This open-access article is distributed under Creative Commons licence CC-BY-NC 4.0.

Suspect until proven human? Rebuilding trust in health professions education in the age of generative artificial intelligence

The introduction of generative artificial intelligence in the public space occurred several years back, and yet we still marvel at or fear it (or, likely, a bit of both). Its use in education is also not novel, albeit new generative artificial intelligence-supported strategies are continuously being developed. Unsurprisingly, in the background, something else changed in our classrooms. A coherent paragraph may now trigger suspicion; an improvement is scanned with a dubious eye, rather than celebrated. Students are asked to disclose their use of generative artificial intelligence, yet disclosure may also be treated less as transparency than as confession, breeding further questions. If honest with myself, I face the same thoughts, and I trust others do too: *how much of this work is my student's, how much did the generative artificial intelligence contribute, and was this used appropriately to still ensure that learning took place?*

The learner may quietly become a suspect, and the educator an investigator. Substituting undisclosed generative artificial intelligence-produced work for a student's own thinking is misconduct. In health professions education, graduates must develop core competencies aligned to their responsibilities to serve their patients, contribute to national needs, and grow within their careers. Standards must be protected. The question though is whether a mistrust- or misconduct-first response protects them, or merely creates an adversarial culture while leaving students and staff poorly prepared to use generative artificial intelligence responsibly.

Several studies have highlighted broad use of generative artificial intelligence in education to support writing, explain concepts, and propose research ideas.[1-3] However, alongside such use we see fear emerge from students about use or the disclosure of use: hesitance due to repercussions, unclear rules or policies, and inconsistent enforcement.[1,2,4] Institutions ask for honesty, but students may experience honesty as self-incrimination, or be ignorant as to what constitutes a disclosure.

Trust is a core value that humans expect from one another, and that constitutes a principle that allows our communities to move forward. Trust can also be fragile: easy to break, and difficult to rebuild. The International Center for Academic Integrity identifies it as a fundamental value alongside honesty, fairness, respect, responsibility and courage.[5] Carless[6] showed that distrust and accountability pressures can constrain learning-orientated assessment and become self-reinforcing, highlighting its impact within our educational ecosystems. A community of trust is therefore not merely a kinder academic culture; it is part of the curriculum. Our graduate outcomes necessitate such levels of professionalism that nurture trust: transparency, ethical judgement, accountability, and the courage to disclose uncertainty. Students learn these attributes partly through how universities exercise authority, and how it is role modelled within our engagements. So, a question I often reflect upon: *can academia convincingly demand professional trustworthiness when we model suspicion as if it is a response?*

Trustworthiness develops through practice, not assertion. Students need opportunities to disclose limitations, explain reasoning, receive feedback and correct errors. These routines mirror clinical professionalism, where safe care depends on admitting uncertainty, reporting mistakes, seeking assistance and documenting decisions. A community that treats declaration as suspicious may teach concealment, or assumes that non-disclosure hides true use; one that responds proportionately teaches that transparency is compatible with accountability. Institutional culture is therefore a form of professional socialisation: what universities role-model in assessment shapes how graduates may later exercise judgement under scrutiny. A community is also reciprocal; staff must explain boundaries, acknowledge their own generative artificial intelligence use, and correct inconsistent practice.

Generative artificial intelligence-detection software was touted as a way to bridge uncertainty of use, but has also been found to be inaccurate or unreliable.[7] Detector output is not proof, constituting a signal for a humane, evidence-based conversation and due process. Various universities have publicly disabled such detection software,[8] or quietly disbanded it. For example, the University of Cape Town's framework cautions against overemphasising misconduct, warns that detector errors can damage trust, and acknowledges that support is needed on all fronts.[9] However, many conversations on the ground in higher education still vie for the ability to detect generative artificial intelligence use, but such actions may also instigate mistrust or false accusations that can damage reputation, or incur an administrative rigamarole of (dis)proving use.

Surveillance also creates a troubling hidden curriculum relating to role modelling and perceived hypocrisy. Educators increasingly use generative artificial intelligence for administrative, educational or research purposes, which may beg similar concerns to authenticity, vetting of information, intellectual ownership and transparency, although procedures remain overtly fixed on students. A large-scale and ironic example of those in positions of power falling victim to declaration and vetting concerns, although not academic, is the retraction of the South African artificial intelligence policy due to hallucinated references.[10] Such large-scale errors in positions of authority may create a perceptual lens of power dictating the conversation, rather than responsible use. If students must disclose generative artificial intelligence use, must educators not do the same within the learning ecosystem?

Notable efforts have been placed on shifting the discussion to authentic, generative artificial intelligence-safe assessments. Assessment must support defensible conclusions about learners' competence,[11] so quality assurance reference points are necessary. Generative artificial intelligence has exposed the fragility of a polished product as a potential surrogate for quality of learning. A mere prohibition or declaration box to such tasks is likely to be insufficient, necessitating structural changes to promote more validity.[12] Assessment strategies have already positioned staged submissions, brief oral explanations, practical demonstrations, context-rich cases and portfolios as the future mechanism for learning and competence judgement without requiring invigilation to control for generative artificial intelligence use.

Assessment should distinguish independent from generative artificial intelligence-supported competence. Where unaided reasoning, communication or procedural skill is essential, secure generative artificial intelligence-free assessment is justified. Where practitioners will use generative artificial intelligence, assessment should test whether students can protect confidential information, interrogate outputs, verify evidence,

identify bias, document use, and accept responsibility for decisions. These are increasingly recognised as competencies and patient-safety obligations.[13,14] As such, students consequently need explicit generative artificial intelligence literacy, not only prohibited-tool lists. They should understand why plausible outputs may be false or biased; what information must not enter public systems; how claims are verified; when use requires declaration; and when generative artificial intelligence should be set aside. However, often such explicit training is not found in programmes, begging the question that if we leave it on the sideline, can we truly expect that generative artificial intelligence literacy and ethics will develop?

A trust-centred institution would co-create guidance with students; state task-level boundaries clearly; teach literacy before demanding compliance; make disclosure routine and non-punitive when use is permitted; expect reciprocal transparency from staff; treat detector output as an unreliable indicator rather than a verdict; and base claims of competence on multiple, meaningful sources of evidence. Fair consequences remain necessary for substantiated misconduct, but their legitimacy depends on clear expectations, credible evidence and an opportunity to respond.

Although difficult to operationalise, these discussions are not new. The choice is not between trust and standards. Trust without verification can be naive; verification without trust may become surveillance. If we build curricula and assessments with an underlying fear or assumption of cheating, students will learn concealment. If we build an academic community around transparency, judgement and responsibility, students may model the trustworthiness expected of health professionals. Our task is not to police learners into authenticity, but to role-model and cultivate a professional culture worthy of trust.

Declaration of generative artificial intelligence use. Now, a pertinent question: *did I use generative artificial intelligence to write this paper?* Why, yes, I did – invariably important to disclose use these days. Claude Opus 4.8 was used for soundboarding, language review, alternative stances, and critical review via a mixture of chain-of-thought, contextual, Socratic and iterative prompting. However, the thoughts surrounding the majority of points are attributed to various reflections from recent informal discussions with friends, colleagues, collaborators, and conference attendees. I hope that you trust me that this was done judiciously, ethically, and appropriately.

Werner Cordier

Department of Pharmacology, School of Medicine, Faculty of Health Sciences, University of Pretoria, South Africa; African Institute for Data Science and Artificial Intelligence (AfriDSAI), University of Pretoria, South Africa

werner.cordier@up.ac.za

1. Freeman J. Student Generative AI Survey 2025. HEPI Policy Note 61. Oxford: Higher Education Policy Institute, 2025. https://www.hepi.ac.uk/wp-content/uploads/2025/02/HEPI-Kortext-Student-Generative-AI-Survey-2025.pdf (accessed 30 June 2026).
2. Smit M, Wagner RF, Bond-Barnard TJ. Ambiguous regulations for dealing with AI in higher education can lead to moral hazards among students. Project Leadersh Soc 2025;6:100187. https://doi.org/10.1016/j.plas.2025.100187
3. Baigi SFM, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: A systematic review. Health Sci Rep 2023;6(3):e11138. https://doi.org/10.1002/hsr2.1138
4. Gonsalves C. Addressing student non-compliance in AI use declarations: Implications for academic integrity and assessment in higher education. Assess Eval High Educ 2025;50(4):592-606. https://doi.org/10.1080/02602938.2024.2415654
5. International Center for Academic Integrity. The Fundamental Values of Academic Integrity. 3rd ed. Clemson, SC: International Center for Academic Integrity, 2021. https://academicintegrity.org/aws/ICAI/asset_manager/get_file/911282?ver=1 (accessed 30 June 2026).
6. Carless D. Trust, distrust and their impact on assessment reform. Assess Eval High Educ 2009;34(1):79-89. https://doi.org/10.1080/02602930801895786
7. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, et al. Testing of detection tools for AI-generated text. Int J Educ Integr 2023;19(1):26. https://doi.org/10.1007/s40979-023-00146-z
8. University World News. South African universities move beyond AI detection tools. 27 June 2026. https://www.universityworldnews.com/post.php?story=20260627115042773 (accessed 30 June 2026).
9. University of Cape Town. UCT Framework for AI in Education. Cape Town: UCT, 2025. https://cilt.uct.ac.za/sites/default/files/media/documents/cilt_uct_ac_za/2486/uct-ai-in-education-framework-june-2025-final.pdf (accessed 6 July 2026).
10. Simpson K, Keetsi N, Shondlani L. The withdrawal of South Africa's Draft Artificial Intelligence Policy: An opportunity? 4 May 2026. https://www.dlapiper.com/en-au/insights/publications/2026/05/withdrawal-of-south-africa-draft-ai-policy (accessed 7 July 2026).
11. Dawson P, Bearman M, Dollinger M, Boud D. Validity matters more than cheating. Assess Eval High Educ 2024;49(7):1005-1016. https://doi.org/10.1080/02602938.2024.2386662
12. Corbin T, Dawson P, Liu D. Talk is cheap: Why structural assessment changes are needed for a time of GenAI. Assess Eval High Educ 2025;50(7):1087-1097. https://doi.org/10.1080/02602938.2025.2503964
13. Masters K, MacNeil H, Benjamin J, et al. Artificial intelligence in health professions education assessment: AMEE Guide No. 178. Med Teach 2025;47(9):1410-1424. https://doi.org/10.1080/0142159X.2024.2445037
14. Rodger D, Mann SP, Earp BD, Savulescu J, Bobier C, Blackshaw BP. Generative AI in healthcare education: How AI literacy gaps could compromise learning and patient safety. Nurse Educ Pract 2025;87:104461. https://doi.org/10.1016/j.nepr.2025.104461