38. https://doi.org/10.1007/s40171-025-00474-8

Sohail, S., Parveen, S., & Dar, T. (2025). Investigating the Role of Generative AI in Transforming Teaching, Learning, and Assessment Practices. Journal of Applied Linguistics and TESOL (JALT),8(4), 159-174.

Tembo T. (2026). Minister Solly Malatsi's withdrawal of draft national AI policy for its fabricated sources 'embarrassing and damning'. Independent Online Webpage (27 April 2026). Minister Solly Malatsi's withdrawal of draft national AI policy for its fabricated sources 'embarrassing and damning'

UNESCO, P. (2021). Reimagining our futures together: A new social contract for education. Paris, France: Educational and Cultural Organisation of the United Nations.

UNESCO's work is recognised as one of the best AI initiatives with social and ethical impact. (2023, April 20). UNESCO. https://www.unesco.org/en/articles/unescos-work-recognized-among-best-ai-initiatives-social-and-ethical-impact

Williamson, B., Eynon, R., & Potter, J. (2020). Pandemic Politics, Pedagogies, and Practices: Digital Technologies and Distance Education during the Coronavirus Emergency. Learning, media and technology,45(2), 107–114. https://doi.org/10.1080/17439884.2020.1761641

Xu, W., Agrawal, S., Briakou, E., Martindale, M. J., & Carpuat, M. (2023). Understanding and detecting hallucinations in neural machine translation via model introspection. Transactions of the Association for Computational Linguistics,11, 546-564. https://doi.org/10.1162/tacl_a_00563

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart learning environments,11(1), 28. https://doi.org/10.1186/s40561-024-00316-7

Zlotnikova, I., Hlomani, H., Mokgetse, T., & Bagai, K. (2025). Establishing ethical standards for GenAI in university education: a roadmap for

academic integrity and fairness. Journal of Information, Communication and Ethics in Society,23(2), 188-216. https://doi.org/10.1108/JICES-07-2024-0104

8. Short biography

Professor Micheal M. van Wyk is a Full Professor of Economics Education at UNISA and an NRF C2-rated researcher. His scholarship focuses on flipped pedagogy, cooperative learning, ODeL, self-directed learning, and generative AI in higher education. With extensive experience across schooling and academia, he has published widely on digital pedagogies and curriculum innovation. An award-winning author and distinguished scholar, he contributes significantly to teacher education, research supervision, and the advancement of ethical AI integration in higher education.

Email: vanwykm4@gmail.com