

The Triple-A Framework: Towards AI Assurance and Accountability in Higher Education

Timothy D. Hubbard (✉), Theo Lynn

DCU Business School, Dublin City University, Dublin 9, Ireland

✉ Corresponding author. E-mail: tim.hubbard@dcu.ie (Timothy D. Hubbard)

Front. Digit. Educ., **Just Accepted Manuscript** • DOI: 10.1007/s44366-027-0120-8
<http://journal.hep.com.cn> on September 30, 2026

© Higher Education Press 2026

Just Accepted

This is a “Just Accepted” manuscript, which has been examined by the peer-review process and has been accepted for publication. A “Just Accepted” manuscript is published online shortly after its acceptance, which is prior to technical editing and formatting and author proofing. Higher Education Press (HEP) provides “Just Accepted” as an optional and free service which allows authors to make their results available to the research community as soon as possible after acceptance. After a manuscript has been technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an Online First article. Please note that technical editing may introduce minor changes to the manuscript text and/or graphics which may affect the content, and all legal disclaimers that apply to the journal pertain. In no event shall HEP be held responsible for errors or consequences arising from the use of any information contained in these “Just Accepted” manuscripts. To cite this manuscript please use its Digital Object Identifier (DOI(r)), which is identical for all formats of publication.

FDE-27011-TH

Front. Digit. Educ., 2027, 4(1): 11

Verso: Front. Digit. Educ., 2027, 4(1): 11

Recto: Hubbard, T.D. & Lynn, T. The Triple-A Framework: Towards AI Assurance and Accountability in Higher Education

RESEARCH ARTICLE

<https://doi.org/10.1007/s44366-027-0120-8>

The Triple-A Framework: Towards AI Assurance and Accountability in Higher Education

Received July 23, 2026; revised August 27, 2026; accepted September 16, 2026

Timothy D. Hubbard (✉), Theo Lynn

DCU Business School, Dublin City University, Dublin 9, Ireland

Corresponding author. Timothy D. Hubbard, tim.hubbard@dcu.ie

© Higher Education Press 2026

Abstract Generative artificial intelligence (GenAI) is reshaping higher education at a pace that outstrips universities' policy, pedagogical, and procurement responses. While guidance documents proliferate, consistent governance does not. This paper adapts assurance and accountability scholarship to GenAI in higher education by proposing the Triple-A Framework (AI Assurance and Accountability in Higher Education), a multi-stakeholder, five-control-function framework. The resulting 5×5 matrix crosses five higher-education stakeholders—students, educators, institutions, technology providers, and policymakers, regulators, and accreditation bodies—with the five control functions: declarative, confirmative, preventative, detective, and corrective. We then apply a trust lens, showing how the assurance functions support trust building, how the accountability functions support trust repair, and how the preventative function deliberately bridges the two. The framework is a diagnostic and design tool, not a compliance checklist, standard, or maturity model: it shows where trust work must

occur. We illustrate the framework with vignettes and set out a research and policy agenda for borderless, equity-aware faculty and institutional development communities aligned with UN Sustainable Development Goal 4.

Keywords generative AI, higher education, faculty development, trust, AI governance, assurance, accountability

1 Introduction

Higher education has become a daily user of generative artificial intelligence (GenAI). Students rely on it for drafting, summarising, coding, and tutoring. Faculty rely on it for lesson planning, feedback, and administrative work. Institutions are quietly integrating it into admissions, student services, and research support. Researchers have already documented the breadth of higher education's response: paradigm-shifting in framing (Wu & Chen, 2025), curriculum-reforming in practice (Ma et al., 2025), and uneven in execution (Krause et al., 2025). How universities proceed next matters.

However, the response that higher education has largely settled on (i.e., calling for future faculty development) is not enough on its own. A focus on faculty makes sense, as they are central to teaching and assessment and, further, are the most visible interface students have with the institution. But faculty cannot, by themselves, supply the institutional grammar of trust that GenAI now demands, i.e., the shared set of rules, roles, evidencing practices, and remediation routes through which an institution (or, where institutional capacity is thin, the sector acting on its behalf) makes its handling of GenAI legible to those affected by it. Faculty cannot be expected to prevent harm the institution has not designed against; they cannot detect provider failures without provider

cooperation; and they cannot make their assurances to students credible without institutional reporting to back them up.

This matters for three reasons. First, faculty efforts will not stick if they rest on unstated rules and unspecified accountability. Second, the failures that have already occurred, including undeclared use in assessment, unauthorised tool adoption, and leaked student data, are interpreted by students and the public as evidence about the institution (Adu & Owusu-Agyeman, 2026; Pesek, 2026). Third, the asymmetry between well-resourced and less well-resourced universities widens month by month because the assurance work that wealthy institutions can absorb internally has no obvious equivalent for everyone else.

Left unaddressed, this asymmetry risks turning a generational capability shift into a generational inequity. Assurance capacity is itself an equity issue. Well-resourced institutions can build governance, procurement, legal, technical, and audit capacity internally; less well-resourced institutions cannot. Unequal assurance capacity could produce unequal GenAI governance capacity, which in turn leads to unequal protection for students and uneven credibility across the sector. Many of the controls that wealthy institutions can host internally must be carried at the sector level for everyone else through shared declarative templates, model procurement clauses, provider trustmarks, accreditation-aligned reporting, and pooled audit mechanisms. These concerns are not hypothetical. A systematic review of the “dark side” of AI in education identifies surveillance, deskilling, narrowing of pedagogy, and the reproduction of inequalities as documented harms already visible in higher education today (Adu & Owusu-Agyeman, 2026). In addition, there has been widespread concern regarding cognitive offloading, the risk that students and staff delegate to GenAI the very cognitive work that education exists to develop (Gerlich, 2025).

To address these concerns, we treat trustworthy GenAI in higher education as a co-designed property of the whole ecosystem. Building on a published framework for trust, assurance, and accountability originally developed for cloud computing (Lynn et al., 2021), we present the Triple-A Framework (AI Assurance and Accountability in Higher Education), a 5×5 matrix that crosses five higher-education stakeholders (students, educators, institutions, technology providers, and policymakers, regulators, and accreditation bodies) with five control functions: declarative, confirmative, preventative, detective, and corrective. Assurance functions support trust building; accountability functions support trust repair. The matrix gives the institutional grammar of trust a concrete object to organise around.

We make three contributions. First, we identify four ways GenAI undermines the assumptions underlying cloud-era assurance and accountability frameworks—novel content generation, non-determinism, anthropomorphisation, and mission-constitutive deployment—and develop a new framework whose topology is informed by the cloud architecture but whose content is tailored for GenAI in higher education. Second, we use the resulting matrix to illustrate one application, reframing future faculty development not as another layer of AI literacy but as fluency across the five control functions; a fuller treatment of programme design is beyond this paper’s scope. Third, we set out a research and policy agenda, including how the control functions may need to be re-weighted as AI capabilities change.

2 Background: Trust, Assurance, and Accountability

Trust is conventionally defined as “a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another” (Rousseau et al., 1998). This definition emphasises two main points relevant to this paper. First, trust involves a willingness to accept vulnerability that is rarely the product of fully deliberate calculation.

Second, people rely on specific cues about the other party and about the surrounding environment when deciding whether to take that risk.

A prominent stream of trust research distinguishes between three knowledge-based antecedents, collectively termed trustworthiness: ability (does the other party have the competence to do what is needed?), integrity (does the other party act on principles I can accept?), and benevolence (does the other party act in my interests?) (Mayer et al., 1995). A complementary tradition, often termed institution-based trust, focuses on contextual antecedents that shape trust when the trustor lacks the expertise or information for fully rational evaluation: structural assurance (the belief that rules, regulations, and safeguards make the setting safe) and situational normality (the belief that the environment is in proper order) (McKnight et al., 1998). A subsequent line of work extends these constructs from trust in people and vendors to trust in the technology artefact itself, distinguishing beliefs about a specific technology's helpfulness, reliability, and functionality from beliefs about its provider (McKnight et al., 2011; Söllner et al., 2016). Both traditions apply here, and the artefact-level extension is especially salient for GenAI: students and faculty are rarely technical experts in the tools they use, and they assess them through a mix of competence judgments about the system, contextual cues about the institutional environment, and inferences about the provider standing behind both.

Trust theory has informed two complementary governance vocabularies. Assurance is typically forward-looking. It concerns mechanisms intended to establish justified confidence before and during reliance. Accountability operates across the lifecycle through the allocation of responsibility, traceability, investigation, remediation, and, where appropriate, sanction. Accountability's role in trust repair is especially evident after failure. As such, it is often backward-

looking. Both are necessary. Assurance without credible accountability is brittle, while accountability without anticipatory assurance is predominantly reactive.

Earlier work proposed an assurance and accountability framework for cloud computing built around five control functions—declarative, confirmative, preventative, detective, and corrective—that span the range from announcing a position to remediating a breach (Lynn et al., 2013; Lynn et al., 2021). That framework was developed for an information-processing utility whose harms were largely about data confidentiality and service availability. GenAI changes the framework's foundations in four ways. First, GenAI produces novel content rather than solely processing data the user has supplied. Second, its outputs are not deterministic. The same input can yield materially different responses across sessions and configurations, complicating both assurance and after-the-fact reconstruction. Third, it invites anthropomorphisation in ways cloud services never did, blurring the boundary between tool use and relational reliance. Anthropomorphic cues can encourage users to treat GenAI as quasi-human counterparts; research links anthropomorphic characteristics to emotional dimensions of trust (Glikson & Woolley, 2020). Yet GenAI remains a technological artefact whose apparent candour, humility, and ethics are parameterised by a provider, so user trust is layered across the artefact, the deploying institution, and the provider. This layered structure is one reason the trust-in-technology constructs are relevant (McKnight et al., 2011). Fourth, it is deployed in a context, the formation of human capability, that is the educational mission itself, so its failures are not adjacent to the institution's purpose but a core feature of it.

Recent scholarship has begun to articulate what education-specific AI governance should look like. Pesek (2026) argues for rethinking education governance from first principles as AI capabilities expand, treating governance not as a layer added on top of existing arrangements but

as a property of the system as a whole. Mobilio and Guglielmini (2026) organise the transformation around four pillars (equity, ethics, pedagogy, and capacity) and warn that the system's governance arrangements lag its technical capabilities. We share these diagnoses and aim to consider them throughout. Our contribution is to build a new framework for GenAI in higher education whose five-control-function lifecycle is informed by the cloud architecture, but whose substantive content reflects the four shifts identified above, and to give faculty development a concrete object to organise around rather than a moving target.

3 The Higher Education GenAI Landscape

GenAI is now embedded in higher education at scale. Recent surveys of individual U.S. institutions indicate that well over 80% of students report using generative tools for academic work at some campuses (e.g., Contractor & Reyes, 2025), though sector-wide prevalence remains unevenly measured. Faculty adoption also appears to be rising more slowly but accelerating (Digital Education Council, 2025; Tyton Partners, 2024). Lau et al. (2026) identify fragmented buy-in, conflicting policy signals, implementation challenges, limited visibility into student-AI interactions, and resource constraints as challenges to adoption. Institutional responses, by contrast, remain uneven. Most universities have published guidance, but fewer have integrated GenAI into other functions. Even at top-ranked universities, formal reviews of AI use in assessment vary widely (Moorhouse et al., 2023). The regulatory environment is also evolving. Under the European Union (EU) AI Act, specified education-related uses, including AI used for admission or allocation, evaluation of learning outcomes, decisions concerning an appropriate education level, and monitoring prohibited behaviour during tests, fall within Annex III high-risk categories subject to Article 6. Following Regulation (EU) 2026/1744 (The European Parliament and the Council of the European Union, 2026), the principal requirements in Sections 1–3 of Chapter III applicable to AI

systems classified as high-risk under Article 6(2) and Annex III apply from 2 December 2027. Obligations are role-specific. For example, providers must meet requirements including risk management and technical documentation, while deployers have distinct obligations concerning appropriate use, human oversight, monitoring, logging, and information. While less evolved, the United States operates a patchwork of agency guidance and state-level legislation (California Legislature, 2024; Biden, 2023; U.S. Department of Education, 2023), and China has issued sector-specific direction on AI in education (Ministry of Education of the People’s Republic of China, 2018, 2024). Documented incidents already span the categories the framework is designed to address; e.g., undeclared use in assessment, unauthorised procurement, training data leakage from student work, and reputational damage from grading tool failures. Researchers have tracked parts of this picture—curriculum reform (Ma et al., 2025), human–AI collaboration (Wu & Chen, 2025), learning-outcome effects (Krause et al., 2025), and AI-literacy instrumentation (Wu et al., 2025). Less developed is the institutional grammar that makes these efforts coherent among stakeholders. We turn to that next.

Across this landscape, five stakeholder groups co-produce the conditions under which university students or researchers encounter GenAI. First, *students* engage as learners, with interests in fair treatment, privacy, and clarity about what constitutes acceptable assistance. Second, *educators* engage as designers and assessors, with interests in pedagogy, assessment integrity, supervisory workload, and professional identity. Third, *institutions* engage as the responsible parties, with interests in policy, procurement, governance, risk, reputation, and mission. Fourth, *technology providers* engage as commercial actors, with interests in model performance, data-handling commitments, service-level agreements, and contractual durability. And fifth,

policymakers, regulators, and accreditation bodies engage as system-level guarantors, with interests in standards, sanctions, sector-wide equity, and the credibility of qualifications.

The five stakeholder groups are not symmetric. Students have the least leverage but the most exposure to outcomes. Educators have the most direct line of sight to harm but the least authority to redesign infrastructure. Institutions have authority but variable capacity. Technology providers hold many of the design choices that shape what is possible, while provider accountability remains fragmented and jurisdiction-dependent. Some jurisdictions now impose education-relevant AI obligations, whereas elsewhere accountability remains largely indirect through data protection, consumer, procurement, contractual, or general AI regulation. Policymakers and accreditation bodies have legitimacy but, in most jurisdictions, no detailed view of what happens inside the classroom. Any framework that hopes to coordinate these groups must, therefore, be explicit about who is being asked to do what at each point in the lifecycle—and about which gaps are gaps in capability, which are gaps in authority, and which are gaps in incentive.

4 The Triple-A Framework: A 5 × 5 Assurance and Accountability Matrix

Before presenting the framework in detail, we clarify the conceptual vocabulary it uses. We treat declarative, confirmative, preventative, detective, and corrective as control functions within an assurance and accountability lifecycle, not as controls in themselves. A control is a specific mechanism—an access restriction, a Data Protection Impact Assessment (DPIA) gate, an audit log, an assessment-review process, an incident-response procedure, a contractual audit right, a monitoring workflow. Control functions describe what such mechanisms do throughout the lifecycle. A useful hierarchy runs as follows: at the highest level sits the framework itself (assurance and accountability); within the framework sit five control functions. Each function has a control objective—what must be achieved under that function. Each objective is realised through

one or more control mechanisms, the specific processes, technical measures, artefacts, or obligations, and each mechanism produces evidence that it exists and works. The matrix in Table 1 is populated at the level of control objectives and example mechanisms, not an exhaustive control catalogue.

*** Insert Table 1 about here *** (此处插入表 1)

The distinction matters because the same mechanism can perform different functions depending on how it is used. A log of GenAI tool access is confirmative when used to evidence alignment with declared expectations; it is detective when used to identify anomalies, breaches, or drift. An approved-tool catalogue is declarative when it lists acceptable tools; it becomes preventative only when linked to access controls, procurement gates, DPIA requirements, contractual restrictions, or workflow barriers that reduce the likelihood of a specified, foreseeable harm.

We name the resulting framework the Triple-A Framework (AI Assurance and Accountability in Higher Education). It has two dimensions: stakeholder groups and control functions. Building on prior frameworks, we retain the five control functions, grouped by whether they support assurance or accountability. The five control functions are not unique to the cloud-assurance tradition from which they derive; the parent framework itself is multi-stakeholder and explicitly concerned with assurance as trust-building and accountability as trust repair (Lynn et al., 2021). These functions also closely align with governance architectures developed independently for other relevant purposes. For example, NIST’s AI Risk Management Framework uses Govern, Map, Measure, and Manage; ISO/IEC 42001 (International Organization for Standardization, 2023) follows a Plan-Do-Check-Act cycle; and the NIST Cybersecurity Framework 2.0 (Pascoe et al., 2024) uses Govern, Identify, Protect, Detect, Respond, and Recover. Broadly, these map onto our

framework's declarative, confirmative, preventative, detective, and corrective functions, while the Map and Identify phases others have used correspond to the named risks discussed later. The contribution is therefore not introducing multi-stakeholder governance or a trust lens de novo, but specifying them for GenAI in higher education. Relative to the parent framework and cited governance architectures, the proposed framework makes two substantive adaptations. First, it decomposes broad ecosystem actors into five higher-education stakeholder groups with distinct forms of authority, exposure, capacity, responsibility, and incentives. Second, it populates the framework with education-specific control objectives, mechanisms, evidence, and named-risk pathways spanning teaching, assessment, research, procurement, and student support. The Triple-A framework is therefore an education-specific extension and reinterpretation of the earlier assurance and accountability architecture rather than a new governance lifecycle.

We retained these five functions as an initial functional vocabulary because we claim they constitute an optimal decomposition of AI governance. First, they provide broad temporal coverage of governance activity before, during, and after potential failure. Second, they map productively onto the trust-building and trust-repair questions examined here. Finally, because they operate at the functional level, rather than the mechanism level, they are likely to remain useful as technologies, risks, and control mechanisms change.

The five control functions we put forward are declarative, confirmative, preventative, detective, and corrective. Declarative control functions make rules, expectations, and commitments visible. Confirmative control functions provide evidence of alignment between what has been declared and what is happening in practice. Preventative control functions design out foreseeable harms before they occur. Detective control functions identify breaches, failures, and

emerging harms. Corrective control functions support trust repair through proportionate, transparent responses, remediation and learning.

The first three are forward-looking assurance and support trust building; the last two are backward-looking accountability and support trust repair. The split is functional rather than absolute: accountability functions repair trust when invoked, but their visible presence (predictable oversight, predictable redress) also does trust-building work before any failure, which is why structural assurance appears on both sides of Table 3. The lifecycle should be seen as a loop rather than a line (Figure 1). Together, the five control functions describe a governance lifecycle that begins with announcements about how GenAI will be used, ends with the institutional- and sector-level learning that follows a breach, and feeds that learning back into the declarative layer.

*** Insert Figure 1 about here *** (此处插入图 1)

Corrective practice deserves separate attention because proportionality matters as much as transparency, and the two types of harm it addresses are not equally well-mapped. Where the harm is to data, correction is comparatively well-mapped: notification, remediation, and regulator engagement. Where the harm is reputational or epistemic. Examples include a cohort of degrees awarded under a flawed assessment regime, public erosion of confidence in a qualification, a research record contaminated by undisclosed generative content, correction is harder. The trust-repair literature suggests why. Responses effective for competence failures, such as apology and remediation, may be insufficient for integrity-adjacent violations, which require demonstrated reform (Kim et al., 2004). At organisational and institutional levels, durable trust repair requires diagnosis, intervention reform, and evaluation capable of demonstrating that the conditions underlying the failure have changed (Bachmann et al., 2015; Gillespie & Dietz, 2009). In this context, the institution must identify what went wrong, remedy what it can reasonably remedy,

communicate transparently with affected parties, and demonstrate substantive changes to the conditions that produced the failure. Because such harms cannot be fully remedied, corrective practice here leans most heavily on feedback from a single incident into the declarative layer, where the credibility of the whole lifecycle is ultimately tested. This staged view is consistent with organisational trust-repair research, in which repair proceeds through immediate response, diagnosis, reforming interventions, and evaluation (Gillespie & Dietz, 2009), and, at institutional scale, depends on re-establishing the legitimacy of the very structures that failed (Bachmann et al., 2015).

Crossing five stakeholders with five control functions produces Table 1. Each cell summarises the expected practice for that stakeholder at that point in the lifecycle. Read horizontally, the matrix describes what a single stakeholder must do across the full lifecycle. Read vertically, it describes the system-wide capability required at each control function. A control function is fully realised only when all five stakeholder rows are populated; an unpopulated row indicates a system-level gap, not just a stakeholder failing.

The named-risk test embedded in the preventative function implies a working risk typology. Table 2 illustrates how a small set of recurring GenAI risks in higher education can be traced across the lifecycle: preventative functions anticipate the named risk, detective functions identify the signals that it has materialised, and corrective functions repair the harm and revise the system. The typology is illustrative rather than exhaustive; its purpose is to make the named-risk test auditable. As such, each preventative mechanism in Table 1 should be traceable to a named risk of this kind.

*** Insert Table 2 about here *** (此处插入表 2)

The matrix can be re-grounded in trust theory. Each control function foregrounds a different trust problem (Table 3). Declarative functions principally express integrity and structural assurance: they expose the principles the stakeholder commits to acting on and establish the rules of engagement. Confirmative functions express integrity and ability together: integrity in the willingness to show what is happening in practice, ability in the competence to make that evidence intelligible. Preventative functions express ability, benevolence, and integrity: ability in the technical and procedural competence to anticipate harm, benevolence in the willingness to act on others' behalf before failure occurs, and integrity in showing that commitments are taken seriously in advance of convenient failure; by making safe use routine and expected, preventative functions also support situational normality. Detective functions initiate trust repair rather than perform it; i.e., detection surfaces the violation that the repair process then addresses (Gillespie & Dietz, 2009). They express integrity and structural assurance: integrity in the willingness to surface failure honestly, structural assurance in the predictability of oversight. Corrective functions express benevolence, integrity, and structural assurance together: benevolence in the intent to make affected parties whole, integrity in transparent response, and structural assurance in the predictability of redress.

*** Insert Table 3 about here *** (此处插入表 3)

To make the cells concrete, consider the Institutions row read across all five control functions. Declaratively, the institution publishes its position on permitted and prohibited use of GenAI across teaching, assessment, research supervision, and student support, along with the scope of approved tools. Confirmatively, it evidences alignment between that position and actual practice through internal review, audit, and public reporting on a cadence proportionate to the risk profile and rate of change of the tools in use. Preventatively, it designs out foreseeable harm at the

level of curriculum, assessment, procurement, and data flow—for example, by requiring Data Protection Impact Assessments before adoption and by approving assessment designs that anticipate GenAI use rather than assume its absence. Detectively, it instruments its environment to surface anomalies, e.g., unauthorised tool deployments, unusual data egress, patterns in assessment submissions, and provides channels through which students and faculty can raise concerns. Correctively, it responds to confirmed breaches proportionately and transparently, with named owners, defined timelines, and a feedback loop into the declarative layer so that the next iteration of policy reflects what was learned. These activities are illustrative rather than exhaustive; the remaining stakeholder rows can be populated analogously each of the other four stakeholder rows analogously. Detective instrumentation must itself be proportionate. The surveillance harms documented in the education literature (Adu & Owusu-Agyeman, 2026) can be produced by exactly these mechanisms, which is why the student-facing detective cell is anomaly-triggered rather than continuous.

Each row $\times$ column intersection represents a specific stakeholder operating in a specific control function. These row $\times$ column intersections are where the work of trustworthy GenAI in higher education must happen. Two features of the framework follow from this. First, it is a whole-ecosystem framework. No single stakeholder can deliver trustworthy GenAI on its own. Faculty cannot prevent harms that the institution has not designed to avoid; institutions cannot detect provider failures without provider cooperation; and regulators cannot impose standards that institutions are not equipped to meet. Second, the framework is intentionally generative rather than prescriptive. We do not claim that the cells of Table 1 are exhaustive or final.

The matrix can be used in three ways. First, diagnostically: an institution can ask which cells are populated and which are empty in its own practice. Empty cells are where trust is most

likely to break down, and they are easy to find. A department that has issued no declarative guidance on GenAI use in assessment has an empty Educator–Declarative cell, and the downstream functions in that row—Confirmative, Preventative—cannot be meaningfully populated without it. Second, developmentally: the matrix sequences investment in faculty, institutional, and provider capacity over time, because each control function depends on the one before it. Confirmative reporting is meaningless without something declared to confirm against. Detective findings are harder to interpret when expectations were never declared and preventions never specified, because there is no baseline against which breach or drift can be read. Third, comparatively: the matrix lets institutions, providers, and sector bodies be benchmarked against each other and tracked over time, supporting the kind of cross-institutional learning that Digital Intelligence International Development Education Alliance and similar networks are designed to enable.

It is worth clarifying what the matrix is not. It is not a compliance checklist and should not be treated as one. The cells are descriptions of practice, not specifications. Two institutions can populate the same cell in radically different ways—a small, undergraduate-focused university and a large, research-intensive one will both need to publish declarative guidance on GenAI use in assessment, but the form, level of detail, and review cadence appropriate to each will differ. The matrix tells institutions where the work has to happen; it does not tell them exactly how. That is by design. A prescriptive framework would freeze practice at a moment of unusually rapid technical change. A descriptive framework lets practice evolve while keeping the underlying trust architecture intact. Notwithstanding this, it is not unreasonable to expect that a given jurisdiction, or a group of institutions, regulators, and accreditation bodies, may adopt a common set of mechanisms for the same control objectives.

5 Worked Vignettes

Three vignettes show the matrix in action. Each describes how a single incident is routed among stakeholders and control functions. They are deliberately mundane. The point of each vignette is not that the situation is exotic, but that the response is. Without the matrix, each case is handled idiosyncratically and slowly; with it, the response is faster, more proportionate, and more legible to those affected.

5.1 Assessment Integrity

A final-year project is submitted with citations that cannot be located. The module's declared rules, such as its course-level GenAI statement, disclosure expectations, and academic integrity procedure, define the expectations against which the submission is judged (declarative). Because the assessment was designed in advance to require retained drafts and version history, process evidence exists before any concern arises (preventative), and the student's declared-use cover sheet records what was disclosed (confirmative). The unlocatable citations are the signal that triggers a case review. The institution's process compares the submission with the retained drafts, version history, and an oral defence (detective). A proportionate outcome is reached, and the case is fed back into module-level guidance and assessment redesign (corrective). The provider's role (what the model produced, under what configuration) becomes evidence. The regulator's role is only triggered if the institutional response is itself inadequate. This concern also arises in doctoral dissertation writing, where GenAI now plays a documented role in drafting, editing, and reference construction, and where prompt literacy itself becomes part of what students are expected to acquire (Rafi & Khalil, 2026).

5.2 Procurement and Data Protection

An academic department adopts a niche GenAI tutoring tool without a DPIA. The institution discovers the deployment through anomaly detection in enterprise single sign-on logs (detective). Preventative measures, e.g., an approved tool catalogue and a procurement gate, were declared but not enforced. Correction includes suspending the deployment, determining whether the processing meets the applicable DPIA threshold, completing any required impact assessment before resumption, assessing notification and remediation obligations, and revising the procurement gate. The provider's contractual commitments regarding data processing become directly relevant.

5.3 Accreditation and Equity—A Case Under Strain

A regional university with limited resources adopts a third-party GenAI grading assistant to ease the burden of grading large cohorts in a first-year quantitative course. Six months later, students raise concerns about inconsistent feedback. The institution moves to investigate (detective) but discovers that its contract grants no inspection rights over the model, no access to logs, and no obligation on the provider to disclose configuration changes; the Provider–Confirmative cell is empty because the institution lacked the procurement capacity to populate it at the declarative stage. Sector-level confirmative mechanisms (e.g., accreditation standards, conformity assessment, or a sector trustmark) would have transferred the inspection work to the sector, but in this jurisdiction those mechanisms do not yet exist; the Regulator–Confirmative cell is also empty. The institution must therefore correct the issue under conditions where it cannot fully diagnose the harm and cannot credibly assure students that it has been remedied. The matrix does not resolve the case; it identifies exactly which cells failed and at which level the remedy must be built. This is the structural pattern the equity literature identifies (Adu & Owusu-Agyeman, 2026; Mobilio & Guglielmini, 2026) and the level at which equity-focused faculty development communities add

the most value: not by enabling each institution to do the work alone, but by pooling the assurance and accountability infrastructure that smaller institutions cannot build in isolation.

6 Sample Metrics

Both AI and the higher education context are rapidly evolving. Thus, recommending specific metrics for each stakeholder group across the full framework scoped in Table 1 would quickly go out of date. However, it's important to provide guidance on example metrics and how they function within the framework. The following examples illustrate one metric for each control function and use each to show how the metrics link to the framework's foundational goals or specific features.

For the declarative control function, a straightforward metric could be the percentage of programmes and courses with a published, current GenAI use statement that references an approved scope for AI tools. Over time, as that metric becomes consistently high, new metrics will need to be introduced. For the confirmative control function, measure the percentage of GenAI tool deployments that undergo a conformance review (including the audit itself, the internal review documents, and a public report) in line with the tool's risk-based process. A metric like this clearly connects risk management with the confirmative process. For the preventative control function, metrics could aim to make named risks auditable. For example, an organisation could measure the percentage of preventative mechanisms (e.g., DPIA gates) that are traceable to specific named risks, such as those provided in Table 2. For the detective function, time-to-detection (the median time between a breach event and its surfacing) can be useful but should not be interpreted in isolation. Reducing detection time through intrusive monitoring could itself create privacy, surveillance, or false-positive harms. Institutions could therefore pair severity-adjusted time-to-detection with measures of false-positive burden, privacy impact, and the proportion of high-risk

uses subject to proportionate monitoring. Together, such metrics can help ensure the trust repair process starts promptly without incentivising disproportionate surveillance. Finally, for the corrective function, a more durable measure could be the percentage of material incidents receiving documented root-cause review and a recorded decision on whether changes to policy, guidance, controls, training, contractual terms, or no change are warranted. This would capture organisational learning without assuming every incident should produce a policy amendment.

7 Implications for Future Faculty Development

The framework reframes future faculty development as fluency across the five control functions, not as the accumulation of tool-specific competence. AI literacy frameworks, including the comprehensive evaluation system developed at Wuhan University (Wu et al., 2025) and the UNESCO AI Competency Framework for Teachers (Miao & Cukurova, 2024), are valuable inputs to the declarative and preventative functions, but they do not, on their own, supply the confirmative, detective, or corrective practice that makes assurance and accountability stick. This reframing aligns with the special issue's emphasis on the deep integration of Faculty–Student–AI: when governance is co-designed across the matrix, faculty are released from the impossible position of policing tool use unilaterally, and the classroom can become a setting in which faculty and students collaboratively pursue higher-order thinking with GenAI as a co-participant. The need for such fluency is sharpened by educators' own use of GenAI, for example, LLM-assisted evaluation of student project reports (Sun et al., 2026), which places educators on both sides of the matrix, i.e., authors of declarative rules as well as subjects of them. Classroom-level evidence points the same way. AI-generated feedback has been found to be practically equivalent to expert teacher feedback only when anchored to explicit, co-constructed criteria and supported by assessment literacy, i.e., when the surrounding pedagogical architecture, rather than the feedback source, does the trust

work (Grion et al., 2026). Notably, that comparability was demonstrated under research conditions in which students were blind to the feedback source; outside an approved research design, the declarative function of the Triple-A Framework would require precisely the disclosure that the study design suspended.

Five diagnostic questions, prompts for institutional reflection rather than pass-fail criteria, can be asked of any faculty development programme:

- 1. What has the institution declared, in writing, about GenAI use across teaching, assessment, research supervision, and student support?*
- 2. How is alignment with that declaration confirmed, by whom, and how often?*
- 3. What preventative system is in place at the curriculum and assessment level, and against which named foreseeable harms?*
- 4. How would the institution detect that something had gone wrong, drifted, or breached expectations, and how quickly?*
- 5. What does proportionate correction look like, and is the route to it clear to faculty and students alike?*

8 Considering Future AI Advancements

The framework should also extend beyond AI's current capabilities. It has been developed for, and illustrated with, GenAI systems that remain instances of artificial narrow intelligence (ANI). ANI systems perform impressively well within bounded task domains such as drafting, summarising, coding, and tutoring, but they neither generalise their competence across arbitrary domains nor set their own goals. Two further capability levels are commonly distinguished.

Artificial general intelligence (AGI) denotes systems able to match or exceed human performance across substantially the full range of cognitive tasks, transferring learning between domains without task-specific engineering (Goertzel, 2014). Artificial superintelligence (ASI) denotes systems that would surpass the best human performance in virtually every domain, including scientific creativity and social judgement (Bostrom, 2014). Today's GenAI sits at the upper end of ANI while borrowing the conversational surface of something more general, such as the anthropomorphic ambiguity noted in Section 2. It is reasonable to ask whether a framework built for the present regime retains its purchase as capabilities increase.

We contend that the framework's grammar persists even as the content of its cells shifts. Under ANI, the weight of trust work falls on the declarative and confirmative control functions: institutions can still specify acceptable use in advance and verify conformance afterwards. As systems approach AGI, that balance shifts. AGI's capabilities begin to outrun the evaluation instruments on which assurance depends, so confirmative control functions give way to preventative and detective mechanisms. At the same time, the need for enhanced accountability sharpens. The more autonomous the system, the less possible it becomes to locate fault in individual users rather than in deployment decisions. At the ASI horizon, the framework's mechanisms strain while its diagnostic value arguably peaks: an intelligence that exceeds its overseers challenges the premise of human control altogether. For this reason, preventative commitments made before deployment and the multi-stakeholder character of the matrix extended to international coordination would become the primary sites of trust work. For higher education, whose mission is the formation of human capability, each step up this ladder raises the stakes of

the question the matrix is designed to keep visible: who assures what, to whom, and who answers when assurance fails.

9 Research Agenda

The framework invites empirical work with the goal of assessing its content, construct, and predictive validity. The first task would be to assess its content validity. A Delphi study with faculty developers, institutional leaders, and regulators would assess whether the cells in Table 1 and risk typology in Table 2 are recognisable, complete, and parallel across institutional contexts. Such a study could include two rounds of assessment, stakeholders from a mix of institutional environments, and clear consensus thresholds. A faculty self-assessment instrument and an institutional implementation-and-evidence profile based on the matrix would enable universities to benchmark themselves and communicate progress to staff, students, and accreditors.

Second, a construct validity exercise could develop instruments to measure within each cell. Case studies and expert evaluations could form the foundational measures. Empirical work could answer whether the self-assessment instruments load onto the five distinguishable functions, and whether we have split the assurance/accountability appropriately. Such instrument development can build on adjacent work. The matrix can also support comparative policy analysis. Recent work demonstrates the utility of cross-jurisdictional comparison: Xu et al. (2025) map the United Kingdom's evolving digital education policy, while Zhang (2025) examines Wuhan University's institutional response to China's policy for the deeper integration of digital intelligence in higher education. Against this backdrop, the matrix offers a stable scaffold for comparing different national approaches—the EU AI Act, the Chinese and the UK strategies just cited, and the U.S. patchwork of agency guidance and state-level legislation—without flattening their differences. Cross-jurisdictional comparison is particularly important for examining how

differences in governance capacity, institutional resources, regulatory arrangements, and infrastructure shape the equity implications of GenAI adoption.

Third, empirical work could assess predictive validity. This work would ask whether institutions with more populated cells in Table 1 experience fewer AI incidents and respond faster. Metrics such as those presented earlier preview what such empirical work could look like and could serve as potential dependent variables. This empirical work would need to balance institutions across different locations, sizes, and levels of AI implementation.

10 Policy Agenda

On the policy side, the framework points to several practical steps. Sector bodies could publish model declarative templates for institutions to adapt locally. Minimum confirmative reporting requirements could be aligned with existing accreditation cycles. Sector trustmarks for both institutions and GenAI providers, building on the cloud trustmark concept (Lynn et al., 2013), could lower the cost of safe adoption for smaller institutions. Further, the matrix can be mapped to external regimes—the EU Artificial Intelligence Act, the UNESCO (Miao & Cukurova, 2024) *AI Competency Framework for Teachers*, and the OECD (OECD, 2019) AI principles—so that institutions are not forced to reconcile multiple frameworks manually. Finally, the framework sets clearer expectations for GenAI providers: universities can expect documentation, contractual commitments, secure-by-design controls, and forensic cooperation when things go wrong; providers can expect that institutional adoption follows declared rules and that accountability is administered through predictable channels.

The framework’s applicability across differently resourced higher-education systems warrants particular attention. In some lower-resource or institutionally fragmented contexts, sector-level assurance capacity may be limited, uneven, or organised differently from the

regulation or accreditation arrangements assumed here. In such settings, the matrix may initially be more useful as a gap and dependency map than as a benchmarking instrument. Relevant constraints may include infrastructure, specialist legal and technical expertise, contractual bargaining power, local-language system performance, and faculty workload. Some declarative actions can be comparatively low cost. In contrast, preventative mechanisms such as technical assessment, procurement review, impact assessment, accessibility evaluation, contractual negotiation, and independent audit may themselves require specialist capacity, which can be scarce and expensive even in the highly resourced higher education systems. Application of the Triple-A framework should therefore begin by identifying which assurance functions can reasonably be held institutionally and which require shared, sector-level, regional, or international support.

Further, at the faculty level, the matrix could also open a route to data-driven reform of professional capability evaluation. Evaluation schemes that measure only declarative outputs—policies issued, training completed—miss most of the capability the framework requires. An evaluation regime aligned to the five control functions could evidence faculty capability across declarative clarity, confirmative practice, preventative design, detective vigilance, and corrective contribution, with e-portfolios serving as the natural confirmative-function artefact. As AI itself improves, it can begin analysing data sources within universities, including annotated assessment samples, anonymised prompt logs, peer observation notes, and post-incident reflections (Uvarov, 2026). This is a substantively different evaluation regime from one that asks only whether faculty have used GenAI tools well.

Finally, to translate the framework into practice without making it prescriptive, we offer a sample workflow organised around control-function dependencies, rather than mandated mechanisms. Institutions first convene the relevant stakeholders and diagnose their current

governance landscape by identifying which cells in the matrix contain active mechanisms and which remain empty. They can then sequence gaps from declarative through corrective controls, treating this order as a matter of dependency rather than importance; map preventative mechanisms to the specific risks they address; select proportionate metrics and review cadences; and, finally, establish how corrective findings feed back into declarative commitments and trigger periodic re-diagnosis. Figure 2 illustrates this iterative process.

Dependency order should not be confused with risk priority. Where several gaps are identified simultaneously, institutions should give immediate priority to non-discretionary legal or regulatory obligations and risks involving imminent or potentially severe harm. Remaining gaps should be prioritised using a documented assessment of impact severity, likelihood, scale of exposure, effects on vulnerable or disadvantaged groups, reversibility, detectability, and the effectiveness of existing controls. Control-function dependencies can then inform the sequencing of interventions within that risk-prioritised set. A high-severity detective or corrective gap may therefore properly take precedence over a lower-severity declarative gap. This approach is consistent with risk-based prioritisation frameworks such as the *NIST AI RMF* (Tabassi, 2023), which specifies that treatment of documented AI risks should be prioritised according to impact, likelihood, and available resources or methods.

*** Insert Figure 2 about here *** (此处插入图 2)

11 Limitations

We recognise limitations along two dimensions: limitations arising from development and structure. Both bound our claims.

Our framework has inherent limitations that arise in how it was developed. First, the framework has not yet been empirically validated; the Delphi and instrument-development work

described above is the necessary next step. Second, although we argue that trustworthy GenAI is a co-designed property of the ecosystem, the framework itself was not co-designed with students or other stakeholders, whose interests enter here only by proxy. Third, our regulatory reference points are drawn from the EU, the UK, the US, and China; jurisdictions with different governance traditions may populate and weight the five control functions differently. Finally, we recognise that this framework can inform a future with AGI/ASI, but that world might look drastically different. As such, we recommend formally updating this framework periodically as these new technologies emerge.

Our framework also has three limitations that stem from how we structured it. First, we present each dimension—stakeholder group, control function, etc.—as categorical. In reality, these boundaries are porous. We recognised earlier that some mechanisms (e.g., a log) can serve multiple functions depending on use. Second, the stakeholder rows we present as homogeneous are, in fact, heterogeneous. For example, a category such as ‘students’ comprises undergraduates and doctoral students, who operate in distinct ways. While this is done to balance specificity with presentability, readers should exercise caution when applying the framework without such considerations. Third, we present our lifecycle as linear with clear feedback loops. The reality is likely much messier. Institutions are simultaneously working on different parts of the process presented in Figure 1. While it might be more appropriate to work through the flow as presented, we recognise this is unlikely to happen neatly.

12 Conclusion

Trustworthy GenAI in higher education is not a property of any single stakeholder. It is a co-designed property of the whole ecosystem (students, educators, institutions, technology providers, policymakers, regulators, and accreditation bodies) acting in concert across a lifecycle

that begins with declared rules and ends with corrective learning, which then revises those rules, which in turn revises what is declared next. Future faculty development matters, but its leverage depends on the institutional grammar of assurance and accountability. The Triple-A Framework we propose is one attempt to make that grammar concrete. In framing the matrix in these pages, we follow the journal's call for paradigm-shifting accounts of human–AI collaboration in higher education (Wu & Chen, 2025) while offering a grammar against which those accounts can be evaluated, instrumented, and operationalised.

The work ahead is not to wait for clearer policy from regulators or better tools from providers. Both will help, but neither is sufficient. The work ahead is to build the institutional muscle that enables any university to declare clearly, confirm honestly, prevent thoughtfully, detect promptly, and correct proportionately. That muscle is the same whether the GenAI in question is a chatbot, a grading assistant, or, eventually, something broader. These claims remain provisional: Triple-A has not yet been empirically validated or co-designed with the full set of stakeholders it represents, and its categorical stakeholder and control-function structure will require testing and adaptation across institutional, cultural, and regulatory contexts. We invite colleagues to test, contest, and refine the framework and to share their findings so the underlying grammar can continue to develop in public. In doing so, our framework provides a basis for future studies on trust in educational AI, serves as a design tool for institutional governance, and acts as a common reference for policy discussions, enabling higher education to influence the direction of GenAI instead of just adapting to it.

Acknowledgments The authors did not receive support from any organisation for the submitted work.

Declaration on the Use of AI AI was not used in the preparation of this manuscript.

Author Contributions Timothy David Hubbard and Theo Lynn: Conceptualisation; methodology; formal analysis; investigation; writing of original draft; writing–review & editing.

Conflict of Interest The authors have no conflict of interests to declare.

Data Availability Statements The authors confirm that data was not collected for this article.

References

- Adu, K. K., & Owusu-Agyeman, Y. (2026). A systematic review of the dark side of AI in education: A critical social theory perspective. *Frontiers of Digital Education*, 3(2), 10.
- Bachmann, R., Gillespie, N., & Priem, R. (2015). Repairing trust in organizations and institutions: Toward a conceptual framework. *Organization Studies*, 36(9), 1123–1142.
- Biden, J. R. (2023). *Safe, secure, and trustworthy development and use of artificial intelligence*. Presidential Document EO 14110. United States Government Publishing Office.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- California Legislature. (2024). *Senate Bill No. 1288: Public schools: Artificial intelligence working group*. California Legislative Information.
- Contractor, Z., & Reyes, G. (2025). Generative AI in higher education: Evidence from an elite college. *arXiv Preprint*, arXiv:2508.00717
- Digital Education Council. (2025, January 20). *Digital education council global AI faculty survey 2025*. Available from Digital Education Council website.
- Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6.
- Gillespie, N., & Dietz, G. (2009). Trust repair after an organization-level failure. *Academy of Management Review*, 34(1), 127–145.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1–48.

Grion, V., Doria, B., Agostini, D., & Slaviero, G. (2026). Artificial intelligence and feedback in university education: Effectiveness and student perceptions. *Assessment & Evaluation in Higher Education*. (in press).

International Organization for Standardization. (2023). Information technology—Artificial intelligence—Management system. ISO/IEC 42001:2023. International Standard published.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology*, 89(1), 104–118.

Krause, S., Panchal, B. H., & Ubhe, N. (2025). Evolution of learning: Assessing the transformative impact of generative AI on higher education. *Frontiers of Digital Education*, 2(2), 21.

Lau, S., Boroojeni, K., Keeling, H., & Marroquin, J. (2026). Barriers that programming instructors face while performing emergency pedagogical design to shape student-AI interactions with generative AI tools. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Barcelona: ACM, 528.

Lynn, T., Healy, P., McClatchey, R., Morrison, J., Pahl, C., & Lee, B. (2013). The case for cloud service trustmarks and assurance-as-a-service. In: *Proceedings of the 3rd International Conference on Cloud Computing and Services Science*. Aachen: SciTePress, 110–115.

Lynn, T., Van Der Werff, L., & Fox, G. (2021). Understanding trust and cloud computing: An integrated framework for assurance and accountability in the cloud. In: Lynn, T., Mooney, J. G., Van Der Werff, L., & Fox, G., eds. *Data privacy and trust in cloud computing*. Palgrave Macmillan, 1–20.

Ma, Y., Su, Y. X., Li, M. D., Zhang, Y., Chai, W. T., Huang, A. M., & Zhao, X. F. (2025).

Preparing students for an AI-driven world: Generative AI and curriculum reform in higher education. *Frontiers of Digital Education*, 2(4), 30.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2(2), 12.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review*, 23(3), 473–490.

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO.

Ministry of Education of the People's Republic of China. (2018). *Artificial Intelligence Innovation Action Plan for Institutions of Higher Education*. Ministry of Education of the People's Republic of China. (in Chinese).

Ministry of Education of the People's Republic of China. (2024). *Guidance on Strengthening Artificial Intelligence Education in Primary and Secondary Schools*. Ministry of Education of the People's Republic of China. (in Chinese).

Mobilio, V., & Guglielmini, G. (2026). Rethinking schooling in the age of AI: Equity, ethics, and the four pillars of transformation. *Frontiers of Digital Education*, 3(1), 7.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. W. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, 5, 100151.

OECD. (2019). *Recommendation of the council on artificial intelligence*. OECD/LEGAL/0449.

OECD.

Pascoe, C., Quinn, S., & Scarfone, K. (2024). *The NIST cybersecurity framework (CSF) 2.0*.

NIST CSWP 29. National Institute of Standards and Technology.

Pesek, I. (2026). Rethinking education governance in the age of AI. *Frontiers of Digital*

Education, 3(1), 11.

Rafi, M. S., & Khalil, L. (2026). Generative AI in doctoral dissertation writing: Applications,

limitations, and the need for prompt literacy. *Frontiers of Digital Education*, 3(2), 14.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A

cross-discipline view of trust. *Academy of Management Review*, 23(3), 393–404.

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter

for information systems users. *European Journal of Information Systems*, 25(3), 274–287.

Sun, Q. Y., Zhang, J. L., Sheng, P., Wang, Q. Y., Wang, T. R., Li, H., Zhan, H. S., Zhang, X. Q.,

& Liu, J. (2026). Evaluating the efficacy of a multifaceted prompt for use with LLMs to

evaluate course project reports. *Frontiers of Digital Education*, 3(2), 12.

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. NIST

Trustworthy and Responsible AI, National Institute of Standards and Technology.

The European Parliament and the Council of the European Union. (2024). *Regulation (EU)*

2024/1689 of the European Parliament and of the Council laying down harmonised rules on

artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

The European Parliament and the Council of the European Union. (2026). *Regulation (EU)*

2026/1744 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as

regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union.

Tyton Partners. (2024, June 11). *Time for class 2024*. Available from Tyton website.

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. U.S. Department of Education.

Uvarov, A. (2026). Developing indicators for school digital renewal in the age of AI. *Frontiers of Digital Education*, 3(1), 6.

Wu, D., Sun, X. J., Liang, S. B., Qiu, C., & Wei, Z. Y. (2025). Construction of AI literacy evaluation system for college students and an empirical study at Wuhan University. *Frontiers of Digital Education*, 2(1), 6.

Wu, F., & Chen, J. Y. (2025). Collaboration of generative AI and human: Paradigm shift for higher education. *Frontiers of Digital Education*, 2(2), 24.

Xu, J. L., Luo, M. Y., & Zhang, X. X. (2025). Towards AI: The evolution of digital education policy in the United Kingdom. *Frontiers of Digital Education*, 2(4), 33.

Zhang, P. W. (2025). Leading the digital transformation of higher education through the reform of digital intelligence education: Exploration and practice at Wuhan University. *Frontiers of Digital Education*, 2(1), 2.

表 1 Table 1 The Triple-A Framework’s 5 stakeholder × 5 control function matrix

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
Students	Objective: students understand institutional expectations, permissions, and prohibitions for GenAI use in their studies Example mechanisms: course-level GenAI statements; disclosure	Objective: declared rules are observable in student practice Example mechanisms: submission checklists; declared-use cover sheets; retained process artefacts (drafts, version	Objective: reduce the likelihood of named student-side harms—undisclosed reliance, sensitive data disclosure, fabricated citations Example mechanisms: AI literacy targeted at named risks; warnings before	Objective: identify named-harm signals—fabricated citations, authorship discontinuity, undeclared use—when they emerge Example mechanisms: case investigations triggered by anomaly (citation	Objective: restore proportionate outcomes and student understanding after failure Example mechanisms: explanation of finding; learning intervention; resubmission/viva/mark adjustment; misconduct route; appeal route Evidence: documented outcomes with

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
	templates; worked examples in module handbooks; learning agreement language Evidence: acknowledged learning agreements; updated module handbooks; orientation materials	history, oral defence) Evidence: completed checklists; declared-use records; process artefacts retained with submissions	sensitive data upload; process- evidence requirements built into assessment Evidence: literacy assessments aligned to named risks; data-warning logs; required- process artefacts	clusters, viva inconsistency, abnormal pattern); not routine surveillance Evidence: investigation records linking signal to outcome	reasoning; appeal records; feedback into module redesign

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
Educators	Objective: express course-level rules on GenAI use, aligned to institutional and accreditation expectations Example mechanisms: course-level GenAI statements; assessment-brief language;	Objective: show that declared rules are operating—students aware, materials aligned, approved tools in use Example mechanisms: classroom acknowledgement; rubric audits; peer observation of assessment design; calibration with	Objective: reduce named foreseeable harms—undisclosed outsourcing in assessment, hallucinated citations, sensitive-data leakage in feedback Example mechanisms: assessment redesign targeted at named risks;	Objective: recognise signals of breach, drift, or named harm in student work or classroom practice Example mechanisms: case review triggered by abnormal signals (citation patterns, authorship continuity, rubric	Objective: repair harm at module level—affected students, assessment design, peer learning Example mechanisms: post-incident reflection; module-level redesign; peer dissemination; engagement with institutional process Evidence: redesigned assessments; module

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
	supervision handbooks; rubrics Evidence: published course statements; updated assessment briefs; supervision agreements	department guidance Evidence: acknowledgement records; rubric audit notes; peer observation reports	approved-tool restrictions on feedback workflows; pre- publication checks on AI-assisted feedback Evidence: assessment- redesign documentation; approved-tool use logs; feedback- review records	drift, classroom incident reports) Evidence: case- review records that distinguish anomaly from confirmation	reports referencing lessons learned

Continued...

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
Institutions	Objective: publish the institutional position on GenAI across teaching, research, IP, accessibility, integrity, and human oversight Example mechanisms: institutional GenAI policy; governance roles and accountability matrix; approval	Objective: show institutional practice aligns with declared policy across teaching, procurement, and research Example mechanisms: governance reporting; approved-tool catalogues; DPIA completion tracking; training-	Objective: reduce foreseeable GenAI-related harms in assessment, procurement, and data governance before deployment or use Example mechanisms: assessment design review; DPIA gate; approved-tool access controls; contractual review	Objective: identify failure signals—data leakage, unauthorised tool deployment, accessibility failure, bias complaint, security incident Example mechanisms: anomaly detection on enterprise systems; incident reporting channels; complaint triage;	Objective: restore trust through proportionate response, remediation, and learning Example mechanisms: incident response; affected-party notification; suspension of unsafe workflows; policy revision; staff retraining;

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
	routes; published codes of practice Evidence: published policy; governance committee records; published code of practice	completion records; compliance dashboards Evidence: governance reports; DPIA registers; tool-catalogue updates; training completion data	before deployment; red-teaming for high-risk uses Evidence: completed design reviews; DPIA records; procurement sign- off; access-control configuration; red- team reports	security monitoring; periodic targeted reviews of high-risk uses Evidence: incident logs; complaint registers; security alerts; targeted- review reports	accessibility remediation; procurement review Evidence: incident response records; notification logs; policy revisions; training delivery; remediation completion
Technology Providers	Objective: publish capabilities, limits, terms, and	Objective: show that service operation aligns	Objective: reduce the likelihood of named harms—data	Objective: identify security incidents, model failures, SLA	Objective: remediate provider- side failures and

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
	contractual commitments relevant to educational deployment Example mechanisms: product documentation; service descriptions; contract clauses on data processing, retention, training-data use, accessibility,	with declared commitments Example mechanisms: SLA reporting; security control reporting; accessibility conformance statements; model-change notifications; vulnerability management reporting	leakage, prompt injection, unsafe outputs, bias drift, service degradation, unauthorised access Example mechanisms: secure-by-design controls; tenant isolation; abuse monitoring; anomaly detection; configurable institutional	breaches, or harmful outputs at provider level Example mechanisms: provider-side monitoring; abuse detection; vulnerability scanning; forensic logging; customer-incident interface Evidence: incident logs; forensic	restore service quality Example mechanisms: patch; rollback; model or feature suspension; customer notification; service credits; contract remediation; updated documentation

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
	uptime, auditability, content safety Evidence: published documentation; signed contractual commitments; statements of service limits	Evidence: SLA reports; security attestations; accessibility VPATs; model- change logs	guardrails; privacy- preserving logging Evidence: security architecture documentation; guardrail configuration records; monitoring dashboards; penetration-test reports	reports; customer- facing incident notifications	Evidence: change records; customer notifications; remediation receipts; SLA credit records

Continued...

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
Policymakers, Regulators & Accreditation	Objective: set sector-level standards and minimum expectations for transparency, privacy, accessibility, integrity, records retention, human oversight, and high-risk educational uses Example mechanisms:	Objective: confirm institutional and provider compliance with sector-level standards Example mechanisms: accreditation cycles; conformity assessment; sector trustmark schemes; mandatory	Objective: reduce the likelihood of systemic, named harms in high-risk educational uses—admissions decisions, grading, progression, proctoring, support triage Example mechanisms: guidance for high-risk uses; regulatory sandboxes; sector	Objective: identify systemic harms, discriminatory outcomes, unlawful surveillance, data misuse, inaccessible systems, or misleading assurance claims Example mechanisms: complaints handling; sector-level investigations; whistleblower	Objective: verify corrective action across institutions and providers; revise standards in light of failure Example mechanisms: sanctions; binding remediation orders; updated standards, model clauses, procurement rules, accreditation

Stakeholder	Declarative	Confirmative	Preventative	Detective	Corrective
	regulations; sector codes; accreditation standards; published guidance; model clauses; transparency obligations Evidence: published regulations; accreditation standards; sector codes; guidance documents	reporting; public registers Evidence: accreditation reports; conformity assessment outcomes; trustmark issuance; sector-level transparency reports	monitoring; early-warning networks; restrictions on prohibited uses Evidence: high-risk-use guidance; sandbox participation records; sector monitoring outputs; restriction notices	channels; cross-jurisdictional information sharing; periodic sector audits Evidence: investigation reports; complaint registers; sector audit findings	requirements, reporting thresholds Evidence: enforcement actions; updated standards; revised model clauses; updated accreditation requirements

Notes: This table provides a detailed overview of the proposed framework that connects stakeholders to the five control functions described in the paper. Each intersection outlines the objectives, an example mechanism, and potential evidence that stakeholders can consider to achieve the goals of a particular cell. Cells state control objectives with illustrative mechanisms and evidence; they are not exhaustive. DPIA: Data Protection Impact Assessment, IP: Intellectual Property, SLA: Service Level Agreement, VPAT: Voluntary Product Accessibility Template

表 2 Table 2 Illustrative risk typology traced across the lifecycle. Preventative functions anticipate the named risk; detective functions identify signals that it has materialised; corrective functions repair the harm and revise the system

Named risk	Preventative (anticipate)	Detective (signals of materialisation)	Corrective (repair and revise)
Undisclosed GenAI use in assessment	Process-evidence assessment design (drafts, version history, oral defence); disclosure templates; declared-use cover sheets	Fabricated citation clusters; authorship discontinuity; inconsistency between submission and demonstrated capability at viva	Proportionate academic-integrity outcome; assessment redesign; revision of module-level guidance
Sensitive or third-party data disclosed to public tools	Approved-tool access controls; pre-upload warnings; DPIA gates; privacy-safe workflows	Abnormal data egress; single sign-on anomalies; unapproved tool traffic	Tool suspension; affected-party notification; retroactive DPIA; procurement- gate revision
Fabricated or inaccurate AI- generated	Pre-publication checks on AI-assisted materials and feedback; approved	Student complaints; citation-validation	Correction of materials, feedback, or grades; notification of affected

content in teaching or feedback	feedback workflows with human oversight	failures; error reports on published materials	students; staff development
Biased or inconsistent AI-mediated grading	Human oversight requirements; re-teaming of high-risk uses; contractual inspection and audit rights at procurement	Complaint trends; rubric drift; audits of grading outputs across cohorts and demographic groups	Regrading; provider remediation under contract; revision of standards, oversight, and procurement clauses
Unauthorised tool procurement or deployment	Procurement gate; approved-tool catalogue linked to access controls and contractual restrictions	Anomaly detection of unapproved deployments in enterprise systems; licence and integration audits	Deployment suspension; contract remediation; enforcement and revision of the procurement gate
Cognitive offloading and deskilling	Assessment designs requiring demonstrated process and unaided capability; AI-literacy targeted at this named risk	Declining performance on unaided or in-person tasks; divergence between assisted and unassisted work	Curriculum and assessment revision; learning interventions; revised declarative guidance on permitted assistance

Overtrust and relational reliance on anthropomorphic GenAI	Interface and deployment choices that calibrate rather than inflate trust; AI literacy addressing anthropomorphic cues	Uncritical acceptance of outputs; declining verification behaviour; use of GenAI for pastoral or wellbeing support outside approved channels	Recalibration interventions; revised deployment configuration; updated declarative guidance on appropriate reliance
--	--	--	---

Notes: This table provides a sample of risks this framework may be well suited to address. Each risk is outlined with samples of how the preventative function links trust building (assurance) and trust repair (detective and corrective). Detective signals are indicators warranting proportionate human review; they should not be treated individually as proof of GenAI use or academic misconduct. Adverse academic decisions should rely on corroborated evidence, transparent reasoning, and an accessible review or appeal route. DPIA: Data Protection Impact

Assessment

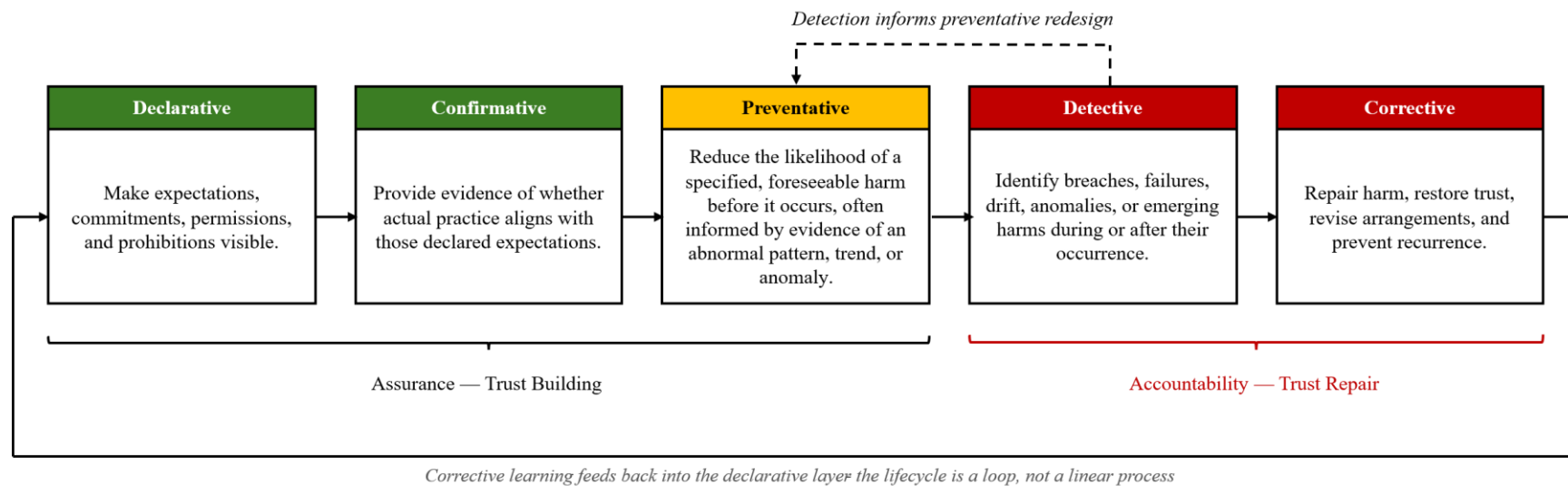
表 3 Table 3 Mapping of control functions to trust issues, roles in the trust lifecycle, and trust antecedents

Control function	Trust issue foregrounded	Role in the trust lifecycle	Primary trust antecedents signalled
Declarative	Opacity of expectations: parties cannot trust what they cannot see committed to	Trust building	Integrity (alignment to stated principles); structural assurance (rules of engagement)
Confirmative	Unverifiable promises: declarations without evidence allow word and practice to decouple	Trust building	Integrity (promise-to-practice consistency); ability (demonstrated competence)
Preventative	Foreseeable but unaddressed harm: known risks left undesigned-against erode confidence before any failure occurs	Bridging: trust building and pre-emption of trust repair	Ability (technical and procedural competence); benevolence (anticipating harm); situational normality
Detective	Hidden failure: breaches and drift that go	Trust repair (trigger): surfaces the violation that initiates repair	Integrity (surfacing failure honestly);

Control function	Trust issue foregrounded	Role in the trust lifecycle	Primary trust antecedents signalled
	unsurfaced make repair impossible		structural assurance (predictable oversight)
Corrective	Unrepaired violation: harm without proportionate response converts a single incident into durable distrust	Trust repair	Benevolence (intent to make affected parties whole); integrity (transparent response); structural assurance (predictable redress)

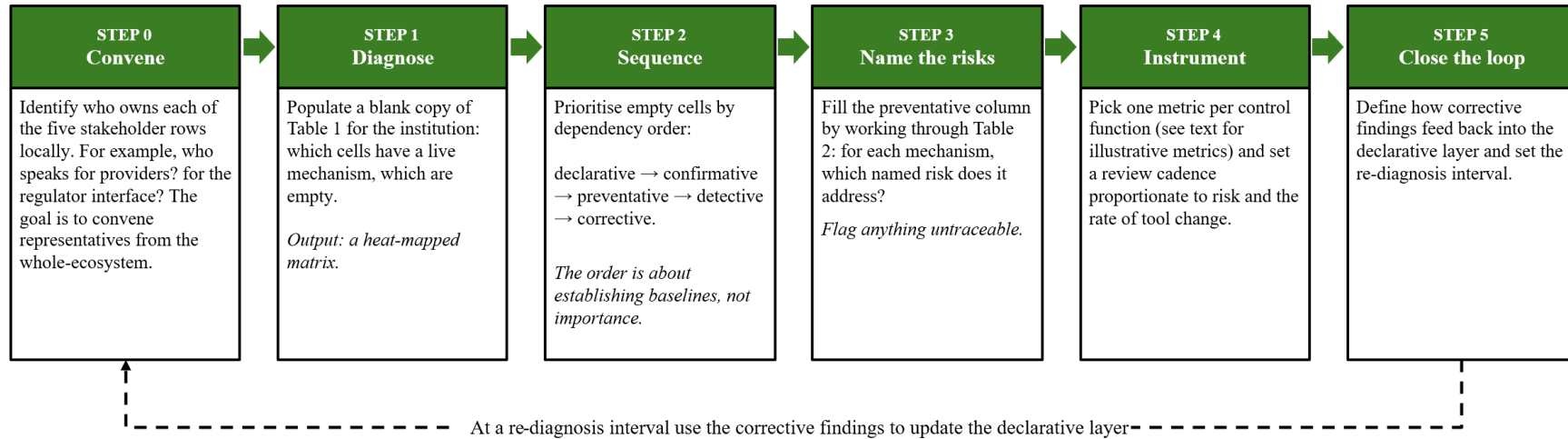
Note: This table is intended to link each control function to the broader trust lifecycle it influences. The table specifies the antecedents of trust that each control function relates to. These mappings are intended to show how the control functions support trust throughout its lifecycle (building and repair) and the signals people use to evaluate trustworthiness (ability, benevolence, and integrity).

图 1 Figure 1 The five control functions across the assurance and accountability lifecycle. The preventative function bridges trust building (assurance) and trust repair (accountability); detection informs preventative redesign, and corrective learning feeds back into the declarative layer.



Note: This figure outlines the framework's flow in relation to trust building and trust repair among stakeholders. It also defines the feedback loops that connect the detective back to the preventative function and the corrective back to the declarative control function. These feedback loops are critical in practice to keep the framework's implementation up to date for those implementing it.

图 2 Figure 2 A six-step illustrative process for implementing the framework.



Note: This figure provides a basic overview of how to implement the process of adopting the framework. It is intentionally broad and is meant only to start a discussion among a group of stakeholders about how to begin enacting the proposed framework.