

FROM TRANSPARENCY TO ACCOUNTABILITY: A FRAMEWORK FOR ETHICAL AND EMPOWERING USE OF GENERATIVE AI IN HIGHER EDUCATION

Matthew Montebello

Department of Artificial Intelligence, University of Malta (Malta)

Abstract

The integration of Generative Artificial Intelligence in higher education marks a pivotal moment in redefining academic practice. While early responses often focused on restriction and detection, emerging evidence suggests that empowering educators and learners through ethical frameworks yields more sustainable and reflective engagement. This paper presents a five-part pedagogical framework, Transparency, Human–AI Collaboration, Contextualised AI Literacy, Institutional Responsibility, and Accountability-Based Assessment, designed to guide responsible GenAI adoption in teaching, learning, and research. Drawing on pre- and post-workshop survey data collected from academic staff at the University of Malta, the study examines shifts in educators’ attitudes, confidence, and ethical reasoning following structured GenAI literacy interventions. Findings reveal significant increases in reflective use, transparency in disclosure, and openness to assessment reform when supported by clear institutional guidance. Two new recommendations, Reflective Transparency and Accountability Over Avoidance, extend the framework to emphasise that learners should be evaluated not for AI use itself, but for their capacity to justify, verify, and critically engage with it. The paper argues that the future of higher education lies in moving from prohibition to participation, from automation to co-agency, and from compliance to ethical creativity. By embedding these five guidelines into curricula and policy, institutions can transform GenAI from a disruptive challenge into an instrument of empowerment, scholarship, and integrity.

Keywords: *Generative AI, academic integrity, ethical pedagogy, assessment reform, AI literacy.*

1. Introduction

The rapid diffusion of Generative Artificial Intelligence (GenAI) tools, such as ChatGPT, Claude, and Google Gemini, into higher education has precipitated a profound shift in academic workflows, particularly in research, writing, teaching, and assessment. Unlike earlier forms of educational AI that focused on adaptation, analytics, or prediction, GenAI systems are capable of generating fluent text, explanations, feedback, and multimodal artefacts that closely resemble human-authored academic work. This qualitative shift has intensified longstanding questions around authorship, epistemic agency, and the role of the human scholar in AI-mediated learning environments (Zawacki-Richter et al., 2019; Knox, 2020; Miao & Holmes, 2023). In response to this rapid technological uptake, many higher education institutions have adopted precautionary or defensive strategies. Early institutional reactions have been characterised by bans on GenAI use, reliance on automated detection tools, and policy discourse framed primarily around academic misconduct and plagiarism (Cotton et al., 2023; Luo, 2024). While such responses are understandable in the face of uncertainty, they risk conflating ethical concerns with technological determinism and may inadvertently discourage transparency, reflection, and responsible engagement. Empirical and conceptual work increasingly suggests that detection-oriented approaches are fragile, inequitable, and pedagogically misaligned, particularly given the documented ease with which GenAI outputs can bypass detection systems (Perkins et al., 2024). A growing body of human-centred AI scholarship argues that GenAI should instead be framed as a pedagogical partner rather than a substitute for human cognition or judgement. From this perspective, the educational value of GenAI does not reside in automation or efficiency alone, but in its capacity to scaffold reflection, critique, and dialogic engagement when embedded within intentional pedagogical design (Biesta, 2020; Cope & Kalantzis, 2017; Montebello, 2020). Recent work has emphasised the importance of human-in-the-loop approaches, where educators retain epistemic authority and learners are supported to critically interrogate, refine, and justify AI-generated outputs rather than passively consume them. Despite this emerging consensus, there remains

a notable gap in the literature: a lack of integrated, practice-oriented frameworks that move beyond principles and policies to offer concrete guidance for educators and institutions. In particular, limited attention has been paid to how transparency, accountability, AI literacy, and assessment design can be aligned to empower learners without penalising responsible technology use. Much of the existing guidance focuses on acceptable use and disclosure, but offers little insight into how learning, authorship, and assessment validity can be preserved when cognitive engagement is no longer structurally enforced by traditional assessment mechanisms. This paper addresses this gap by proposing and empirically grounding a five-part framework for the ethical and empowering use of GenAI in higher education. The framework comprises: (1) Transparency in use, (2) Human–AI collaboration, (3) Contextualised AI literacy, (4) Institutional responsibility, and (5) Accountability-based assessment. The framework is informed by pre- and post-workshop survey data collected from academic staff participating in GenAI professional development initiatives at the University of Malta, complemented by qualitative reflections emerging from these engagements. The findings provide evidence of a shift from fear-based avoidance toward reflective, accountable, and pedagogically meaningful GenAI adoption when educators are supported by clear ethical guidance and assessment-aligned practices.

The remainder of this paper is structured as follows. Section 2 situates the study within relevant theoretical and pedagogical traditions, including constructivist learning, human-centred AI, and assessment-for-learning. Section 3 outlines the methodological design of the GenAI workshops and survey instruments. Section 4 presents and discusses the empirical findings. Section 5 elaborates the proposed five-part framework, with illustrative examples drawn from practice. The paper concludes by discussing implications for educators, institutions, and policy makers seeking to move from avoidance to accountability in the age of Generative AI.

2. Theoretical and pedagogical background

The integration of GenAI into higher education represents not merely a technological development, but a pedagogical and epistemic transformation. Unlike earlier Artificial Intelligence in Education (AIED) systems, GenAI actively participates in meaning-making through language generation, explanation, and feedback, reshaping how academic knowledge is produced, evaluated, and legitimised (Montebello, 2025a). Constructivist and sociocultural learning theories provide a critical lens through which to interpret this transformation. Learning is conceptualised as an active process of interpretation and sense-making rather than passive content acquisition (Cope & Kalantzis, 2017). Biesta’s (2020) notion of world-centred education further emphasises judgement, responsibility, and subjectivity as central educational outcomes, privileging human engagement over performative efficiency. Within these traditions, assessment functions as a mechanism for eliciting cognitive engagement and reflective judgement. GenAI disrupts this alignment by enabling learners to produce academically acceptable artefacts without necessarily engaging in the cognitive processes traditionally associated with learning (Montebello, 2025b). This decoupling shifts engagement from cognitive necessity to cognitive choice, a structural consequence of technological affordances rather than learner misconduct. Human-centred approaches respond by reasserting human agency through educator-in-the-loop design, transparency, and accountability, positioning AI outputs as objects of critique rather than proxies for thinking. Assessment theory reinforces this position. Assessment-for-learning scholarship emphasises process, feedback, and metacognition over product (Black & Wiliam, 1998; Nicol & Macfarlane-Dick, 2006). However, many contemporary assessment practices remain artefact-focused and thus vulnerable to unreflective GenAI use. Dialogic and oral assessment approaches offer more resilient alternatives by foregrounding explanation, justification, and interaction. Policy guidance from UNESCO and the European Union increasingly reflects these concerns, emphasising transparency, ethical literacy, and human oversight (European Commission, 2024; Miao & Holmes, 2023). Nevertheless, such guidance often lacks pedagogically actionable models. This paper builds on human-centred AI, constructivist pedagogy, and assessment theory to address this gap through an integrated, practice-grounded framework.

3. Methodology

This study adopts a practice-based, interpretivist research design to examine how structured pedagogical interventions can support ethical and empowering use of Generative Artificial Intelligence in higher education. Rather than treating GenAI adoption as a purely technical or behavioural phenomenon, the methodology is grounded in the assumption that educators’ beliefs, confidence, and pedagogical intent play a decisive role in shaping how AI is integrated into teaching, learning, and assessment. This approach aligns with human-centred AI scholarship, which emphasises situated practice, reflexivity, and educator agency as core dimensions of responsible GenAI use in education. The study was conducted at the

University of Malta between 2024 and 2025 as part of a series of professional development workshops focused on GenAI in teaching, research, and assessment. Participants comprised academic staff from a wide range of disciplines, including the arts and humanities, social sciences, life sciences, physical sciences, and technology-related fields. This disciplinary diversity was intentional, reflecting the premise that GenAI literacy and ethical use must be contextualised rather than treated as a generic skill. Participation was voluntary, and workshops were advertised institutionally as exploratory and developmental rather than evaluative. This framing was critical in fostering openness, reducing fear of surveillance, and encouraging honest engagement with ethical concerns around GenAI use. Each workshop followed a structured yet flexible design informed by constructivist and inquiry-based pedagogy. Sessions typically combined conceptual grounding in human-centred GenAI principles, including transparency, educator-in-the-loop design, and learner agency. Hands-on engagement with GenAI tools (Gemini, ChatGPT, Claude) was an integral part of every session, applied to authentic academic tasks such as literature synthesis, feedback generation, assessment redesign, and reflective writing. This was followed by a guided discussion and reflection on ethical implications, authorship, assessment validity, and institutional responsibility. Rather than prescribing ‘correct’ uses of GenAI, the workshops deliberately foregrounded critical interrogation of AI outputs, comparative prompting, and justification of decisions. This design choice aligns with prior work emphasising that GenAI should function as a pedagogical partner rather than an outsourcing mechanism for cognition or judgement. Data were collected using a pre- and post-workshop survey design, complemented by qualitative reflections generated during and after workshop activities. The pre-workshop survey captured baseline attitudes toward GenAI, including levels of confidence, perceived risks, ethical concerns, and beliefs about assessment and academic integrity. The post-workshop survey repeated core items while also probing changes in perception, perceived pedagogical value, and willingness to adopt transparent and accountable GenAI practices. Survey instruments employed Likert-scale items and open-ended questions to capture both measurable shifts and nuanced reflections. Particular attention was paid to attitudes toward transparency, disclosure of AI use, and assessment fairness, areas that later informed the development of the accountability-based recommendations. In addition, anonymised qualitative data were drawn from reflective prompts, group discussions, and written comments produced during workshop activities. These reflections provided insight into how educators reasoned about GenAI use, how tensions between innovation and integrity were negotiated, and how assessment practices were reimagined in light of AI capabilities. Quantitative survey data were analysed descriptively to identify pre-post shifts in confidence, ethical awareness, and pedagogical orientation. Qualitative data were analysed thematically using an inductive approach, focusing on recurring patterns related to transparency, agency, assessment design, and institutional responsibility. These themes were then mapped against existing human-centred GenAI principles and assessment theory to inform the development of the five-part framework. Importantly, the purpose of the analysis was not to generalise statistically, but to empirically ground a pedagogical framework in authentic academic practice. In this sense, the methodology serves as a bridge between theory and application, providing evidence that ethical and empowering GenAI integration is both pedagogically feasible and professionally transformative when supported by intentional design.

4. Findings and discussion

Analysis of the pre- and post-workshop survey data, supported by qualitative reflections, reveals a consistent shift in how academic staff conceptualised Generative AI in relation to teaching, learning, and assessment. Prior to the workshops, participants commonly expressed uncertainty and concern, particularly around academic integrity, assessment validity, and the perceived loss of control over authorship. GenAI was frequently framed as a disruptive or risky technology, with several respondents indicating that avoidance or restriction appeared to be the safest institutional response. Post-workshop data, however, indicates a marked change in orientation. Participants reported increased confidence in engaging with GenAI, greater clarity regarding ethical use, and stronger alignment between GenAI practices and pedagogical intent. Rather than viewing GenAI primarily as a threat, many academics began to frame it as a pedagogical resource—provided its use was transparent, contextualised, and accountable. This shift supports existing human-centred GenAI research, which emphasises that educator confidence and ethical literacy are prerequisites for responsible integration. A dominant qualitative theme concerned transparency. Participants consistently reported that structured discussion of AI disclosure reduced anxiety and reframed transparency as a scholarly practice rather than a confession of misconduct. Several academics noted that once transparency was normalised, conversations could move beyond detection toward learning design and assessment quality. This finding aligns with broader critiques of surveillance-driven integrity models and supports the argument that openness fosters ethical engagement more effectively than prohibition. A second theme related to assessment and accountability. Initially, many participants assumed that GenAI use would

inevitably undermine assessment fairness. Following the workshops, however, there was growing agreement that students should not be penalised for GenAI use per se, but rather evaluated on their ability to justify, explain, and critically engage with AI-assisted outputs. This reframing reflects a broader pedagogical shift from product-oriented assessment toward process-oriented and dialogic forms of evaluation, including reflective and oral assessment practices. Participants also highlighted the importance of contextualised AI literacy, noting that ethical use cannot be meaningfully taught in abstraction from disciplinary norms and assessment practices. Generic AI policies were widely perceived as insufficient without concrete pedagogical models that educators could adapt to their own contexts. Taken together, the findings suggest that ethical GenAI integration is less a technical problem than a pedagogical one. When educators are supported through structured professional development, reflective dialogue, and principled frameworks, fear-based responses give way to intentional, accountable practice. These findings empirically ground the five-part framework presented in the next section, positioning it as a response not only to technological change, but to the evolving educational responsibilities of educators and institutions in AI-mediated environments.

5. A five-part framework for ethical and empowering GenAI integration

Building on the theoretical foundations outlined in Section 2 and the empirical findings presented in Section 4, this section introduces a five-part framework designed to support the ethical and empowering integration of Generative Artificial Intelligence in higher education. The framework is intentionally pedagogical rather than prescriptive, offering design principles that educators and institutions can adapt to their specific disciplinary, institutional, and assessment contexts. Its central premise is that ethical GenAI use is best achieved through transparency, accountability, and intentional learning design rather than restriction or surveillance. Transparency constitutes the foundational principle of the framework. Rather than treating disclosure of GenAI use as an admission of misconduct, transparency is reframed as a scholarly practice aligned with academic reflexivity and methodological integrity. Educators and learners are encouraged to explicitly acknowledge where, how, and why GenAI tools have been used in the production of academic work. Findings from the workshops indicate that transparent disclosure reduces anxiety and fosters trust when decoupled from punitive consequences. Participants reported that reflective AI-use statements enabled more productive discussions about authorship, epistemic contribution, and learning processes. This aligns with human-centred GenAI principles that position transparency and explainability as pedagogical affordances rather than compliance mechanisms. In practice, transparency can be operationalised through short AI-use declarations, reflective appendices, or methodological statements that mirror existing academic conventions such as ethics disclosures. The second guideline emphasises that GenAI should augment rather than replace human cognition and judgement. Learners remain epistemically responsible for interpreting, validating, and contextualising AI-generated outputs. GenAI is positioned as a dialogic partner that can support idea generation, surface alternative perspectives, and challenge assumptions, but not as a substitute for scholarly reasoning. Workshop participants highlighted the pedagogical value of comparative prompting, critical interrogation of AI responses, and iterative refinement of outputs. These practices reinforce the role of the learner as an active decision-maker and align with educator-in-the-loop design approaches that prevent over-reliance on automation. Importantly, this guideline reinforces that responsibility does not diminish with tool use, rather, GenAI increases the need for human judgement, particularly in contexts requiring nuance, interpretation, and ethical reasoning. The third guideline recognises that GenAI literacy cannot be meaningfully taught as a generic or decontextualised skill. Ethical and effective GenAI use is inherently discipline-specific, shaped by epistemic norms, assessment practices, and professional expectations. What constitutes responsible AI use in creative writing, legal reasoning, scientific reporting, or data analysis varies considerably. Participants consistently rejected one-size-fits-all AI policies, noting that ethical engagement requires contextual understanding. Contextualised AI literacy therefore involves embedding critical engagement with GenAI within disciplinary curricula, assessment tasks, and scholarly practices. Educators play a crucial role in modelling discipline-appropriate GenAI use and in making evaluative criteria explicit. This approach aligns with constructivist pedagogy and human-centred AI design, both of which emphasise situated judgement over procedural compliance. The fourth guideline shifts attention from individual behaviour to institutional responsibility. Ethical GenAI integration cannot be sustained through individual effort alone, it requires coherent institutional frameworks that align policy, pedagogy, assessment, and professional development. Workshop findings revealed strong demand for institutional clarity, not in the form of restrictive rules, but through shared principles, exemplars, and pedagogical support. Without such alignment, transparent and innovative GenAI practices remain uneven and risky. Institutions therefore have a responsibility to legitimise ethical GenAI use, provide professional development opportunities, and ensure that policies support accountability rather than fear-based compliance. The final guideline addresses

assessment directly and represents a critical extension of the framework. Rather than penalising GenAI use, assessment should prioritise accountability, justification, and cognitive engagement. Participants increasingly converged on the view that students should not be marked down for using GenAI, provided they can explain, defend, and critically reflect on their work. Accountability-based assessment shifts the evaluative focus from whether GenAI was used to how it was used. Practical strategies include reflective commentaries, annotated submissions, oral explanations, and transparency statements that make student reasoning visible. Such approaches restore alignment between learning and assessment while remaining compatible with responsible GenAI use. This guideline reinforces that academic integrity in AI-mediated contexts is demonstrated through accountable engagement rather than tool avoidance. Together, these five guidelines form an integrated framework that moves higher education from avoidance to accountability, from detection to design, and from automation to agency. They reassert that ethical GenAI integration is achieved not by restricting tools, but by cultivating transparent, reflective, and accountable scholarly practice grounded in human judgement and pedagogical intent.

6. Implications, recommendations, and conclusion

The framework proposed in this paper demonstrates that the challenges posed by GenAI in higher education are primarily pedagogical and organisational rather than technical. Restrictive and detection-driven responses risk undermining transparency, trust, and learning, whereas accountability-based approaches foreground reflection, agency, and educational integrity. For educators, this entails a shift toward modelling transparent GenAI use and designing learning tasks that require justification, critique, and reflective engagement. Assessment practices must evolve from product-oriented evaluation to process-focused accountability. Institutions play a critical role by aligning policy, pedagogy, and professional development around human-centred principles that legitimise ethical GenAI use. In conclusion, academic integrity in the age of GenAI is best preserved through accountability rather than avoidance. By integrating transparency, contextualised AI literacy, and assessment redesign, GenAI can be repositioned as a catalyst for pedagogical renewal rather than a threat to scholarly values.

References

- Biesta, G. (2020). *World-centred education: A view for the present*. Routledge.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
- Cope, B., & Kalantzis, M. (2017). *E-learning ecologies: Principles for new learning and assessment*. Routledge.
- European Commission (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L1689. European Parliament & Council of the European Union. Brussels.
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.
- Montebello, M. (2025a) Educating the Educators: Guidelines for Ethical and Empowering Use of Generative AI in Higher Education. Proceedings of the 9th Int. Conf. on Modern Research in Education, Teaching, and Learning (ICMETL).
- Montebello, M. (2025b). Generative AI and the Human Research Scholar. *18th International Conference on Education, Research and Innovation, ICERI25*. (str. 901-909). Seville, Spain.: ICERI.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–221.