CRITICAL / PRACTICE NOTES

Stop Detecting AI. Start Assessing Scholarship.

Rethinking academic misconduct, fairness and assessment in the age of AI

NO. 06 / NOVEMBER 2026 ROY HANNEY READING TIME 6 MIN

The provocation

We have become remarkably good at spotting poor AI-assisted writing. Generic argumentation, fabricated references, superficial analysis and suspiciously formulaic prose can all attract attention. But what exactly are we identifying: academic misconduct, or simply students who are not very good at using AI?

A student who uses AI clumsily may face an allegation and the burden of an investigation. Another student who uses AI skilfully, revises extensively or conceals its involvement may submit work without attracting suspicion. Meanwhile, independent student writing can be mistaken for AI-generated text. All the while, detection-led enforcement risks rewarding concealment while penalising visible weaknesses in AI literacy.

Where AI is permitted, poor AI-assisted work should ordinarily be assessed as poor scholarship. Where AI is prohibited but an assessment remains readily completable using AI, the institution also has an assessment-design problem. Neither situation is solved by trying to infer a text's origins from its style.

The critical question

Are we protecting academic integrity, or selectively penalising students whose AI use is easiest to notice? What would change if we assessed scholarship first and redesigned assessments that cannot establish the intended learning?

What the research suggests

Weber-Wulff et al. (2023) tested 14 AI-text detection tools against 54 texts of known provenance, including human writing, machine-translated writing, AI-generated writing and AI text modified through editing or paraphrasing. Across 756 tests, no tool reached 80% overall accuracy under the authors' strict classification method; the strongest reached 76%. Edited and paraphrased AI text was much harder to identify (pp. 9–13). These results concern tools tested in 2023, not every current version, but they challenge the use of detector scores as evidence about an individual student.

Kofinas et al. (2025) investigated what happens when academics themselves attempt that judgement. Eight markers at two UK business schools assessed student-authored, AI-modified and AI-generated submissions against existing rubrics. Their ability to identify AI involvement varied substantially, with pair-level judgements ranging from 33.3% to 85.7% accuracy (pp. 2528–2534). The authors also report that suspicion of AI could affect markers' approaches to evaluating work. This small exploratory study does not establish how frequently such effects occur across higher education, but it exposes a tension between assessing quality and guessing authorship.

Bassett et al. (2026) move from technical reliability to procedural fairness. They argue that detector scores, stylistic hallmarks, comparisons with earlier writing and repeated testing do not independently establish a student's use of AI (pp. 2–7). More strikingly, they suggest that detection may identify students' limited skill at concealing AI involvement, or their unequal access to stronger tools, rather than consistently identifying misconduct (p. 9). This is a critical policy argument, not a direct empirical measurement of who is referred for investigation.

The papers expose a connected problem: classification is fallible; academic suspicion is inconsistent; and the resulting procedures can distribute scrutiny unevenly. They do not show that every apparently weak AI-

assisted submission is misconduct, nor that replacing detection with ordinary marking would itself solve assessment validity. That final step is a pedagogical proposition requiring judgement and, where necessary, redesign.

An approach to try

Before referring a submission because it "looks like AI", revisit the assessment brief and rubric. The aim is not to relax scholarly standards but to apply them consistently, while ensuring that the conditions of assessment actually support the capabilities being judged.

01 REVIEW — Audit the rubric. Identify how existing criteria address the weaknesses commonly associated with poor AI use: unsupported claims, shallow reasoning, inaccurate sources, generic analysis and failure to exercise disciplinary judgement. These are assessable scholarly deficiencies regardless of how the text was produced.

02 CLARIFY — Make the rules and intended capabilities explicit. Tell students which forms of AI assistance are permitted, what must be acknowledged, and what evidence of learning the task requires. Do not make students infer boundaries from inconsistent advice or retrospectively imposed expectations.

03 ASSESS — Mark what the work demonstrates. Where AI is permitted, use the existing criteria to evaluate argument, evidence, critical engagement and scholarly quality. Do not turn polished prose, weak prose or an AI-detector score into proof of unauthorised use. Independently evidenced fabrication or breaches of clearly stated rules remain matters for established procedures.

04 REDESIGN — Where AI is prohibited, or a polished submission cannot establish the intended learning, alter the task or the conditions in which it is completed. Consider supervised performance, staged or observed work, or opportunities for students to explain and defend their decisions. A prohibition unsupported by assessment design transfers an institutional problem onto individual students.

Where AI fits — and where it does not

AI can help students draft, interrogate arguments, locate weaknesses and revise work. Some students will use these capabilities more effectively than others. Academic assessment should not mistake that difference in proficiency for reliable evidence of dishonesty. Equally, a strong final product does not automatically demonstrate that a student has developed the intended independent capability: that is a question of assessment validity, not detection.

This is not an argument for ignoring dishonesty or guaranteeing that every AI-assisted submission deserves a passing mark. It is an argument for distinguishing poor scholarship from demonstrable misconduct, and for expecting institutions to design assessments they can meaningfully evaluate.

Questions for adaptation

- Which weaknesses in AI-assisted submissions are already covered by our existing marking criteria?
- When do our assessment rules invite suspicions that cannot be independently substantiated?
- If AI use is prohibited, what in the assessment conditions—not merely its wording—makes that boundary meaningful?

Three sources to explore

Bassett, M. A., Bradshaw, W., Bornsztejn, H., Hogg, A., Murdoch, K., Pearce, B., & Webber, C. (2026). Heads we win, tails you lose: AI detectors in education. Journal of Higher Education Policy and Management. https://doi.org/10.1080/1360080X.2026.2622146

Kofinas, A. K., Tsay, C. H.-H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. British Journal of Educational Technology, 56, 2522–2549. https://doi.org/10.1111/bjet.13585

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19, Article 26. https://doi.org/10.1007/s40979-023-00146-z

Method & provenance

This note developed from a researcher-led provocation concerning the unequal consequences of detecting visibly poor AI use. Four original research papers were supplied and critically compared; three were selected for their complementary empirical, marking and procedural contributions. AI assisted with reading, annotation, comparison and drafting; the selected papers' methods, principal findings and limitations were checked against the supplied full texts. The central selective-enforcement interpretation and four-part teaching approach are analytical and pedagogical propositions, not empirically validated interventions.

ABOUT THIS SERIES

Short, evidence-informed provocations that move from critical enquiry to adaptable teaching practice. Each note draws on a small, accessible reading list and makes the role of AI-assisted scholarship transparent.