

CONFLUENCE
Journal of Multidisciplinary Studies

e-ISSN: 3085-5411
Journal Homepage: https://journals.imist.ma/index.php/confluence/index

Generative AI and the Risk of Algorithmic Compliance in University Academic Integrity Policies

Elhadj Moussa BenMoussa
University of East London, London, United Kingdom

https://doi.org/10.34874/PRSM.cjms-vol2iss4.9922

Elhadj Moussa BenMoussa is affiliated with the University of East London. His research examines digital governance, education, and institutional power, with particular attention to academic integrity, algorithmic accountability, and equitable participation in higher education.

e.benmoussa@uel.ac.uk

https://orcid.org/0009-0001-2186-1403

Received: 28-04-2026 Accepted: 20-10-2026 Published: 30-10-2026

Generative AI and the Risk of Algorithmic Compliance in University Academic Integrity Policies

Elhadj Moussa BenMoussa
University of East London, London, United Kingdom

ABSTRACT

Generative artificial intelligence has made the relationship between academic integrity and institutional verification more difficult to settle. A declaration, an oral discussion, or a record of drafts can support learning, but it can also become a demand that students continually demonstrate their legitimacy. This article develops the concept of algorithmic compliance pedagogy through a critical reading of ten public documents from six universities in the United Kingdom, continental Europe, and North America, retrieved on 23 October 2026. The analysis combines reflexive thematic reading with attention to policy wording, the allocation of responsibility, and the grounds on which authorship is questioned. The documents do not support a uniform movement toward automated surveillance. Several explicitly reject AI detectors, while preserving disclosure requirements, authorship interviews, and instructor judgments about writing. Five themes examine conditional permission, restrictions on detection, professional discretion, documentary authenticity, and the relationship between compliance and educational purpose. The central concern is that suspicion can migrate from a detector score to judgments about style or the performance of a writing process. This is a policy risk rather than an observed effect on students. The article distinguishes necessary assessment assurance from disproportionate verification through its connection to learning, the burden of documentation, institutional responsibility, and opportunities to contest an inference. It argues that protecting standards requires universities to account for the grounds of suspicion as well as requiring students to account for their work.

KEYWORDS

Generative Artificial Intelligence

Academic Integrity

Higher Education

Algorithmic Governance

Pedagogical Trust

Multilingual Writers

Introduction

The arrival of generative artificial intelligence (GenAI) has made familiar questions about authorship harder to answer. A fluent assignment may reflect a student's understanding, extensive external assistance, or some combination of the two. Universities need evidence that their assessments represent the learning they claim to certify. They also need to explain how students can use the tools increasingly available within ordinary writing environments. These responsibilities meet in institutional policies, where the boundary between acceptable assistance and misconduct is translated into instructions about writing, disclosure, and proof.

The difficulty is that verification has an educational cost as well as an institutional purpose. Keeping drafts can help a student understand revision. The same requirement can become a precaution against being accused. An oral discussion can reveal understanding, while an

Contact ✉ e.benmoussa@uel.ac.uk

©2026

unexpected authorship interview can require a student to defend a written performance under different conditions. The question is what makes these practices proportionate to the learning being assessed, and what happens when avoiding suspicion becomes an additional condition of academic participation.

This article introduces algorithmic compliance pedagogy to examine that problem. The term describes a risk within AI governance: educational credibility becomes dependent on satisfying classifications of acceptable assistance or signs of supposedly human authorship, while the relationship between those classifications and the learning outcome receives insufficient scrutiny. It is not a name for every integrity rule or every request for an explanation. The concern arises when a platform judgment, a writing stereotype, or an excessive requirement for process evidence becomes more authoritative than a reasoned appraisal of the student's knowledge.

The distinction matters because current institutional guidance contains substantial resistance to automated suspicion. University College London (UCL) advises against detector use, Vanderbilt and Waterloo explain decisions to disable it, and ETH Zurich questions the reliability of technical recognition. A credible critique must account for these positions. The analysis therefore asks how suspicion and responsibility are organized when institutions restrict automated detection but retain other ways of examining authorship. The documents are identified in Appendix A and examined in the Findings section.

The equity concern remains central. Liang and colleagues' study of early GPT detectors found misclassification affecting non-native English writing. That evidence establishes a problem in the tools and samples tested; it does not establish that every contemporary detector disadvantages every multilingual writer. The wider question is whether a student's language becomes a reason to demand additional proof. Policy advice that treats formulaic wording or a change in style as a signal needs particular care, because the inference from textual appearance to unauthorized assistance is open to alternative explanations (Liang et al. 1-2; "Academic Integrity and Generative AI").

This article does not argue for unrestricted AI use, nor does it dismiss concerns about academic standards. It examines the conditions under which verification supports those standards without making students responsible for resolving every uncertainty about authorship. The contribution is a close account of how declarations, permissions, and evidentiary signals distribute obligations between students and institutions. Its claims concern the language of a bounded documentary corpus and the practices that language recommends. It does not claim to measure institutional implementation or students' experiences.

Literature Review

Policy Responses after Generative AI

Research on university AI policy already shows that institutional responses are varied. Jin and colleagues analyze adoption policies and guidance from forty universities across six global regions, identifying attention to academic integrity alongside teaching, training, and equitable access (3-6, 8). Li and colleagues screen 110 universities across three countries and identify 124 policy documents from 64 institutions that had issued guidelines. Their comparison demonstrates differences in policy priorities and governance arrangements (6, 9-11). These studies make a general account of universities responding only through prohibition difficult to sustain. They also establish a basis for examining how broad commitments become specific obligations in assessment.

Dotan, Parker, and Radzilowicz develop a faculty-informed approach to responsible adoption, drawing on collaborative work at the University of Pittsburgh. Their concern with academic freedom and institutional governance supports a discussion of who should decide how AI is used. A participatory process can make the reasons for a rule open to discussion, rather than leaving its educational purpose implicit. This approach directs attention to the authority behind an assessment requirement as well as the student conduct it seeks to regulate (Dotan et al., secs. 2, 5.1-5.2).

The change in students' use also needs careful interpretation. Freeman's 2025 HEPI survey reports extensive use among 1,041 full-time UK undergraduates. Its findings include substantial perceived policy clarity alongside limited institutional training, which complicates the assumption that guidance is uniformly absent or reactive. Stephenson and Armstrong's 2026 report, based on 1,054 UK undergraduates surveyed in December 2025, likewise places widespread use alongside uneven support and differing views about learning. These surveys provide context; they neither verify the practices of the six universities examined here nor show that using AI is equivalent to misconduct (Freeman 1-2; Stephenson and Armstrong 5-6, 9).

International guidance connects governance to educational responsibility. Miao and Holmes's 2023 UNESCO guidance advocates a human-centered approach to generative AI, including protection of agency, inclusion, and linguistic diversity. The 2026 European University Association report by Jørgensen and Phelan places human agency, institutional values, and participatory adoption at the center of its recommendations. These are normative resources for the present argument, rather than additional university cases or evidence that their recommendations have been implemented (Miao and Holmes 24-25; Jørgensen and Phelan 5, 9).

Detection Performance and Unequal Risk

Detector performance is an empirical question that changes with the tool, model, task, and text. Liang and colleagues compare native and non-native English writing and document a serious fairness problem in early detectors. Weber-Wulff and colleagues test multiple systems and show limitations in distinguishing human writing from generated text, including the effects of translation and obfuscation. These studies justify scrutiny of detection evidence. Their results should not be converted into a timeless error rate for every product (Liang et al. 1-2; Weber-Wulff et al. 26-27).

Perkins and colleagues' 2024 preprint examines how adversarial changes affect detector performance on a constructed test set. It extends the evidence on susceptibility to manipulation, but is identified as a preprint and is not used to estimate the frequency of misconduct in universities. A detector's performance on a constructed test set and the prevalence of prohibited assistance in student work are separate matters. The ethical concern arises when a technical classification is allowed to stand in for the second without adequate evidence (Perkins et al. 6-10, 16-17).

There is important contrary evidence. Van Vlasselaer and colleagues' 2026 study compares four tools on 160 documents with known authorship conditions (7-10). It reports comparatively strong performance by one tool and few false positives, with differences across fully generated, hybrid, and humanized texts. Its additional analysis of 1,163 theses lacks known ground truth and cannot establish a confirmed prevalence of AI use. The study supports attention to differences between tools and conditions; it challenges the assertion that poor performance or linguistic bias must be identical across all generations of detectors (Van Vlasselaer et al. 15-17).

Baker and Hawn's review locates algorithmic bias within the development and deployment of educational systems. Its relevance is broader than text detection: fairness depends on the

conditions in which a classification is used and the consequences attached to it. Even an improving detector does not decide whether the assistance was permitted, whether an accommodation applies, or whether the available evidence supports a misconduct finding. Those decisions require an assessment context and an accountable institutional process (Baker and Hawn 1054-57).

Trust and the Contribution of Policy Reading

Bearman, Ryan, and Ajjawi's critical review shows why AI should be examined through the discourses that give it educational meaning, rather than treated as a single neutral intervention. Slade and Prinsloo's earlier work on learning analytics similarly questions the assumption that collecting more information about a learner necessarily benefits that learner. These perspectives direct attention to the relationship between visibility, interpretation, and institutional power. They do not establish that every documentation practice is surveillance (Bearman et al. 379-81; Slade and Prinsloo 1510-11).

The edited collection *Higher Education for Good*, by Czerniewicz and Cronin, provides a further resource for thinking about educational purposes and collective responsibility. Its value here lies in sustaining a critical discussion of what universities should enable, alongside what they must regulate. The present article adds a focused question to that discussion: how can an institution make a student's work examinable without treating uncertainty about its production as a presumption against the student? The answer requires attention to the wording of the rule and to the safeguards accompanying it (Czerniewicz and Cronin).

Theoretical Framework

Discipline and the Demand for Visibility

Foucault's account of disciplinary power offers a way to examine how expectations of observation organize conduct. Applied to writing assessment, the relevant question is how a requirement to remain explainable might influence a student's preparation before any accusation occurs. The framework directs attention to draft retention, declarations, and the comparison of present work with previous writing. It does not establish what students actually feel or do in response to those requirements. Those effects would need evidence beyond policy documents (Foucault 200-03).

Biesta's account of qualification, socialization, and subjectification helps specify the educational concern. A university legitimately qualifies students, yet education also involves their capacity to exercise judgment and act responsibly. A verification practice can serve qualification while constraining other purposes, or support several purposes at once. An oral explanation, for example, can test understanding and invite reflection. Its value depends on the question asked, the conditions of the discussion and how the answer is interpreted (Biesta 77-79).

Williamson's analysis of datafication adds a question about authority: what changes when a record or classification becomes the means through which a learner is made institutionally knowable? The policy corpus does not demonstrate a complete transfer of judgment to platforms. Instead, it enables examination of how technical outputs, documentary evidence, and professional discretion are brought together. The distinction is necessary because a decision to reject a detector may protect academic judgment while leaving other classificatory assumptions intact (Williamson 12-14).

Defining Algorithmic Compliance Pedagogy

Algorithmic compliance pedagogy names a particular risk in this arrangement. It arises when satisfying rules or signals associated with machine authorship becomes an educational performance in itself, with insufficient attention to what the student understands. The algorithmic element concerns the influence of computational classification and expectations about generated text. The compliance element concerns the demand to remain institutionally legible. Where a practice is simply ordinary academic attribution, or a clearly justified assessment of competence, the concept should not be applied without further argument.

Three features warrant examination. First, does the institution rely on an inference about authorship that the evidence cannot adequately support? Second, does the requirement to explain a process exceed what is needed to assess the intended learning? Third, can the student question the inference and receive an explanation of the institutional decision? These questions make the concept open to challenge. A clear, limited declaration accompanied by accessible guidance may be evidence against the critique, rather than another instance of it.

The framework also distinguishes a policy invitation from an observed outcome. Instructions can make a form of conduct reasonable, expected, or risky within an assessment. They do not prove that teachers enforce the instruction consistently or that students become compliant performers. The findings therefore analyze the conditions that documents establish for judgment. The discussion considers the educational consequences as possibilities requiring further investigation, while retaining the importance of protecting academic standards.

Methodology

Corpus and Selection

The study uses qualitative critical policy document analysis to examine how institutions describe acceptable AI assistance and the evidence through which authorship may be questioned. The corpus consists of ten publicly accessible, English-language documents from six named universities: University College London and the University of Oxford in the United Kingdom; Utrecht University and ETH Zurich in continental Europe; and Vanderbilt University and the University of Waterloo in North America. The country coverage is the United Kingdom, the Netherlands, Switzerland, the United States, and Canada.

The ten documents were retrieved on 23 October 2026 and assigned identifiers P1–P10. They include one university assessment policy, institutional guidelines, student and library advice, a teaching resource, and dated notices explaining detector withdrawal. These genres are distinguished in Appendix A; a recommendation on a teaching webpage is not assumed to have the same status as a formal assessment policy. Dated notices and visible update markers range from August 2023 to July 2026. Several pages are undated, and the latest timestamp on a containing guide does not establish when every section was revised.

Selection was purposive and directed toward documents that made explicit statements about assessment permission, disclosure, detection, professional responsibility, or authorship checks. Searches combined institution names with terms including generative AI, assessment, academic integrity, detection, declaration, and oral interview. Eligible items had to be publicly accessible on an institutional site and relevant to assessment or academic integrity after the public emergence of ChatGPT. The selected set deliberately includes restrictions on detection as well as instructions for verification. It does not claim to represent the distribution of policy positions across universities.

The document, rather than the institution or individual webpage section, is the counting unit. UCL contributes three items, Oxford one, Utrecht one, ETH Zurich two, Vanderbilt two, and Waterloo one. This gives four documents in the United Kingdom group and three in each of the other geographic groups. Multiple pages from one university are analytically related, not independent observations. The sample is weighted toward research-intensive institutions with accessible English-language guidance; no claim of institutional-type diversity or thematic saturation is made.

Analytical Procedure

The analysis uses a reflexive thematic document reading informed by Braun and Clarke's later account of reflexive thematic analysis (Braun and Clarke). Theory informed the initial questions, but the themes were revised in response to contradictory passages. Close reading attended to modal instructions, the allocation of responsibility, and the circumstances in which a concern could become an authorship inquiry. This makes the interpretation examinable without treating themes as discoveries that exist independently of the reading.

Three questions guided the reading: how is acceptable assistance defined; what evidence can justify questioning authorship; and which obligations fall on students, teachers, and institutions? Initial codes included permission tied to an assignment, disclosure, draft retention, detector restriction, oral verification, judgments about writing style, teacher discretion, procedural safeguards, and pedagogical support. These codes were grouped into five themes concerning permission, detection, professional authority, process evidence, and educational purpose. Restrictions on detection were treated as counterevidence to an account of detection preceding dialogue.

For each document, the record links the source, relevant heading or clause, code, and thematic contribution. A second comparative reading considered whether an interpretation survived the safeguards and qualifications elsewhere in the same document. Table 1 shows the route from codes to themes and identifies every contributing document. Appendix B records the document-to-theme assignments. Counts refer to documents used to develop a theme, including counterevidence; they are not estimates of institutional prevalence, implementation, or the strength of a finding.

Table 1
Initial Codes and Contributions to the Five Themes

Theme	Initial codes grouped in the theme	Contributing documents	Interpretive qualification
T1 Integrity as suspicion and conditional permission	Task-specific permission; disclosure; expectations of legitimate authorship	P1, P2, P4, P6, P7, P9 (n = 6)	Permission and clear instructions qualify an interpretation of presumed suspicion.
T2 Detection before dialogue reconsidered	Detector restriction or withdrawal; privacy; dialogue; negative cases	P1, P2, P4, P7, P8, P9, P10 (n = 7)	These documents mainly challenge detector-led governance; they do not establish implementation.

Theme	Initial codes grouped in the theme	Contributing documents	Interpretive qualification
T3 Teacher authority and the limits of delegation	Instructor decision rights; style signals; oral inquiry; procedural limits	P1, P4, P5, P6, P7, P8, P9 (n = 7)	Judgment remains with educators, whose inferences still require support and review.
T4 Performative authenticity and process evidence	Draft retention; authorship interviews; declarations; reflective process accounts	P2, P3, P4, P5, P7, P9 (n = 6)	Process evidence can support learning or verification; its effect is not measured.
T5 Compliance as learning and institutional responsibility	Learning purpose; institutional duties; access; participation in rules	P1, P4, P5, P6, P7, P8, P9, P10 (n = 8)	The documents establish competing expectations, not observed obedience or learning gains.

Note. The counts include documents that challenge or qualify a theme. They describe this analysis and are not institutional prevalence estimates.

Scope and Credibility

Examples of policy becoming recommended practice are drawn from the identified documents. UCL's student guidance describes investigatory vivas; its library guidance recommends retaining drafts. Utrecht's teaching resource distinguishes formative conversation from authorship checking. Vanderbilt's guidance recommends staged assignments and personal reflection while also discussing warning signals. Turnitin's own report guide is used as contextual vendor evidence, outside the ten-document corpus. These sources establish what is recommended or stated publicly; they do not establish how often a practice occurs.

Credibility rests on named sources, visible quotations, locators, and an explicit record of counterexamples. The analysis is interpretive and does not claim independent coding, inter-rater reliability, or validation through student testimony. Generative AI assisted source discovery, organization, and drafting during revision, as declared below. The corpus and coding matrix make the resulting claims available for examination and correction. A reader can disagree with a code assignment or the proposed connection between a clause and an educational risk.

There are limitations. Institutional policy documents often present idealized visions rather than everyday academic practice, and public guidance may conceal internal disagreement or uneven implementation across departments. This study therefore does not claim to capture the lived experience of students or staff. It examines the formal expectations through which integrity is described and authorship is made examinable. Interviews, observations, and analysis of actual decisions would be needed to investigate how these expectations affect trust, workload, or the distribution of suspicion.

The corpus is cross-sectional. A dated withdrawal notice and a current guidance page do not constitute a longitudinal record of institutional change. English-language publication and public accessibility shape what can be examined, and the sample does not represent the Global South or all forms of higher education. A missing statement on a selected page is not evidence that an institution lacks the relevant protection elsewhere. These limits also constrain the conclusions drawn from the five themes.

Findings

The corpus shows a tension between making work accountable and limiting unjustified suspicion. It does not substantiate a uniform replacement of trust by detection. The five themes below examine that tension through the wording of named documents. Each theme includes protections or qualifications that constrain the interpretation, and the numerical counts describe the documents contributing to it rather than the number of institutions endorsing a single position.

Integrity as Suspicion and Conditional Permission

UCL's three-category framework distinguishes assessment in which AI is prohibited, assistive, or integral. Its own description, "not a formal policy," matters: the categories support communication, while assessment instructions determine what a student may do. Read alongside the student guidance, this positions authorship as a contextual relationship between assistance, the task, and the student's responsibility. It does not make all assisted work presumptively dishonest ("Three Categories"; "Engaging").

Oxford places obligations on assessment setters as well as students. Staff must define permission, align the task and criteria with it, and specify disclosure. The requirement for "equality of baseline provision of appropriate AI tools" makes the conditions of access part of the institutional responsibility. The analytical point is that an integrity rule can constrain the university's conduct as well as the student's. A demand for transparency is less defensible if the institution has not first made the permitted assistance and available provision clear ("AI Use").

Suspicion remains a possible feature of the same governance arrangement. Vanderbilt's integrity guidance includes the instruction "Review tone" and treats "Inconsistencies between the voice and writing style of prior submissions" as relevant to reporting concerns. These are judgments about appearance and continuity, not direct observations of prohibited assistance. Their inclusion creates a route from a marker's perception of writing to an institutional inquiry. The page also requires alignment with assignment instructions and describes further evidence and testimony, so its language cannot be reduced to automatic conviction on stylistic grounds ("Academic Integrity and Generative AI").

The risk is therefore conditional. A change in writing may reasonably invite an educational question, but it can reflect feedback, revision, language development, or permitted support. For a multilingual writer, being compared with an earlier performance may make improvement itself difficult to explain. That is an interpretation of the possible burden created by the guidance, informed by the detector literature; the corpus contains no student-level evidence that such discrimination occurred. Integrity becomes suspicion when an uncertain signal is treated as a reason the student must disprove, rather than an inference the institution must substantiate.

Detection before Dialogue Reconsidered

UCL's student guidance states, "UCL does not use GenAI detectors when marking." It then distinguishes a marker's request for information from an automatic accusation. ETH Zurich's integrity guidance states, "Trust in such a method is therefore not appropriate." Its referent is technical recognition of generated output. These passages directly contradict a claim that all institutional integrity advice gives detector scores priority over professional interpretation ("Engaging"; "Academic Integrity").

Waterloo's September 2025 decision notice is similarly explicit: "Time and effort are best spent on education rather than policing misuse of GenAI." The notice identifies reliability, bias,

and costs among its reasons for withdrawing the functionality and points instructors toward redesign and support. The wording does educational work by directing institutional effort away from the tool. It is evidence of a stated alternative to detector-led governance, while remaining a decision notice rather than an evaluation of the effects of that alternative ("Discontinuing").

These restrictions sit within a wider evidentiary problem. Turnitin's own report guide says the output "should not be used as the sole basis" for adverse action. It also distinguishes the AI percentage from a similarity score. The percentage concerns text classified by the model; it is not a finding that a student breached a particular assessment rule. This distinction matters even where a detector performs well under a given test. The institution must connect the evidence to the assistance permitted for the task ("Using").

Detection before dialogue is consequently a proposition the selected documents often resist. The remaining concern is how an institution responds to uncertainty once the score has been withdrawn or restricted. A commitment to dialogue can protect the student, or become an informal investigation without clear limits. The policy question is whether the conversation has an identified purpose, an appropriate evidentiary basis, and a route into the ordinary procedures if a misconduct concern persists. The absence of a detector does not answer those questions.

Teacher Authority and the Limits of Delegation

ETH Zurich's teaching guidelines state that "Performance assessments must be conducted independently and personally." They also make lecturers responsible for their materials and for deciding the permitted role of AI. UCL's categories leave the final assessment decision with the setter. These provisions preserve professional responsibility rather than documenting its wholesale transfer to a platform. The work of defining what independence means for a particular task remains with educators ("AI in Teaching"; "Three Categories").

Vanderbilt's withdrawal notice instructs staff, "If you have concerns about students misusing AI, talk to them about it." The advice restores a role for the educational relationship, but the accompanying suggestion to compare present writing with earlier work still requires a defensible interpretation. A teacher's familiarity can contribute relevant context. It can also carry assumptions about the style a student is expected to sustain. Replacing a machine score with a human impression therefore changes the form of judgment without settling its fairness ("Guidance").

Utrecht's resource by Koenders and Hsiao offers a further limit: "Having full certainty is therefore not the goal of this meeting." The statement concerns oral inquiry into authorship and understanding. Its significance is that the conversation is not presented as an instrument capable of establishing an entirely pure boundary between human and machine contributions. When concern remains, the resource directs staff toward established examination procedures. Professional judgment is retained, but its purpose and authority require explanation (Koenders and Hsiao).

These provisions distribute substantial discretion to teachers. The risk is that autonomy is paired with responsibility for interpreting uncertain indicators and explaining a decision under pressure. Academic judgment needs support through clear expectations and procedural review; it should not be romanticized as automatically impartial. The proposed concept concerns the conditions of that judgment, including cases where institutional demands make defensibility more salient than educational understanding.

Performative Authenticity and Process Evidence

UCL's library guidance instructs students to "Retain drafts of your work so you can evidence your process." This is a concrete movement from a finished submission toward an account of its production. A draft can illuminate revision, but the quoted rationale also anticipates a need to substantiate that process. The same artifact can support feedback or a defense of authorship. The wording alone does not show which function dominates in a particular course ("Acknowledging").

Utrecht makes another distinction explicit. Its account of a sampled oral check states that this is "not a grade determining part of the examination." This separates a verification procedure from the original marking judgment, while the resource also requires clarity about its purpose and handling. The separation can protect students from an undeclared change in assessment. It also makes visible an additional institutional task that a student may have to complete after submitting written work (Koenders and Hsiao).

Vanderbilt's recommendations to "Incorporate personal reflections" and "Scaffold assignments" connect process evidence to the prevention of misuse. These forms can create opportunities for learning. Their use as signs of authentic production introduces another possibility: students may learn to narrate a credible process without that narrative establishing the quality of their understanding. Reflection is not self-validating. Its assessment needs to consider the reasons offered and their relationship to the work, rather than its fluency or conformity to an expected personal voice ("Academic Integrity and Generative AI").

Performative authenticity describes that possible separation between an explanation's institutional credibility and its educational substance. It does not imply that declarations or staged submissions are inherently oppressive. A selective explanation tied to the learning outcome can be valuable. A demand for an extensive history that cannot be questioned may impose a different burden, especially where students have unequal confidence in describing their work in the expected language. The corpus establishes the availability of these documentary expectations, not the distribution of their effects.

Compliance as Learning and Institutional Responsibility

ETH Zurich calls for AI use to be "made transparent at all times," while also allowing course rules to be "jointly developed and subsequently documented." The first formulation extends visibility; the second permits participation in determining the rule. Their coexistence is important. A requirement that appears extensive when isolated may be accompanied by a practice through which students can understand or influence it. The task for analysis is to examine that relationship, rather than classify every demand for transparency as obedience ("Academic Integrity").

UCL's categories quote an institutional protection under which evidence of use without clear direction "will not be penalised." Here the university accepts a consequence for failing to explain its own expectations. That provision places a limit on the idea that students alone must manage uncertainty. Oxford likewise requires assessment setters to define permission and declarations. The relevant form of accountability is reciprocal: students must follow a clear rule, while staff must make the rule educationally and procedurally defensible ("Three Categories"; "AI Use").

The five themes therefore do not show that compliance has replaced learning throughout the selected institutions. They identify a contest over what makes work credible and who must explain a decision. The risk of algorithmic compliance pedagogy becomes sharper where

writing signals or process records are detached from the assessment purpose and treated as proof of a student's legitimacy. The counterexamples show that institutions can also limit this risk through explicit permission, restrictions on detection, dialogue, and duties placed on the assessor.

This interpretation identifies a risk of algorithmic compliance while setting a boundary around the claim. The policies invite particular forms of conduct; they do not demonstrate that students have become obedient or that educational relationships have ceased to matter. Whether students learn primarily to defend themselves, and whether some groups face more frequent demands to do so, remains an empirical question. The documents specify where that question should be investigated and what competing institutional commitments need to be considered.

Discussion

What the Concept Explains

The analysis locates the problem in the organization of judgment. Detector withdrawal can coexist with requirements to keep records, narrate a process, and answer questions about changes in style. This combination shows why the debate should extend beyond the accuracy of a particular tool. An institution may restrict automated classification while preserving expectations shaped by the effort to distinguish generated from supposedly authentic writing. The educational issue is how those expectations are connected to an inference about competence or conduct.

Algorithmic compliance pedagogy is useful when it identifies that connection precisely. It would become too broad if every disclosure counted as surveillance, or every request for an explanation counted as mistrust. The corpus contains examples of guidance that seeks to protect students and clarify learning. The concept should therefore distinguish a justified account of work from an unlimited demand to remain above suspicion. The interpretation depends on the task, the evidence sought, and the consequences of failing to provide it.

The strongest contribution of the comparison is the possibility of suspicion migrating across forms of verification. A score can be rejected while a stereotype about generated writing remains influential in human assessment. This does not make the stereotype an algorithm in the technical sense. It shows how expectations associated with machine classification may travel into professional practice. The analytical connection must be established through particular instructions, such as those concerning tone and continuity of style, rather than inferred from the mere presence of an integrity policy.

The Legitimate Case for Verification

Universities have obligations that trust alone cannot discharge. A qualification communicates a claim about attainment to other students, employers, professional bodies, and the public. Where a student outsources the competence being assessed, awarding credit can misrepresent what has been demonstrated and disadvantage peers who completed the task within its rules. There is therefore a strong educational and institutional case for checking whether an assessment supports the judgment made from it. Vanderbilt's guidance explicitly connects integrity with the credibility of a degree; Utrecht distinguishes the consequences of formative and summative assessment ("Academic Integrity and Generative AI"; Koenders and Hsiao).

That case does not justify every method of checking. A detector classification, an inconsistency in style, and an inability to answer an unexpected question have different evidentiary limits. The institution needs to identify the intended learning, the assistance

permitted, and the reason for concern before deciding what further evidence is proportionate. It also needs to explain how a student can respond. Verification can sustain trust when it makes a defensible judgment possible; it can damage trust when uncertainty is converted into a burden that the student cannot reasonably discharge.

The 2026 detector study makes this argument more demanding. If some systems perform better in specified conditions, criticism cannot rest only on the assertion that detection never works. The relevant institutional question remains whether a tool is suitable for the particular use and whether its output supports the decision under consideration. Evidence of AI-associated text is not, by itself, evidence of unauthorized assistance. Even a more accurate instrument does not settle the educational rule or replace an accountable assessment process (Van Vlasselaer et al. 15-17).

Proportionate Documentation and Academic Judgment

Four considerations follow from the comparison. The first is the connection to learning. A revision account should show something relevant to the outcome, such as how a student evaluated evidence or changed an argument. If the record adds little information about understanding, its purpose needs examination. The second is proportionality: a limited selection of drafts or decisions may be enough, whereas a complete interaction history can expose unnecessary material and create substantial work for both student and assessor. These are proposals arising from the analysis, not evaluated interventions.

The third consideration is institutional responsibility. Students should not have to infer the acceptable role of AI from conflicting examples or the preferences of an individual marker. The selected guidance contains commitments to clarify permission and support access. Those commitments should be examinable alongside compliance requirements. An institution's use of AI in teaching or feedback also needs explanation. Reciprocal transparency gives students a basis for understanding why assistance is legitimate in one part of an educational relationship and restricted in another.

The fourth is the opportunity to contest an inference. A student should be able to understand why further proof is being requested and offer an account that is considered on its merits. A concern about writing needs to remain open to other explanations, including development, feedback, and permitted support. This does not oblige an assessor to accept every explanation. It requires an institution to state the grounds for its judgment and to connect any consequence to the relevant rule and procedure. Documentation becomes educationally defensible when it contributes to that process.

Language and the Limits of the Evidence

The concern about multilingual writers must be kept distinct from a claim about all international students. Language background, nationality, and familiarity with an institution are different dimensions. The earlier detector research establishes specific classification problems. The policy passages on tone and style show possible occasions for a burden of explanation. Neither source demonstrates the rate at which particular student groups are investigated across these six universities. Treating the risk as a research question is stronger than repeating an unequal-impact claim without the evidence needed to assess it.

Future research should examine the routes through which concerns are raised, the explanations students provide, and how decisions are reviewed. It should distinguish selection for an interview from a misconduct finding, and both from a learning outcome. Work with

multilingual writers could explore how students explain changes in their writing and whether the requested evidence is accessible. Such research would test whether the formal safeguards identified here operate in practice, as well as whether informal judgments produce burdens the published guidance does not acknowledge.

The same restraint applies to teacher authority. The corpus shows instructions and responsibilities, not staff behavior or workload measured over time. Professional discretion can be supported by the rejection of an unreliable score, yet made difficult by ambiguous expectations about authorship. A critical approach needs to investigate both possibilities. Educational leadership should provide the conditions for defensible judgment instead of assuming that a policy, a platform, or an individual teacher can remove uncertainty altogether.

Conclusion

This study has examined how ten public documents from six universities make AI assistance permissible, visible, and open to inquiry. The documents do not support a general claim that universities have replaced pedagogy with automated policing. They contain restrictions on detectors, recognition of professional judgment, and commitments to clearer assessment design. At the same time, their instructions about style, draft retention, and oral explanation show how authorship can remain an additional object of institutional verification after automated detection is rejected.

Algorithmic compliance pedagogy identifies a risk within that arrangement. It concerns situations in which conformity to an authorship signal or a documentary expectation becomes more significant than the learning being assessed. The present evidence supports examination of this risk, not a conclusion that the selected universities produce obedient learners. Its effects on students and staff require investigation through experiences and decisions that policy documents alone cannot provide.

The practical implication is that verification should be connected to a stated learning purpose, proportionate evidence, and an accountable institutional judgment. Students have responsibilities for their work, while universities have responsibilities for the clarity of their rules and the grounds on which they question a submission. Protecting academic integrity requires both. A university can ask a student to explain a piece of work while remaining willing to explain, and review, its own reasons for doubt.

Works Cited

"Academic Integrity." *AI in Teaching and Learning*, ETH Zurich, ethz.ch/en/the-eth-zurich/education/ai-in-education/academic-integrity.html. Accessed 23 Oct. 2026.

"Academic Integrity and Generative AI." *Generative AI at Vanderbilt*, Vanderbilt University, www.vanderbilt.edu/generative-ai/academic-integrity/. Accessed 23 Oct. 2026.

"Acknowledging the Use of Generative AI and Referencing Generative AI." *Library Guides and Databases*, University College London, library-guides.ucl.ac.uk/referencing-plagiarism/acknowledging-genAI. Accessed 23 Oct. 2026.

"AI in Teaching and Learning." *ETH Zurich*, updated Dec. 2024, ethz.ch/en/the-eth-zurich/education/ai-in-education.html. Accessed 23 Oct. 2026.

"AI Use in Summative Assessment." *University of Oxford*, updated 27 Oct. 2025, www.ox.ac.uk/about/how-we-are-run/policies-and-statements/policy-hub/ai-use-in-summative-assessment. Accessed 23 Oct. 2026.

Baker, Ryan S., and Aaron Hawn. "Algorithmic Bias in Education." *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, 2022, pp. 1052-92, https://doi.org/10.1007/s40593-021-00285-9.

Bearman, Margaret, et al. "Discourses of Artificial Intelligence in Higher Education: A Critical Literature Review." *Higher Education*, vol. 86, no. 2, 2023, pp. 369-85, https://doi.org/10.1007/s10734-022-00937-2.

Biesta, Gert. "What Is Education For? On Good Education, Teacher Judgement, and Educational Professionalism." *European Journal of Education*, vol. 50, no. 1, 2015, pp. 75-87, https://doi.org/10.1111/ejed.12109.

Braun, Virginia, and Victoria Clarke. "Reflecting on Reflexive Thematic Analysis." *Qualitative Research in Sport, Exercise and Health*, vol. 11, no. 4, 2019, pp. 589-97, https://doi.org/10.1080/2159676X.2019.1628806.

Czerniewicz, Laura, and Catherine Cronin, editors. *Higher Education for Good: Teaching and Learning Futures*. Open Book Publishers, 2023, https://doi.org/10.11647/OBP.0363.

"Discontinuing Use of AI Detection Functionality in Turnitin: September 2025." *Associate Vice-President, Academic*, University of Waterloo, uwaterloo.ca/associate-vice-president-academic/discontinuing-use-ai-detection-functionality-turnitin. Accessed 23 Oct. 2026.

Dotan, Ravit, et al. "Responsible Adoption of Generative AI in Higher Education: Developing a 'Points to Consider' Approach Based on Faculty Perspectives." *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2024, pp. 2033-46, https://doi.org/10.1145/3630106.3659023.

"Engaging with Generative AI in Your Education and Assessment." *Study at UCL*, University College London, www.ucl.ac.uk/study/current-students/exams-and-assessments/assessment-success-guide/engaging-generative-ai-your-education-and-assessment. Accessed 23 Oct. 2026.

Foucault, Michel. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan, 2nd Vintage Books ed., Vintage Books, 1995.

Freeman, Josh. *Student Generative AI Survey 2025*. HEPI Policy Note 61, Higher Education Policy Institute, Feb. 2025, www.hepi.ac.uk/wp-content/uploads/2025/02/HEPI-Kortext-Student-Generative-AI-Survey-2025.pdf. Accessed 23 Oct. 2026.

"Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector." *Brightspace*, Vanderbilt University, 16 Aug. 2023, www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/. Accessed 23 Oct. 2026.

Jin, Yueqiao, et al. "Generative AI in Higher Education: A Global Perspective of Institutional Adoption Policies and Guidelines." *Computers and Education: Artificial Intelligence*, vol. 8, 2025, article 100348, https://doi.org/10.1016/j.caeai.2024.100348.

Jørgensen, Thomas E., and Clare Phelan. *Adopting AI That Serves the Needs and Values of Universities: Final Report of the EUA Task-and-Finish Group on Artificial Intelligence*. European University Association, 29 Jan. 2026, www.eua.eu/publications/reports/adopting-ai-that-serves-the-needs-and-values-of-universities.html. Accessed 23 Oct. 2026.

Koenders, Laura, and Ya-Ping (Amy) Hsiao. "A Follow-up Oral Process Interview: What It Is and What It Should Not Be." *Educational Development & Training*, Utrecht University, www.uu.nl/en/education/educational-development-training/knowledge-

dossiers/knowledge-dossier-generative-ai-in-education/a-follow-up-oral-process-interview-what-it-is-and-what-it-should-not-be. Accessed 23 Oct. 2026.

Li, Ming, et al. "A Framework for Developing University Policies on Generative AI Governance: A Cross-National Comparative Study." *Studies in Higher Education*, 13 July 2026, pp. 1-29, https://doi.org/10.1080/03075079.2026.2696496.

Liang, Weixin, et al. "GPT Detectors Are Biased against Non-Native English Writers." *Patterns*, vol. 4, no. 7, 2023, article 100779, https://doi.org/10.1016/j.patter.2023.100779.

Miao, Fengchun, and Wayne Holmes. *Guidance for Generative AI in Education and Research*. UNESCO, 2023, https://doi.org/10.54675/EWZM9535.

Perkins, Mike, et al. "GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education." *arXiv*, 28 Mar. 2024, arxiv.org/abs/2403.19148v1. Preprint.

Slade, Sharon, and Paul Prinsloo. "Learning Analytics: Ethical Issues and Dilemmas." *American Behavioral Scientist*, vol. 57, no. 10, 2013, pp. 1510-29, https://doi.org/10.1177/0002764213479366.

Stephenson, Rose, and Charlotte Armstrong. *Student Generative AI Survey 2026*. HEPI Report 199, Higher Education Policy Institute, 12 Mar. 2026, www.hepi.ac.uk/reports/student-generative-ai-survey-2026/. Accessed 23 Oct. 2026.

"Three Categories of GenAI Use in Assessment." *Teaching & Learning*, University College London, www.ucl.ac.uk/teaching-learning/generative-ai-hub/three-categories-genai-use-assessment. Accessed 23 Oct. 2026.

"Using the AI Writing Report." *Turnitin Guides*, Turnitin, guides.turnitin.com/hc/en-us/articles/22774058814093-Using-the-AI-Writing-Report. Accessed 23 Oct. 2026.

Van Vlasselaer, Marijke, et al. "Who Wrote This? Evaluating the Reliability of AI Detection Tools in Higher Education." *International Journal for Educational Integrity*, vol. 22, no. 1, 2026, article 16, https://doi.org/10.1007/s40979-026-00226-w. Corrected 6 Oct. 2026.

Weber-Wulff, Debora, et al. "Testing of Detection Tools for AI-Generated Text." *International Journal for Educational Integrity*, vol. 19, no. 1, 2023, article 26, https://doi.org/10.1007/s40979-023-00146-z.

Williamson, Ben. *Big Data in Education: The Digital Future of Learning, Policy and Practice*. SAGE, 2017, https://doi.org/10.4135/9781529714920.

Appendix A: Corpus Register

A containing guide's timestamp is not treated as a date for every individual section. Full source details, including URLs, appear in Works Cited. Linked documents and external references mentioned by a page are not automatically additional corpus items.

ID	Institution and country	Document and status	Visible date or update information
P1	University College London, United Kingdom	Three Categories of GenAI Use in Assessment Assessment design guidance; explicitly not a formal policy Passages examined: guidance; categories of assessment; notes on academic integrity	Undated live page; announces review during 2026/27
P2	University College London, United Kingdom	Engaging with Generative AI in Your Education and Assessment Student assessment and integrity guidance Passages examined: GenAI and academic integrity	Undated live page; announces review during 2026/27
P3	University College London, United Kingdom	Acknowledging the Use of Generative AI and Referencing Generative AI Library guidance on acknowledgment and documentation Passages examined: considerations for using generative AI; how to acknowledge use	Section undated; containing guide timestamp 29 July 2026
P4	University of Oxford, United Kingdom	AI Use in Summative Assessment University assessment policy Passages examined: staff requirements 1-6; student requirements 7-9; three principles	Page last updated 27 October 2025; displayed review date 2 July 2026
P5	Utrecht University, Netherlands	A Follow-up Oral Process Interview: What It Is and What It Should Not Be Teaching and examination guidance by Laura Koenders and Ya-Ping (Amy) Hsiao Passages examined: assessment purposes; random interview before grading; targeted cases; communication	Undated live page; publication date not inferred from acknowledgment
P6	ETH Zurich, Switzerland	AI in Teaching and Learning Institutional teaching and learning guidelines Passages examined: generative artificial intelligence; responsibility; transparency; fairness	Explicit update notice December 2024
P7	ETH Zurich, Switzerland	Academic Integrity Guidance on integrity and originality declarations Passages examined: integrity as an attitude; scientific work; declaration options; technical recognition	Undated live page
P8	Vanderbilt University, United States	Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector Dated decision notice and teaching guidance Passages examined: decision rationale; advice following withdrawal of detector	16 August 2023
P9	Vanderbilt University, United States	Academic Integrity and Generative AI Institutional teaching and integrity FAQ Passages examined: instructor policy; proactive strategies; detection advice; Honor Council guidance	Undated live page
P10	University of Waterloo, Canada	Discontinuing Use of AI Detection Functionality in Turnitin: September 2025 Decision notice and guidance for instructors Passages examined: rationale for this decision; content redesign and academic integrity	September 2025 (effective date in the title; no publication date shown)

Appendix B: Document Contributions to the Five Themes

The matrix records documents used in developing each theme, including protections and counterexamples. A mark indicates a contribution to the interpretation, not endorsement of surveillance or evidence of an observed educational effect. An empty cell means the document was not used for that thematic comparison; it is not a claim that the institution lacks the relevant practice. Each document is counted once within a theme.

Document	T1	T2	T3	T4	T5
P1	1	1	1		1
P2	1	1		1	
P3				1	
P4	1	1	1	1	1
P5			1	1	1
P6	1		1		1
P7	1	1	1	1	1
P8		1	1		1
P9	1	1	1	1	1
P10		1			1
Total	6	7	7	6	8

T1 = Integrity as suspicion and conditional permission; T2 = Detection before dialogue reconsidered; T3 = Teacher authority and the limits of delegation; T4 = Performative authenticity and process evidence; T5 = Compliance as learning and institutional responsibility.