

From Agent Labels to Accountable Deployment: Evidence Maturity and Meaningful Human Oversight for Agentic AI in Higher Education

Abhishek Dey

hello.deyabhishek@gmail.com

South Ural State University

Systematic Review

Keywords: agentic AI, human oversight, AI governance, higher education, accountability, evidence maturity

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11020094/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Agentic artificial intelligence is entering higher education as systems progress from generating answers to planning and executing actions. This transition creates an ethical problem that aggregate accuracy cannot resolve: institutions may delegate consequential discretion before they possess evidence of durable benefit, fairness, or effective human control. This critical integrative review examines 15 peer-reviewed reviews, one recent preprint retained separately, and selected primary studies to determine what the evidence supports and what remains ethically unjustified. The literature is conceptually unstable, often conflating perceived agency, static generative AI, and systems with goal-directed execution. Evidence is strongest for bounded artifact quality and weakest for durable learning, equity, workload, and institution-level outcomes. Authentic deployments remain mostly short, single-site, and weakly connected to tested oversight arrangements. The article develops the Autonomy-Oversight-Evidence framework, comprising five autonomy levels, five oversight levels, and six stages of evidence maturity. Its normative principle is that delegated autonomy must not exceed either the maturity of the supporting evidence or the strength of accountable human oversight. This principle follows from duties to prevent avoidable harm, preserve learner autonomy, distribute risks fairly, maintain contestability, and keep institutional accountability with responsible persons. The framework converts these duties into reporting requirements, permission boundaries, approval checkpoints, logging, escalation, appeal, and rollback. Agentic AI may improve selected educational processes, but current evidence does not justify broad claims of sustained effectiveness or autonomous use in rights-affecting decisions.

1. Introduction

Generative AI in higher education is moving from answering questions to taking actions. Emerging systems can decompose goals, choose tools, retrieve information, maintain state, evaluate intermediate outputs, and coordinate specialized agents. This shift matters because an answer-generating system and an action-taking system create different educational opportunities and failure surfaces, and we need to be aware of the difference. A chatbot may produce an inaccurate explanation; an agent embedded in assessment, advising, admissions, or research workflows may also propagate that error, expose data, trigger downstream actions, or obscure who is accountable.

The literature has expanded accordingly. Recent reviews map generative AI agents across educational levels, trace the transition from rule-based to goal-oriented systems, catalog architectures and pedagogical roles, examine instructor practice, and synthesize institutional responses [1–4]. A dedicated systematic review of agentic AI in higher education now reports promising academic, engagement, and efficiency outcomes alongside risks [5]. That publication changes the useful question. Another broad inventory of applications would duplicate an emerging synthesis space. The more consequential unresolved problem is whether evidence, autonomy, and oversight are being evaluated on the same scale.

Three forms of inflation make the answer difficult. First, label inflation occurs when any AI interface, chatbot, avatar, or automated workflow is described as agentic. Second, outcome inflation occurs when model metrics or user perceptions are generalized to learning, equity, or institutional effectiveness. Third, maturity inflation occurs when a prototype, a brief classroom exercise, and a semester-long deployment are treated

as equivalent evidence. These problems are not simply terminological. They influence procurement, research design, risk classification, and the degree of discretion an institution delegates to a system.

This article therefore offers a critical integrative review and a normative governance framework, not a new claim of exhaustive primary-study coverage. It asks four questions: (1) what changes when agentic AI is defined by demonstrated execution capabilities; (2) what kinds of outcomes are presently supported; (3) when does human oversight constitute effective control rather than ceremonial approval; and (4) what ethical conditions must be satisfied before institutions delegate educational discretion? The contribution is the Autonomy-Oversight-Evidence (AOE) framework. It connects empirical uncertainty to duties of non-maleficence, respect for learner autonomy, justice, accountability, and contestability. Its central claim is that institutions cannot ethically justify more autonomy than their evidence and oversight arrangements can jointly support.

2. Ethical stakes and evidentiary obligations

Agency is not a property conferred by a name. For this review, an AI system is treated as agentic when it pursues a goal through an execution loop and demonstrates at least two of the following capabilities: planning or task decomposition, contextual or persistent memory, external tool use, reflection or iteration, and multi-agent coordination. The definition is intentionally behavioral. It excludes one-shot prompting, ordinary content generation, predictive analytics, and dialog-based systems that only react to the next user turn. It also avoids assuming that a system is agentic because users perceive it as independent.

This boundary has methodological consequences. Static systems are normally evaluated at the level of their outputs and interactions. Agentic systems must additionally be evaluated as processes: what goals they receive, which actions they may take, which data and tools they can access, how intermediate states are logged, when they stop, and how people can interrupt or reverse their actions. Multi-agent systems introduce coordination failures, such as error amplification, role confusion, circular validation, and false confidence (Hallucination or Context Snowballing), arising from agreement between participants who share the same underlying model or data.

Educational evidence must also be separated by level. Improvements in answer accuracy, rubric alignment, or feedback style are artifact outcomes. Changes in knowledge, transfer, metacognition, motivation, or cognitive demand are learner outcomes. Time savings, workload redistribution, professional judgment, and adoption are educator outcomes. Admissions, progression, resource assignment, compliance, and cost are institutional outcomes. A favorable artifact result can be necessary for a learner benefit, but it is not sufficient evidence of one. Similarly, high acceptance is not evidence that delegation is safe or that learning is durable.

These distinctions carry ethical weight because educational institutions exercise authority over learning, assessment, access, progression, and the distribution of opportunity. When an institution delegates part of that authority to an agentic system, it does not delegate its responsibility for foreseeable harms. Average performance cannot by itself legitimate a deployment if errors are concentrated among particular groups,

affected persons cannot understand or challenge decisions, or staff lack the information and authority required to intervene.

Five normative commitments therefore guide the analysis. Non-maleficence requires limiting avoidable educational, informational, and procedural harm. Respect for autonomy requires preserving meaningful learner and educator agency rather than disguising delegation as assistance. Justice requires examining how benefits, errors, surveillance, and verification labor are distributed. Accountability requires a responsible person or institution with both knowledge of the system and power to correct it. Contestability requires reasons, review, and accessible remedies when an AI-enabled process affects a person's educational interests. These commitments make uncertainty ethically relevant: weak evidence is not merely a research limitation when institutions use it to justify consequential delegation.

3. Review design

3.1 Scope and source set

The review used a critical integrative design to compare the assumptions, boundaries, and conclusions of recent secondary research and to connect them to selected empirical exemplars. The analytic set comprised the 15 peer-reviewed reviews listed in Appendix A. The source set was assembled purposefully and updated through targeted, citation-led searching rather than an exhaustive database search. Records were eligible when they examined agentic AI, AI agents, or closely related generative-AI systems in education; higher-education-specific evidence had to be identifiable for higher-education-focused coding. Publisher, DOI, or institutional metadata were verified through 4 September 2026. Baradziej [5] and Saleem [6] were added in the 4 September update. Li and Zheng [7], a preprint, and Lan and Zhou [8], an adjacent self-regulated-learning review, were retained only as contextual sources. The author screened and coded the sources; no second-reviewer reliability estimate was available.

This design does not claim the coverage of a prospectively registered systematic review of primary studies. It provides a transparent, current critique of the review landscape and uses purposefully selected primary studies to examine the evidentiary chain from agent capability to educational or institutional outcome. The distinction is important because a prospectively registered, multi-database primary-study review would be required to estimate effects and prevalence.

3.2 Coding logic

Each review was coded on six questions: whether higher education was exclusive or separable; whether systems met the behavioral agentic definition; whether LLM-based and multi-agent capabilities were distinguished; whether measured outcomes were synthesized; whether evidence came from authentic educational or institutional deployment; and whether a named appraisal procedure was applied. Findings were synthesized by comparison rather than pooled because the reviews used different definitions, populations, units of analysis, and publication windows. Primary examples were chosen because they illuminate recurring inferential problems: artifact quality versus learning, perceived versus demonstrated agency, and short deployment versus mature evidence.

3.3 Interpretive safeguards

Four safeguards shaped the synthesis. First, a review's use of the word agent did not determine classification. Second, dual screening or joint coding was not treated as risk-of-bias appraisal. Third, positive directional findings were not converted into a common effect estimate. Fourth, absence language was bounded to the verified source set. Accordingly, this article states that evidence is limited or not systematically synthesized; it does not claim that no evidence exists.

3.4 Generative AI assistance

Generative AI tools assisted with literature organization, drafting, and language editing. The author retained responsibility for the inclusion criteria, source verification, coding decisions, interpretation, normative argument, and final text. No AI system is listed as an author.

4. Findings

4.1 A crowded secondary literature with unstable boundaries

The review landscape is already substantial. Wang et al. [4] mapped 474 publications on generative-AI-powered agentic AI across education and organized the field around architectures, roles, theories, and agentic capabilities. Kamalov and Sulieman [1] tracked the progression from rule-driven systems to goal-oriented and multi-agent designs. Kostopoulos et al. [2] supplied a technology-oriented state-of-the-art survey. Liu et al. [3] synthesized 33 studies across macro, meso, and micro levels. These sources establish that agentic education is no longer a speculative niche.

However, the same corpus shows that scale is sensitive to definition. Reviews of collaborative learning, pedagogical agents, broad generative AI adoption, learner agency, or institutional governance all contain relevant material, but not every included system achieves its goals autonomously. Ba et al. [9], for example, offer a valuable synthesis of AI agents in computer-supported collaborative learning across 46 empirical studies, while the category spans technologies developed before contemporary tool-using LLM agents. Reviews of generative AI in higher education synthesize authentic implementation and perception data but often concern static interaction rather than delegated execution [10–12]. Conflating these literatures enlarges the apparent evidence base while weakening causal and governance conclusions.

Table 1
Boundaries and contributions of the highest-relevance reviews

Review	Scope	Agent boundary	Outcomes	Appraisal	Primary contribution
Wang et al. [4]	Mixed levels	Strict	Partial	No	Capabilities and research landscape
Kamalov and Sulieman [1]	Mixed levels	Strict	Partial	Unclear	Architectural transition
Kostopoulos et al. [2]	Mixed levels	Strict	Partial	No	Taxonomy and technology
Liu et al. [3]	Mixed levels	Strict	Yes	Yes	Roles, theory, future agendas
Baradziej [5]	HE	Strict/broad	Yes	Not central	Roles and reported effects
Doszhanov et al. [13]	HE	Partial	Yes	Yes	Instructor professional practice
Elsayed et al. [14]	HE	Partial	Yes	Partial	Human–AI teaching synergy
Ba et al. [9]	Mixed levels	Broad agent	Yes	Yes	Collaborative learning
G. Wang et al. [15]	HE	Static GenAI	Yes	No	Offloading and learner agency
Bacalso et al. [16]	HE	Static GenAI	Partial	No	Institutional governance

4.2 Evidence is strongest closest to the artifact

The clearest empirical advantages concern bounded outputs. Guo et al. [17] compared a two-agent feedback system with a single-agent approach across 240 middle-school science responses. The multi-agent design reduced over-praise from 15.42% to 1.25%, over-inference from 27.20% to 7.08%, and their co-occurrence from 9.58% to 0.83%. Lai and Li [18], evaluating 800 feedback outputs derived from 200 undergraduate essays, similarly reported lower error and hallucination rates for a tri-agent strategy than for zero-shot single-agent prompting. These studies show that role separation and structured checking can improve feedback artifacts under defined conditions.

Their interpretation should stay bounded. Better feedback does not automatically lead to better learning. Neither evaluation establishes whether students understood, used, or transferred the feedback, whether instructors saved time after verification, or whether performance gains persisted. The multi-agent architecture may also increase latency, cost, and the risk surface for privacy or orchestration failures. The appropriate claim is therefore that agent decomposition improved selected feedback-quality metrics—not that multi-agent systems are generally educationally superior.

4.3 Learner outcomes are promising but context-dependent

More direct evidence of learning is beginning to appear. Hao et al. [19] analyzed data from 305 university students and 19,365 dialogue lines in a multi-agent learning environment, using pre/post measures. Students with lower prior knowledge showed comparatively strong learning gains and motivation, and technology acceptance increased. Wang et al. [20] reported that AI-agent-supported collaborative learning in university programming was associated with higher achievement, self-efficacy, and interest, and lower mental effort than the comparison condition, although mental effort did not differ. Such studies move beyond satisfaction and model effectiveness.

Even here, the evidence does not yet support a universal effect. Results are embedded in particular disciplines, task structures, scaffolds, durations, and student populations. Interaction logs can reveal productive patterns, but they do not by themselves establish if students are developing independent competence or becoming dependent on delegated cognitive work. Reviews of cognitive offloading and learner agency warn that generative AI may support self-regulation under certain conditions and weaken it under others [15]. Agentic systems intensify that tension because they can complete a larger share of the planning, monitoring, and revision cycle that learners are expected to internalize.

4.4 Perceived agency is not demonstrated autonomy

Alqurni [21] reported a survey of 280 higher-education students in Saudi Arabia; the demographic analysis used 256 cases, while other sections of the article report differing sample figures. Perceived agency predicted perceived usefulness, ease of use, and autonomy support; self-efficacy and autonomy support, in turn, predicted self-learning motivation. The study is valuable for understanding adoption and motivational pathways. It should not, however, be cited as evidence that a technically autonomous agent caused better learning. The key independent construct is perceived agency. This distinction illustrates a broader measurement problem: mental reactions to an AI's presentation can be meaningful yet remain separate from the system capabilities that pose operational risk.

4.5 Authentic deployment remains shallow

Some research has entered genuine university workflows. Nguyen-Duc et al. [22] evaluated a multi-agent retrieval-augmented system for admissions counseling at a Vietnamese university using 6,079 question-answer pairs. Reported accuracy was 92%, hallucination declined from 15% to 1.45%, response time remained below four seconds, and the two-week operating cost was US\$11.58 using GPT-4o mini. The study shows technical practicability in a consequential, realistic setting and reports operational metrics that many educational papers omit.

Yet the deployment was short and reported as a preprint. It does not establish effects on applicant decisions, access equity, appeal rates, staff workload over an admissions cycle, or institutional liability. This pattern recurs across the literature: authentic settings are appearing, but semester-long, multi-site, replicated, or institution-scale evaluations are rare. Reviews therefore often combine mature and immature evidence in a single unified category, which can make the field appear more deployment-ready than it is.

Table 2
Illustrative empirical studies and the level of inference they support

Study	System/use	Evidence	Outcome level	Maturity	Primary constraint
Guo et al. [17]	Two-agent formative feedback	240 middle-school science responses	Artifact errors	Prototype/evaluation	Preprint; K-12 artifact study; no learner outcome
Lai and Li [18]	Single vs multi-agent feedback	200 essays; 800 outputs	Error/hallucination	Controlled artifact study	Journal and external validity need caution
Hao et al. [19]	Multi-agent learning environment	305 students	Learning, motivation, acceptance	Authentic learner study	Single context; durability unclear
Alqurni [21]	Perceived agentic AI	256 demographic-analysis cases	Motivation pathways	Cross-sectional survey	Source reports inconsistent sample counts; perception is not demonstrated autonomy
Nguyen-Duc et al. [22]	Multi-agent admissions RAG	6,079 Q–A pairs	Accuracy, hallucination, latency, cost	Two-week real workflow	Preprint; no applicant outcome

4.6 Meaningful human oversight is usually described, not tested

Meaningful human oversight is a recurrent recommendation but an uncommon independent variable. Reviews advocate teacher control, human–AI synergy, validation, transparency, and institutional governance [13, 14, 16]. Few studies compare oversight designs, such as output approval versus checkpointed intervention, or measure whether reviewers detect errors under time pressure. Human presence is often assumed to equal effective control, even when the person cannot inspect tool calls, reconstruct the agent’s trajectory, limit permissions, or reverse an action.

This gap matters because oversight has costs and failure modes. Continuous review can erase expected efficiency gains; sparse review can normalize automation bias. A nominal approver may rubber-stamp high-volume outputs. Conversely, well-designed checkpoints can focus attention on consequential transitions rather than every low-risk step. Effective oversight must therefore be specified in operational terms: who is accountable, what they can see, when they can intervene, which permissions are bounded, what is logged, and how actions are escalated or rolled back.

4.7 From performance evidence to ethical justification

Established governance sources support a proportional approach. The NIST Generative AI Profile identifies confabulation, privacy, bias, information integrity, intellectual property, security, value-chain, and human-AI configuration risks and connects them to govern-map-measure-manage practices [23]. UNESCO emphasizes human agency, privacy, inclusion, and ethical validation in education [24]. Under the European Union AI Act, certain educational uses, including systems involved in access, admissions, evaluation, or assignment, may be classified as high-risk and subject to risk management, data governance, transparency, record-keeping, and human-oversight requirements [25]. These instruments do not prove that an intervention works, but they clarify why performance evidence alone cannot justify deployment. Ethical justification also depends on purpose, proportionality, distribution of risk, effective control, and remedies for affected persons.

An agentic system may be accurate on average and still be ethically unacceptable. It may obscure responsibility, shift verification work onto already burdened staff, intensify surveillance, weaken a learner's control over personal data, or make decisions that cannot be meaningfully contested. Conversely, uncertainty does not require a blanket ban on low-risk experimentation. It requires constrained permissions, explicit ownership, transparent evaluation, and a burden of justification that rises with both autonomy and consequence.

5. The Autonomy–Oversight–Evidence framework

The AOE framework links three dimensions that are usually reported separately. Autonomy describes what the system can decide and execute. Oversight describes the strength and timing of accountable human control. Evidence maturity describes how closely the evaluation approximates durable, scalable educational practice. The framework is descriptive before it is normative: it gives researchers and institutions a common language to describe a study or use case. Its decision rule is normative: allowable autonomy should not outrun either evidence maturity or oversight strength.

The proportionality rule rests on three propositions. First, delegation does not transfer institutional moral responsibility to a model, vendor, or nominal human approver. Second, oversight is meaningful only when the responsible person can understand relevant system behavior, refuse or interrupt action, and obtain correction or rollback. Third, people subject to consequential AI-enabled decisions must retain notice, explanation, review, and remedy. These propositions convert human oversight from a reassuring label into an accountable control relationship.

5.1 Normative basis for proportionality

The rule is precautionary but not prohibitory. It permits low-risk, reversible experimentation under modest evidence when permissions remain narrow and outputs are reviewed. It withholds consequential discretion when evidence is immature, oversight is nominal, or affected persons lack recourse. The institution bears the burden of showing why a proposed level of autonomy is necessary, what evidence supports it, who remains accountable, and how errors will be detected and remedied.

Table 3
Operational scales in the Autonomy–Oversight–Evidence framework

Dimension	Level	Label	Operational meaning
Autonomy	A0	Reactive generation	Produces content in response to a prompt; no execution loop
Autonomy	A1	Bounded orchestration	Follows a fixed sequence of agent roles or checks
Autonomy	A2	Delegated task autonomy	Selects tools or substeps within a bounded task
Autonomy	A3	Workflow autonomy	Maintains state and executes multi-step actions in a real workflow
Autonomy	A4	Consequential autonomy	Influences admissions, grading, progression, allocation, or other rights-bearing decisions
Oversight	O0	Unspecified	No accountable review or intervention arrangement reported
Oversight	O1	Retrospective audit	Outputs or logs are sampled after use
Oversight	O2	Pre-use approval	A responsible person approves outputs before they affect users
Oversight	O3	Checkpointed control	Human approval is required at defined consequential transitions
Oversight	O4	Continuous bounded supervision	Permission limits, live monitoring, logging, rollback, and escalation are integrated
Evidence	M0	Concept	Proposal, framework, or use case without an implemented system
Evidence	M1	Prototype/benchmark	Synthetic, technical, or artifact-level evaluation
Evidence	M2	Participant evaluation	Laboratory or short evaluation with higher-education participants
Evidence	M3	Authentic deployment	Use in an actual course or university workflow
Evidence	M4	Extended or multi-site	Semester-long, multi-course, or multi-institution evaluation
Evidence	M5	Replicated/institution scale	Independent replication or sustained institution-scale deployment

5.2 Applying the proportionality rule

The framework does not require every low-risk classroom tool to reach M5. It requires claims and permissions to be proportionate. An A1 feedback checker at M1 may be appropriate for instructor-supervised experimentation if outputs are approved before students receive them. An A3 advising system connected to live data requires at least evidence of authentic deployment, strong permission boundaries, meaningful checkpoints, and rollback. An A4 system should not make final high-impact decisions merely because its benchmark accuracy is high. Consequential uses require mature, context-relevant validation and continuous accountable control; in many cases, the system should remain advisory rather than decisional.

AOE also clarifies research comparisons. A single-agent versus multi-agent experiment may compare architectures while holding oversight constant. A study comparing O2 and O3 can assess whether checkpoints improve error detection without eliminating efficiency gains. Longitudinal studies can test whether an A2 tutor shifts from productive scaffolding to cognitive offloading over time. Reporting the triplet—for example, A2–O3–M3—makes otherwise vague claims about an ‘agentic deployment’ clearer.

Table 4
Proportionate evidence and oversight targets by autonomy level

Autonomy	Risk posture	Minimum evidence target	Oversight target	Illustrative controls
A0–A1	Low	M1–M2	O1–O2	Disclosure, output checks, data minimization
A2	Moderate	M2–M3	O2–O3	Tool allowlists, checkpoints, logs, stop conditions
A3	High	M3–M4	O3–O4	Sandboxing, least privilege, monitoring, rollback, incident response
A4	Consequential	M4–M5 plus legal validation	O4; human final authority	Impact assessment, appeal, auditability, independent validation

6. Implications

6.1 For research

Research should begin by reporting actual capabilities rather than product labels. A minimum system description should identify model and version, planning mechanism, memory, tools and permissions, orchestration, stopping conditions, logging, human intervention spots, and failure handling. Studies should separate artifact, learner, educator, and institutional outcomes; state the comparator; report null and adverse findings; and disclose latency, cost, and verification labor. Because agent behavior can change with model updates and tool availability, configuration and evaluation dates are part of reproducibility.

The field also needs stronger designs. Priority studies include randomized or well-controlled comparisons against both non-AI and static-AI alternatives; semester-long follow-up; multi-site replication; subgroup analyses for equity; measurement of skill retention and transfer after agent withdrawal; and factorial tests of autonomy and oversight. Quality appraisal and evidence maturity should remain separate. A rigorously conducted short laboratory study can have low maturity but low risk of bias; a large authentic deployment can be mature yet methodologically confounded.

6.2 For educators

Educators should decide which cognitive work the learner must retain before deciding what an agent may perform. Delegation is more defensible when it removes administrative friction or supplies formative checks while leaving goal setting, justification, evaluation, and reflection visible to the learner. Assessment planning should include process evidence, oral explanations, staged drafts, or critiques of agent outputs, where

appropriate. Agent logs can support reflection, but they should not become intrusive surveillance or a substitute for pedagogical judgment.

6.3 For institutions and policy

Institutions should inventory agentic uses by action, affected interest, and consequence, not only by vendor. Procurement should require documentation of data flows, tool permissions, retention, model updates, oversight, incident handling, and exit options. Pilots should specify success and harm criteria, named accountable owners, limited populations, and a termination plan. High-impact educational decisions should preserve human final authority, explanation, contestability, and accessible appeal. Evidence from another institution or discipline should be treated as transferable only after local validation, including subgroup analysis for unequal error and burden.

Governance must also account for hidden labor. If instructors must verify every output, claimed time savings may be displaced rather than realized. If verification is relaxed to preserve efficiency, the risks of errors and accountability violations increase. Institutions should measure both automation gains and oversight burden, including who performs that work and whether it is recognized. The AOE triplet can be added to project registers and ethics or procurement reviews so that rising autonomy triggers stronger evidence and control requirements.

7. Limitations

This article has five principal limitations. First, it is a critical integrative review of a purposefully assembled secondary-literature source set, not a completed, prospectively registered systematic review of primary studies. It cannot estimate the prevalence of particular systems or pool educational effects. Second, the primary studies are illustrative and were selected to test recurring claims, not to constitute an exhaustive sample. Third, terminology and system capabilities change rapidly; a product name or model family does not guarantee stable conduct across versions. Fourth, several recent sources are preprints or advance publications, and the evidence base is concentrated in brief, single-context evaluations. Fifth, source selection and coding were conducted by one author, so inter-rater reliability was not estimated. The AOE framework is therefore a proposal for validation, not a validated risk score. Subsequent research should test inter-rater reliability, predictive usefulness, and feasibility across institutional contexts.

8. Conclusion

Agentic AI in higher education is moving from interface to infrastructure, but the evidence has not matured at the same speed. The current literature supports a restrained conclusion: structured and multi-agent designs can improve selected artifacts, and some authentic deployments and learner studies report benefits. It does not yet justify broad claims of sustained learning effectiveness, equitable impact, workload reduction, or safe institutional autonomy. The central governance problem is a mismatch among what systems may do, how people can control them, and how mature the supporting evidence is.

The AOE framework answers this mismatch by reporting autonomy, oversight, and evidence maturity together. Its ethical claim is stricter than a call for responsible innovation: delegated discretion is justified

only when accountable control and context-relevant evidence are jointly sufficient for the stakes involved. Where either condition fails, the system's permissions should remain narrow, reversible, and advisory. Higher education need not reject agentic systems by default, but it should require them to earn discretion and preserve remedy for the people subject to their actions.

Declarations

Competing interests:

The author has no relevant financial or non-financial interests to disclose.

Ethics approval:

Ethical approval was not required because the study analyzed published literature and public bibliographic records.

Use of generative AI

Generative AI tools assisted with literature organization, drafting, and language editing. The author verified the sources, coding, interpretation, argument, and final text and accepts full responsibility for the manuscript.

Funding:

No funding was received for conducting this study.

Author Contribution

A.D. conceived and designed the study, conducted the literature search, screening, data extraction, and synthesis, developed the normative governance framework, and wrote and revised the manuscript. A.D. approved the final manuscript and accepts full accountability for the work.

Data Availability

The data analyzed during this study were extracted from published literature and publicly available bibliographic records. The review-level coding and evidence map supporting the findings are included in the article and Appendix A. The underlying publications are identified through the numbered reference list. No confidential or participant-level data were generated.

References

1. Kamalov F, Sulieman H. From rule-based to goal-oriented: A systematic review of agentic AI in education. IEEE Access. 2026;14:95314–27. <https://doi.org/10.1109/ACCESS.2026.3704764>.

2. Kostopoulos G, Gkamas V, Rigou M, Kotsiantis S. Agentic AI in education: State of the art and future directions. *IEEE Access*. 2025;13:177467–91. <https://doi.org/10.1109/ACCESS.2025.3620473>.
3. Liu X, Tang Q, Wu W, Chen X, Wang C. (2026). GenAI agents in education: A systematic review of macro-level trends, educational roles, conceptual foundations, and future agendas. *Interactive Learning Environments*. Advance online publication. <https://doi.org/10.1080/10494820.2026.2672562>
4. Wang N, Zou D, Xie H, Qin SJ. A scoping review of generative AI-powered agentic AI in education: Research landscape, agentic capabilities, together with insights from the frontier agent paradigm, exemplified by OpenClaw. *Computers Education: Artif Intell*. 2026;11:100653. <https://doi.org/10.1016/j.caeai.2026.100653>.
5. Baradziej S. A systematic review of the roles of agentic AI in higher education. *Social Sci Humanit Open*. 2026;14:103418. <https://doi.org/10.1016/j.ssaho.2026.103418>.
6. Saleem I. (2026). Agentic artificial intelligence in business higher education—a bibliometric analysis to emphasize current understandings and future trends. *Higher Education, Skills and Work-Based Learning*, 16(1), 219–236. <https://doi.org/10.1108/HESWBL-08-2025-0330>
7. Li S, Zheng J. (2026). A scoping review of large language model-based pedagogical agents. *arXiv*. <https://doi.org/10.48550/arXiv.2604.12253>.
8. Lan M, Zhou X. A qualitative systematic review on AI-empowered self-regulated learning in higher education. *npj Sci Learn*. 2025;10:21. <https://doi.org/10.1038/s41539-025-00319-0>.
9. Ba S, Shi X, Wu S, Lu G. Artificial intelligence agents in computer-supported collaborative learning: A systematic literature review. *Computers Education: Artif Intell*. 2026;10:100579. <https://doi.org/10.1016/j.caeai.2026.100579>.
10. Segura Altamirano SF, Maquen-Niño GLE, Guzmán Roldán CM, Pérez Herrera A, Castro Cárdenas DM. Generative artificial intelligence in higher education: A systematic review of perceptions, implementation, and pedagogical transformation. *Rev Educ*. 2026;14(1):e70152. <https://doi.org/10.1002/rev3.70152>.
11. Shi Y, Yu K, Dong Y, Chen F. (2026). Large language models in education: A systematic review of empirical applications, benefits, and challenges. *Computers and Education: Artificial Intelligence*, 10, 100529. <https://doi.org/10.1016/j.caeai.2025.100529>
12. Tillmanns T, Salomão Filho A, Rudra S, Weber P, Dawitz J, Wiersma E, Dudenaite D, Reynolds S. Mapping tomorrow's teaching and learning spaces: A systematic review on GenAI in higher education. *Trends High Educ*. 2025;4(1):2. <https://doi.org/10.3390/higheredu4010002>.
13. Doszhanov B, Zhumadilla B, Zhapparbergenov R. Optimization of university instructors' professional practice through AI agents: A systematic review. *Front Educ*. 2026;11:1893124. <https://doi.org/10.3389/educ.2026.1893124>.
14. Elsayed AA, Malkawi G, Wardat Y, Kamalov F. Beyond automation: A systematic review of AI instructional methods and a framework for human–AI synergy in higher education. *IEEE Access*. 2026;14:45948–63. <https://doi.org/10.1109/ACCESS.2026.3675937>.
15. Wang G, Wang W, Yang D, Ren J. Generative AI, cognitive offloading, and learner agency in higher education: A scoping review. *Behav Sci*. 2026;16(7):1150. <https://doi.org/10.3390/bs16071150>.

16. Bacalso FD, Villamor FE, Paclipan CP, Orquia CJB, Rendon IP, Albiso AT, Himang CM. Institutional approaches to generative AI management in higher education: A systematic review. *Front Educ.* 2026;11:1814426. <https://doi.org/10.3389/feduc.2026.1814426>.
17. Guo S, Latif E, Zhou Y, Huang X, Zhai X. (2024). Using generative AI and multi-agents to provide automatic feedback. *arXiv*. <https://doi.org/10.48550/arXiv.2411.07407>
18. Lai J, Li Z. A comparative empirical evaluation of single-agent and multi-agent LLM prompting strategies for automated formative feedback in education. *J Intell Eng Technol.* 2026;1(2):29–38. <https://doi.org/10.70393/6a696574.343135>.
19. Hao Z, Cao J, Li R, Yu J, Liu Z, Zhang Y. Mapping student–AI interaction relations within multi-agent learning environments: Supporting bespoke learning and reducing performance gaps. *Comput Educ.* 2026;241:105472. <https://doi.org/10.1016/j.compedu.2025.105472>.
20. Wang H, Wang C, Chen Z, Liu F, Bao C, Xu X. Impact of AI-agent-supported collaborative learning on the learning outcomes of university programming courses. *Educ Inform Technol.* 2025;30(12):17717–49. <https://doi.org/10.1007/s10639-025-13487-8>.
21. Alqurni J. Exploring the role of agentic AI in encouraging self-efficacy, autonomy support, and self-learning motivation in higher education. *Front Artif Intell.* 2026;9:1738774. <https://doi.org/10.3389/frai.2026.1738774>.
22. Nguyen-Duc A, Manh CV, Tran BA, Ngo VP, Le Chi L, Nguyen AQ. (2025). An empirical study of multi-agent RAG for actual university admissions counseling. *arXiv*. <https://doi.org/10.48550/arXiv.2507.11272>
23. National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.600-1>
24. Miao F, Holmes W. (2023). Guidance for generative AI in education and research. UNESCO. <https://doi.org/10.54675/EWZM9535>
25. European Union. (2024). Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). Official J Eur Union, L 2024/1689. <http://data.europa.eu/eli/reg/2024/1689/oj>
26. Istrate O. AI agents in education: An early systematic review of emerging roles, potential, and limitations. *Revista de Pedagogía Digit.* 2024;3(1):24–30. <https://doi.org/10.61071/RPD.2496>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixA.docx](#)