

The Intellectual Structure of Research on Agentic AI Ethics and Governance in Higher Education: A Systematic Evidence Synthesis and Research Agenda

Jie Ke*, Jun Sun, Ying Wang

Jackson State University, United States¹

University of Austin, Rio Grande, United States²

*Corresponding author email: jie.ke@jsums.edu

Abstract:

Research on the ethics and governance of artificial intelligence in higher education has expanded rapidly since 2022, but the field's intellectual structure, its dominant themes, internal debates, and blind spots have not been systematically mapped since AI tools began exhibiting agentic properties such as multistep planning and autonomous task execution. This evidence synthesis maps the field using a PRISMA guided systematic search and bibliometric analysis of 312 publications (2019 to 2026) drawn from Web of Science and Scopus. Keyword co-occurrence and co-citation analyses identify five thematic clusters: transparency and explainability; distributed accountability and liability; human oversight and controllability; value alignment and epistemic integrity; and equity, access, and algorithmic fairness. For each cluster, the synthesis characterizes the constituent literature, its convergent findings, and its unresolved disagreements. Across clusters, six structural gaps emerge: a deficit of empirical implementation studies relative to conceptual work; the absence of validated measurement instruments for responsible AI use; the underrepresentation of community colleges, minority-serving institutions, and Global South contexts; an imbalance that favors accountability directed at students over accountability directed at institutions and developers; uncertain transfer of findings from before 2023 to agentic systems; and a weak connection between education research and technical AI governance venues. The paper concludes with a research agenda addressing each gap in turn, intended to guide researchers, editors, and funders as agentic AI deployment in higher education accelerates.

Received:

26 August 2026

First Decision:

9 September 2026

Revised:

17 September 2026

Accepted:

24 September 2026

Published:

5 October 2026

Copyright © 2026

by Dr. Jie Ke, Dr. Jun Sun, Dr. Ying Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution license (CC BY NC).



Keywords: bibliometric analysis, academic integrity, institutional policy, accountability, equity

1. Introduction:

The arrival of generative AI in higher education produced one of the fastest growing literatures in educational research. Within three years of ChatGPT's release, scholarship on the ethics and governance of AI in universities had moved from a specialist concern to a mainstream preoccupation of journals across educational technology, higher education studies, and applied ethics. Rapid growth of this kind carries a familiar cost: literature accumulates faster than anyone can read it, duplicated effort proliferates, and the questions that matter most are not necessarily the questions being studied. Periodic systematic mapping is the discipline's corrective. Existing syntheses, however, either predate the field's turn toward agentic AI, systems that plan, execute, and adapt multistep tasks with limited human oversight, or address single slices of the problem, such as academic integrity (Bittle & El-Gayar, 2025) or AI ethics in education generally (Yu & Yu, 2023). No published synthesis has mapped the full ethics and governance literature as it has reorganized itself around agentic capabilities.

This paper provides that map with three questions. First, what is the intellectual structure of research on the ethics and governance of AI in higher education between 2019 and 2026: which themes organize the field, which works anchor them, and how has the structure shifted over time? Second, within each thematic cluster, what has the literature established, and where does it disagree? Third, what does the field systematically fail to study, and what research agenda follows from those gaps?

The corpus analyzed here was originally assembled for a companion study (Author, 2026, manuscript submitted for publication) that used bibliometric clustering instrumentally, as the empirical foundation for a governance framework. The present paper has a different purpose and makes a different contribution: the synthesis itself. Where the companion paper asks how institutions should govern agentic AI, this paper asks what the research field knows, how that knowledge is organized, and where it is thin. What this paper offers instead is the field level evidence assessment that framework building necessarily compresses: a cluster by cluster reading of the literature's composition and quality, an explicit gaps analysis, and a prioritized agenda for the next phase of research.

1.1. Purpose and Significance of the Study

This study provides the first systematic, PRISMA guided map of ethics and governance literature since AI tools in higher education began exhibiting agentic properties. Its purpose is threefold: to identify the thematic clusters, anchor works, and points of agreement and disagreement that organize the field between 2019 and 2026; to expose the structural gaps, in measurement, implementation evidence, institutional representation, and accountability balance, that the field has not yet addressed; and to translate those gaps into a prioritized research agenda.

The significance of this synthesis extends beyond a literature review. As agentic AI systems assume greater autonomy over academic tasks, institutions, regulators, and funders are already making governance decisions on the strength of literature this paper shows to be conceptually mature but empirically thin. By identifying exactly where the evidence base is strongest and where it is weakest, this study equips researchers to direct their next studies productively, gives editors and reviewers a basis for raising the field's evidentiary standards, and gives funders and institutional leaders a concrete, prioritized target for investment.

2. Background: The Agentic Turn in Educational AI

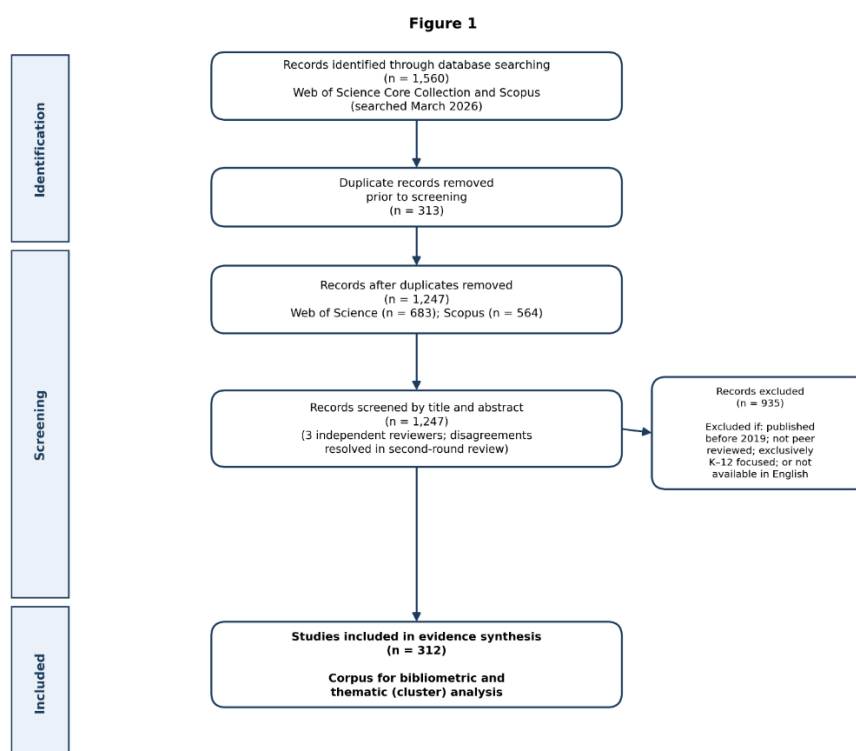
Two developments make this systematic literature review timely. The first is empirical. Adoption of generative AI in higher education outran both policy and research almost immediately: by late 2023, nearly half of American high school students reported using generative AI for schoolwork (ACT, 2023), while many educators reported working without clear institutional AI policies, and a quarter of higher education institutions surveyed globally had already encountered ethical incidents related to AI (UNESCO, 2025). Evidence on learning effects accumulated just as quickly and just as unevenly: meta-analyses report a moderate-to-large positive effect of ChatGPT use on achievement (Liu et al., 2025) and a large positive effect on undergraduate learning outcomes (Mo et al., 2026), yet primary studies caution that effects reverse when AI substitutes for rather than scaffolds student reasoning (Imran & Almusharraf, 2023), and research on student voice reveals simultaneous enthusiasm and anxiety about dependency and fairness (Baidoo-Anu & Owusu Ansah, 2023; Chan & Hu, 2023).

The second development is conceptual. The ethics literature's object of study changed underneath it. Early work addressed reactive systems, tools that answer prompts, inheriting its categories from the AI literacy tradition (Long & Magerko, 2020; Ng et al., 2021). Beginning around 2024, a new theoretical vocabulary entered the field through work on advanced AI assistants (Gabriel et al., 2024) and the dimensional characterization of AI agents by autonomy, efficacy, goal complexity, and generality (Kasirzadeh & Gabriel, 2025). Systems now common in universities occupy the lower reaches of this agentic spectrum: they decompose assignments, carry context across sessions, and act on a student's behalf across tasks. Whether the findings and frameworks built for reactive tools survive this transition is, as the synthesis below shows, one of the field's most consequential open questions. Institutions, meanwhile, are governing in the interim: policy inventories find enormous variation in scope and enforceability across universities (Azevedo et al., 2025), and global survey data indicate that only about a fifth of institutions have formal AI policies while twice that share report frameworks still under construction (UNESCO, 2025). The literature mapped below is therefore not an academic exercise; it is the evidence base on which a live, worldwide effort to build policy is drawing.

3. Methodology:

3.1. Search Strategy and Screening

Records were retrieved from the Web of Science Core Collection and Scopus in March 2026 using a Boolean string combining three concept groups: AI ("agentic AI" OR "generative AI" OR "large language model*" OR "LLM"), setting ("higher education" OR "universit*" OR "academic*"), and governance ("ethic*" OR "governance" OR "accountabilit*" OR "responsibilit*"), with truncation wildcards to capture variant word forms. The search using key words and combination of keywords was conducted in title, abstract, and text of the literature. Eligible records were peer-reviewed journal articles, conference papers, editorials, books and book chapters. other non-peer-reviewed materials were excluded. Before deduplication, 1,560 records were retrieved; 313 duplicates were identified and removed by Doi matching and content scanning. The search returned 1,247 unique records after deduplication (WoS: 683; Scopus: 564). Screening followed PRISMA 2020 procedures: records were excluded if published before 2019, not peer reviewed, exclusively K–12 focused, or unavailable in English, yielding a final corpus of 312 articles/literature. Title and abstract screening were conducted by 3 researchers, with disagreements resolved by the second-round of reviews and discussions. The screening process is documented in the PRISMA 2020 flow diagram (Figure 1).



Note. Flow diagram adapted from the PRISMA 2020 statement (Page et al., 2021). Screening applied the four exclusion criteria reported in the text (publication date, peer-review status, educational level, and language); the source manuscript does not report a per-criterion breakdown of the 935 records excluded at this stage.

PRISMA 2020 Flow Diagram of Study Identification, Screening, and Inclusion

3.2. Bibliometric and Thematic Analysis

VOSviewer (version 1.6.20) was used for keyword co-occurrence analysis (minimum occurrence threshold: five) and co-citation analysis with cited references as the unit of analysis (minimum co-citation threshold: three). Cluster solutions were generated with VOSviewer's default association strength normalization. The quantitative clusters were then subjected to qualitative thematic synthesis: the top 5 most-cited and top 5 most-co-cited works in each cluster were read in full and coded for research design, population, findings, and claims structure.

The corpus analyzed here is the same corpus of 312 articles/literature assembled for a companion framework study (Author, 2026, manuscript submitted for publication). That study reports the cluster solution summarily as the empirical basis for a governance matrix; the present paper reports the full synthesis, including per cluster evidence assessment and gap analysis, which the companion paper does not contain. This reuse of the same corpus of 312 articles/literature is disclosed in the submission cover letter.

4. Results:

4.1. Corpus Structure and Temporal Evolution

The keyword co-occurrence analysis identified 89 keywords meeting the threshold, organized into six clusters. The highest total link strength (TLS) values belonged to "artificial intelligence" (TLS = 847), "academic integrity" (TLS = 612), "higher education" (TLS = 589), "AI literacy" (TLS = 534), and "accountability" (TLS = 498). That two of the five most central terms concern responsibility and trust, rather than technology or pedagogy, is itself a finding: the field's center of gravity is normative. The six clusters and their most central keywords are summarized in Table 1.

Table 1.

Six Keyword Co-occurrence Clusters (89 keywords meeting the threshold)

Keyword cluster	Keywords (n)	Most central keywords (by total link strength)
AI technologies and tools	13	artificial intelligence (847), generative AI, large language models, ChatGPT, agentic AI
Transparency, explainability, and AI literacy	20	higher education (589), AI literacy (534), transparency, explainability, disclosure
Academic integrity and assessment	17	academic integrity (612), plagiarism detection, academic dishonesty, assessment design, AI detection

Accountability and institutional governance	14	accountability (498), institutional policy, AI governance, responsibility, regulation
Human oversight and value alignment	16	human oversight, value alignment, epistemic integrity, autonomy, human-in-the-loop
Equity, access, and algorithmic fairness	9	equity, algorithmic fairness, digital divide, minority-serving institutions, access

Co-citation analysis (cited references as the unit of analysis) produced five clusters, summarized in Table 2: transparency and explainability ($n = 89$); distributed accountability and liability ($n = 74$); human oversight and controllability ($n = 61$); value alignment and epistemic integrity ($n = 58$); and equity, access, and algorithmic fairness ($n = 30$). Publication dates within clusters reveal a clear temporal structure. Work published 2019 to 2022 concentrates on predictive and reactive systems, plagiarism detection, learning analytics ethics, and AI literacy definition; from 2023 onward the corpus pivots sharply, with the steepest growth in papers organized around agentic properties.

The equity cluster is the smallest but fastest growing, with the steepest year over year citation trajectory in the corpus. The keyword and co-citation solutions are related but not identical: the six keyword clusters partition the field by topic vocabulary, while the five co-citation clusters partition it by shared intellectual ancestry. Their alignment on integrity, accountability, and literacy themes, and their divergence at the margins, where keyword clusters separate tool specific discussion that co-citation analysis merges, indicates a field whose terminology is still ahead of its theoretical consolidation.

Table 2.

Five Co-Citation Clusters in the Corpus ($N = 312$)

Cluster	n	Anchor Works (Most Co-cited)	Dominant Concern	Principal Gap
Transparency and explainability	89	Gabriel et al. (2024); Kasirzadeh & Gabriel (2025); Yu & Yu (2023)	Interrogability of AI reasoning in academic work	Principles rarely operationalized
Distributed accountability and liability	74	Mukherjee & Chang (2025); Kapoor et al. (2025); Chan (2023)	Who answers when agentic AI causes academic harm	Almost no empirical accountability studies

Human oversight and controllability	61	Gabriel et al. (2024); Kasirzadeh & Gabriel (2025)	Designed (versus assumed) human control	Student rights largely untheorized
Value alignment and epistemic integrity	58	Durand-Folco & Martineau (2023); Bean (2011); Furze (2024)	Epistemic development under AI assistance	Long term learning outcomes unmeasured
Equity, access, and algorithmic fairness	30	UNESCO (2021); Arum et al. (2025); IDRC (2024)	Differential access, support, and bias exposure	Smallest cluster; MSIs and Global South missing

Note. n = documents per co-citation cluster. MSIs = minority-serving institutions; IDRC = Inclusive Design Research Centre.

4.2. Cluster 1: Transparency and Explainability

The largest cluster is anchored by the theoretical program of Gabriel et al. (2024) and Kasirzadeh and Gabriel (2025) and, on the education side, by Yu and Yu's (2023) earlier bibliometric mapping of AI ethics in education. Its organizing problem is interrogability: whether the humans who must evaluate AI assisted academic work, instructors grading it and students defending it, can access the reasoning that produced it. The literature converges on three points. It agrees that visibility of agent behavior and interpretability of agent reasoning are distinct requirements, frequently conflated in institutional policy. It agrees that the burden of explanation should rise with the system's autonomy, since more autonomous systems perform more of the cognitive work whose provenance matters. And it agrees that in practice neither requirement is met: universities license and permit tools whose reasoning neither party to the assessment relationship can inspect (Hibbert et al., 2024).

Where the cluster is weakest is in operationalization and evidence. Most contributions state transparency as a principle rather than a testable requirement, a pattern Yu and Yu (2023) already observed and which persists in the literature that followed the agentic turn. Empirical work is dominated by AI literacy studies, competency definitions descending from Long and Magerko (2020) and Ng et al. (2021), assessment frameworks (Chee et al., 2025), and gray literature institutional guides (Digital Education Council, 2025; EDUCAUSE, 2024; European Commission & OECD, 2025). These document what students and staff should understand about AI systems, but almost no studies test whether transparency mechanisms, disclosure logs, provenance records, reasoning traces, actually change instructor judgment or student learning. Methodologically, the cluster is dominated by conceptual analysis, framework proposals, and survey research; experimental and design-based studies

are rare, and the few competency instruments that exist (Chee et al., 2025) have been validated for reactive tools rather than agents. The cluster thus exhibits a characteristic shape: theoretically sophisticated, prescriptively rich, empirically thin.

4.3. Cluster 2: Distributed Accountability and Liability

The second cluster joins two literatures that citation patterns show becoming one: a philosophical literature on responsibility for autonomous systems and an applied literature on academic integrity. From the first side, Mukherjee and Chang (2025) supply the cluster's most borrowed concept, the diffusion of responsibility across developers, deployers, and users when autonomous systems cause harm, and Kapoor et al. (2025) supply its most cited design argument, that agentic systems should be built to serve user rather than platform interests. From the second side, the empirical record is comparatively strong: systematic reviews document both the efficiency benefits and the dishonesty risks of generative AI (Bittle & El-Gayar, 2025), studies trace integrity failures to assessment design rather than student character (Cotton et al., 2023), and demonstrations of detector unreliability, including false positives on canonical historical texts, have effectively closed the debate on detection as a governance strategy (Edwards, 2023).

The cluster's live disagreement concerns where integrity governance should reside. One position treats integrity as a rules problem: clearer syllabus taxonomies of permitted, restricted, and prohibited AI use (Chan, 2023) and graduated scales of AI involvement (Furze, 2024). A second treats it as a development problem: what matters is not compliance but whether students' epistemic agency grows or atrophies under AI assistance (Ardito, 2025; Durand-Folco & Martineau, 2023). Citation practice suggests the field is migrating from the first position to the second, consistent with professional bodies' shift from prohibition to disclosure norms (American Psychological Association, 2023). What neither position yet possesses is evidence: the cluster contains almost no studies that follow an accountability arrangement through an actual integrity incident, and none that test whether taxonomies, disclosure memos, or attestation regimes change behavior. The cluster is also conspicuously silent on liability in the legal sense. Although its vocabulary is juridical, responsibility, liability, accountability, contributions from legal scholarship are nearly absent from the co-citation network, so questions that will eventually be settled in contracts and courts, such as a vendor's obligations when its agent fabricates sources in a graded submission, are being theorized without the discipline best equipped to answer them.

4.4. Cluster 3: Human Oversight and Controllability

The third cluster is the most theoretically consolidated and the most empirically deserted. Its consensus claim, articulated by Gabriel et al. (2024) and formalized by Kasirzadeh and Gabriel (2025), is that

human control over AI systems is a designed property, not a default one, and that governance should specify autonomy thresholds above which human in the loop review is mandatory. Education scholarship in this cluster imports the distinction between nominal control (a human formally initiates and approves) and substantive control (a human can competently interrupt, redirect, and reject), and observes that student initiated agentic work often satisfies only the former.

Three features of this cluster stand out in the synthesis. First, nearly all of its work addressing education is conceptual or policy analytic; oversight mechanisms such as review checkpoints, permission scoping, and override rights are proposed frequently and evaluated never. Second, the oversight relation studied runs almost exclusively from institution to student; the reciprocal relation, students' rights against AI systems imposed on them, in grading, feedback, and advising, appears only at the margins, chiefly via Kapoor et al.'s (2025) agent advocate argument, and remains untheorized in terms specific to education. Third, the cluster has begun to attract policy attention faster than research attention, with governance proposals surfacing in legislative and regulatory venues (Executive Order No. 14277, 2025; Federation of American Scientists, 2026) that cite the same thin conceptual base. The risk this creates, policy consolidating around unevaluated mechanisms, is discussed in Section 5.

4.5. Cluster 4: Value Alignment and Epistemic Integrity

The fourth cluster is the field's pedagogical conscience. Its anchor works are older than the others', because its central concept predates AI: Bean's (2011) argument that intellectual development requires productive struggle, imported into the AI debate to ask what happens when an agent absorbs the struggle on the student's behalf. Durand-Folco and Martineau (2023) give the cluster its political economy, warning of credentials accumulating without the capacities they certify, and the practitioner wing supplies its instruments: staged models of institutional adaptation (Maloney, 2023), graduated AI involvement scales (Furze, 2024), and ethics integrated course design frameworks (Online Learning Consortium, 2025).

The cluster's convergent finding, consistent across otherwise disparate work, is that assessment design, not policy language, is the operative variable: tasks eliciting generic responses invite substitution, while tasks demanding localized knowledge, iterative process evidence, or oral defense make AI use visible and pedagogically tractable (Cotton et al., 2023). Faculty development research reaches the parallel conclusion that practice-based programs outperform policy dissemination (Alier et al., 2024; Monzon & Hays, 2025; UNESCO, 2024), and institutional case studies confirm that policy without capacity produces confusion rather than compliance (Azevedo et al., 2025; Wu et al., 2024). The gap is longitudinal: no study in the corpus measures whether redesigned assessment actually protects learning outcomes over semesters or degrees, so the cluster's central causal claim, that

alignment oriented design preserves epistemic development where prohibition does not, rests on theory and short-term indicators rather than direct evidence. The mechanisms the cluster proposes are nonetheless the field's most concrete: process portfolios, annotated revision of AI output, reflective disclosure memos, and oral defense all have clear operational definitions and could be evaluated with standard educational research designs, which makes this cluster the natural starting point for the implementation studies the field lacks.

4.6. Cluster 5: Equity, Access, and Algorithmic Fairness

The smallest cluster carries the largest normative weight and the steepest growth curve. Its empirical anchor is Arum et al.'s (2025) documentation of stratified adoption: students at less selective institutions receive less institutional support for AI use even as they face equal or greater exposure to its risks. Its normative anchor is UNESCO's (2021) recommendation on AI ethics, and its design anchor is the Inclusive Design Research Centre's (2024) triad of access, support, and bias audit. The cluster's distinctive analytical move is to treat AI harms as compounding rather than additive: access gaps produce competency gaps, which amplify exposure to algorithmic bias in the AI assisted grading and adaptive systems most common at under resourced institutions.

Two silences define this cluster more than its contents. Community colleges, historically Black colleges and universities, Hispanic-serving institutions, and tribal colleges, precisely the institutions where the cluster's compounding logic predicts the greatest harm, are nearly absent as research sites rather than as objects of concern. And the Global South appears almost exclusively through intergovernmental gray literature rather than empirical study. A literature whose fastest growing theme is inequity, studied overwhelmingly at well-resourced institutions in the Global North, is a literature measuring the problem where it is smallest.

5. Gaps and Research Agenda

Read across rather than within clusters, the synthesis yields six structural gaps. The gaps are not independent: the measurement gap blocks the implementation studies that would close the evidence deficit, the population blind spots guarantee that whatever evidence does accumulate will generalize poorly, and disciplinary insulation keeps the expertise needed for accountability questions outside the field. Closing them in isolation is therefore unlikely to work; the agenda below is ordered so that measurement and implementation work enable the rest. Table 3 pairs each with the research designs most likely to close it.

First, an implementation deficit: conceptual, prescriptive, and policy analytic work outnumbers empirical implementation study in every cluster; proposed mechanisms (disclosure memos, oversight checkpoints, bias audits) circulate widely without ever being evaluated in a live course or institution.

Second, a measurement gap: the field possesses no validated instruments for its own core constructs, responsible use, epistemic integrity under AI assistance, substantive oversight capacity, leaving studies to rely on self-reported measures and ad hoc rubrics.

Third, population blind spots: minority-serving institutions, community colleges, and Global South universities are objects of equity argument but rarely sites of data collection.

Fourth, an accountability asymmetry: student obligations are elaborately specified across the corpus while institutional and developer obligations remain gestural, an inversion of where power over agentic systems lies.

Fifth, temporal fragility: a large share of the corpus's empirical base predates agentic capabilities, and no study tests whether findings from research on reactive tools transfer to systems that plan and act; the field may be governing tomorrow's systems with evidence about yesterday's.

Sixth, disciplinary insulation: co-citation patterns show education research and technical AI governance scholarship citing largely disjoint literatures, meeting only through a handful of boundary works (Gabriel et al., 2024; Kasirzadeh & Gabriel, 2025), which slows the translation of technical governance advances into educational practice and of educational evidence into system design.

Table 3.

Gap by Gap Research Agenda

Gap	Exemplar research questions	Priority designs
Implementation deficit	Do disclosure memos or oversight checkpoints change instructor and student behavior?	Field experiments; design-based implementation research in live courses
Measurement gap	What observable indicators validly distinguish responsible from substitutive AI use?	Instrument development and validation; measurement invariance studies
Population blind spots	How do agentic AI governance needs differ at MSIs, community colleges, and Global South institutions?	Multi-site comparative case studies; partnerships with under resourced institutions
Accountability asymmetry	What developer and institutional obligations are enforceable, and how?	Policy analysis with legal scholarship; procurement contract studies

Temporal fragility	Do reactive tool findings replicate with agentic systems?	Systematic replications varying system autonomy
Disciplinary insulation	What educational evidence should inform agent design, and vice versa?	Cross venue collaborations; boundary spanning special issues

Note. MSIs = minority-serving institutions.

6. Discussion:

Three implications follow from the map. For researchers, the synthesis reorders the field's incentives: the marginal value of another conceptual framework or policy commentary is now low in every cluster, while the marginal value of implementation studies, validated instruments, and research at underrepresented institutions is high. The comparison with adjacent syntheses sharpens the point: integrity focused reviews (Bittle & El-Gayar, 2025) and mappings that precede the agentic turn (Yu & Yu, 2023) each found their literatures long on concern and short on intervention evidence, and the present synthesis shows that the agentic turn has widened rather than narrowed that ratio, because theory has moved faster than fieldwork.

For editors and reviewers, the temporal fragility gap suggests a concrete practice: requiring authors to specify the agentic profile of the systems they study, at minimum the autonomy and generality of the tools involved, so that future syntheses can distinguish evidence about reactive tools from evidence about agents. Much of the current corpus is ambiguous on exactly this point, which is why the transfer question cannot presently be answered.

For funders and institutional leaders, the population blind spots and the accountability asymmetry are the actionable findings. Both reflect resource allocation rather than intellectual difficulty: studying implementation at a community college costs more relational investment than theorizing at a research university, and specifying developer obligations requires legal and procurement expertise that education faculties rarely hold. Targeted funding and interdisciplinary teams are the plausible correctives; the research agenda in Table 3 is designed to be fundable in precisely that form.

For practitioners, finally, the map is usable now even where the evidence is thin, provided it is used honestly. Institutions selecting among governance mechanisms should prefer those the literature defines operationally (the assessment design instruments of the value alignment cluster) over those it defines only aspirationally (transparency principles without logging or provenance requirements) and should treat their own implementations as evaluable pilots rather than settled policy, since they will in effect be generating the evidence the research base is missing.

The synthesis also clarifies what should not be concluded. The dominance of normative over empirical work is not evidence that the field is unserious; it is the ordinary shape of a literature responding to a fast technological shock, in which conceptual triage precedes evaluation. The argument here is that the triage phase has done its work. The categories are stable, five clusters with clear anchors and boundaries, and the field's next phase should be evaluative.

7. Policy and Societal Implications

The societal stakes of this synthesis lie in what universities certify. Degrees function as public signals of capability; if agentic AI erodes the link between credential and competence faster than governance research can respond, the cost falls not only on institutions but on employers, professions, and the publics they serve. This synthesis shows that the research base underpinning institutional AI governance is thinnest exactly where societal exposure is greatest: at under resourced institutions serving the most vulnerable student populations, and at the level of developer and institutional obligation where power over agentic systems is concentrated. Policymakers drafting rules for AI in education, a process already underway in federal directives and legislative proposals, are therefore building on evidence about yesterday's tools, gathered at unrepresentative institutions, with student conduct specified in excessive detail and institutional conduct left thin.

The practical implications are threefold: regulatory frameworks should mandate transparency about the agentic profile of educational AI systems so that evidence and rules can be matched to actual capabilities; public research funding should be directed toward implementation and measurement studies at minority-serving and open access institutions; and procurement policy should be recognized as a governance instrument, since contract terms between universities and AI developers are currently an important and immediately actionable mechanism of developer accountability for universities. Without these corrections, AI governance in higher education risks consolidating around plausible but unevaluated mechanisms.

8. Conclusion:

This synthesis mapped 312 peer reviewed publications on the ethics and governance of AI in higher education onto five stable thematic clusters, assessed the evidence within each, and identified six structural gaps that jointly define the field's next research phase. The picture that emerges is of a literature that has completed its conceptual work and not yet begun its empirical work: rich in frameworks, principles, and warnings; poor in implementation evidence, validated measures, and representative research sites. As AI systems in universities become more autonomous and more general, the cost of that imbalance compounds, because governance decisions are being made now,

at every institutional level, on the strength of the conceptual literature alone. The agenda proposed here, evaluative, measured, and deliberately aimed at the institutions and actors the field has overlooked, is offered as the map for that work.

9. Limitations:

This synthesis is limited to English-language, peer-reviewed records indexed in Web of Science and Scopus. Database coverage bias is therefore a constraint: preprints, policy gray literature (except where cited contextually), and scholarship not published in English are excluded—a material constraint given the Global South gap the synthesis itself identifies. Search-term sensitivity is an additional limitation; alternative keyword choices might have retrieved a somewhat different corpus. Citation-indexing lag means the newest scholarship is underrepresented: co-citation analysis privileges older, well-cited works and gives less weight to recent publications, so the 2025–2026 literature is represented conservatively. Bibliographic metadata errors in the source databases may have introduced noise into the corpus. The cluster solutions depend on VOSviewer parameter choices (occurrence and co-citation thresholds); alternative thresholds could yield somewhat different partitions. Cluster labeling involves interpretive judgment, and although anchored in quantitative co-citation structure, alternative labels are defensible. Finally, the corpus was assembled in March 2026; in a field growing this quickly, the map ages fast and should be rerun periodically. The corpus is shared with a companion framework study (Author, 2026, manuscript submitted for publication), as disclosed in Section 3.2.

Ethics Approval and Consent:

Not applicable. This study synthesizes published literature and involved no human or animal participants.

Data Availability:

The search strings, screening log, and the full corpus of 312 records with cluster assignments are available from the corresponding author upon reasonable request.

Code Availability:

Not applicable. Analyses used VOSviewer 1.6.20 with parameter settings reported in Section 3.2; no custom code was written.

Competing Interests:

The author(s) declare no competing interests.

Funding:

This research received no external funding.

Use of Generative AI Tools:

Generative AI tools (Gemini Pro and ChatGPT) were used to assist with copy editing, formatting, and reference verification during the preparation of this manuscript. All searches, screening decisions, coding, analysis, interpretations, and conclusions are the authors' own, and the author verified and take full responsibility for all content.

10. References

- ACT. (2023, December 11). High school students' use and impressions of AI tools. *ACT Research*. <https://www.act.org/content/act/en/research/pdfs/High-School-Students-Use-and-Impressions-of-AI-Tools-Accessible.html>
- Alier, M., García-Peñalvo, F. J., & Camba, J. D. (2024). Generative artificial intelligence in education: From deceptive to disruptive. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5), 5–14. <https://doi.org/10.9781/ijimai.2024.02.011>
- American Psychological Association. (2023). How to cite ChatGPT. *APA Style*. <https://apastyle.apa.org/blog/how-to-cite-chatgpt>
- Ardito, C. G. (2025). Generative AI detection in higher education assessments. *New Directions for Teaching and Learning*, 2025(182), 11–28. <https://doi.org/10.1002/tl.20624>
- Arum, R., Calderon Leon, M., Li, X., & Lopes, J. (2025). ChatGPT early adoption in higher education: Variation in student usage, instructional support, and educational equity. *AERA Open*, 11. <https://doi.org/10.1177/23328584251331956>
- Author. (2026, Under Review). Governing AI agents in higher education: A framework for evaluation, accountability, and responsible integration in the agentic era. *Journal of Accreditation and Quality Assurance in Higher Education*.
- Azevedo, L., Robles, P., Best, E., & Mallinson, D. J. (2025). Institutional policies on artificial intelligence in higher education: Frameworks and best practices for faculty. *New Directions for Adult and Continuing Education*, 2025(188), 70–78. <https://doi.org/10.1002/ace.70013>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62. <https://doi.org/10.61969/jai.1337500>
- Bean, J. C. (2011). *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom* (2nd ed.). Jossey-Bass.

- Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4), 296.
<https://doi.org/10.3390/info16040296>
- Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20(38).
<https://doi.org/10.1186/s41239-023-00408-3>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(43). <https://doi.org/10.1186/s41239-023-00411-8>
- Chee, H., Ahn, S., & Lee, J. (2025). A competency framework for AI literacy: Variations by different learner groups and an implied learning pathway. *British Journal of Educational Technology*, 56(5), 2146-2182. <https://doi.org/10.1111/bjet.13556>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Digital Education Council. (2025). Digital Education Council AI literacy framework. <https://www.digitaleducationcouncil.com/post/digital-education-council-ai-literacy-framework>
- Durand-Folco, J., & Martineau, J. (2023). Le capital algorithmique [The algorithmic capital]. *Écosociété*.
- EDUCAUSE. (2024). AI literacy in teaching and learning: A durable framework for higher education. *EDUCAUSE Review*. <https://er.educause.edu/articles/2024/10/ai-literacy-in-teaching-and-learning-a-durable-framework-for-higher-education>
- Edwards, B. (2023, July 14). Why AI detectors think the U.S. Constitution was written by AI. *Ars Technica*. <https://arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/>
- European Commission & OECD. (2025). AI literacy framework for primary and secondary education (AILit). <https://ailiteracyframework.org/>
- Executive Order No. 14277. (2025, April 28). Advancing artificial intelligence education for American youth. 90 Fed. Reg. 17519. <https://www.federalregister.gov/documents/2025/04/28/2025-07368/advancing-artificial-intelligence-education-for-american-youth>
- Federation of American Scientists. (2026, January). Federal coordination to support responsible AI deployment: FAIR in Education Act proposal. <https://fas.org/publication/fair-in-education-act/>

- Furze, L. (2024, August 28). Updating the AI assessment scale.
<https://leonfurze.com/2024/08/28/updating-the-ai-assessment-scale/>
- Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomšev, N., Ktena, I., Kenton, Z., Rodriguez, M., et al. (2024). *The ethics of advanced AI assistants*. arXiv.
<https://doi.org/10.48550/arXiv.2404.16244>
- Hibbert, M., Altman, E., Shippen, T., & Wright, M. (2024). A framework for AI literacy. *EDUCAUSE Review*. <https://er.educause.edu/articles/2024/6/a-framework-for-ai-literacy>
- Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 15(4), ep464. <https://doi.org/10.30935/cedtech/13605>
- Inclusive Design Research Centre. (2024). *Framework for accessible and equitable artificial intelligence (AI) in education*. OCAD University. <https://idrc.ocadu.ca/>
- Kapoor, S., Kolt, N., & Lazar, S. (2025). Position: Build agent advocates, not platform agents. In *Proceedings of the 42nd International Conference on Machine Learning* (Vol. 267, pp. 81617–81633). PMLR. <https://proceedings.mlr.press/v267/kapoor25a.html>
- Kasirzadeh, A., & Gabriel, I. (2025). Characterizing AI agents for alignment and governance. *arXiv*. <https://doi.org/10.48550/arXiv.2504.21848>
- Liu, Z., Zuo, H., & Lu, Y. (2025). The impact of ChatGPT on students' academic achievement: A meta-analysis. *Journal of Computer Assisted Learning*, 41(4), e70096.
<https://doi.org/10.1111/jcal.70096>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
<https://doi.org/10.1145/3313831.3376727>
- Maloney, E. J. (2023, April 4). The 4 stages of AI. *Inside Higher Ed*.
<https://www.insidehighered.com/blogs/learning-innovation/4-stages-ai>
- Mo, F., Huang, J., Yang, Y., Özen, Z., Maeda, Y., & Olenchak, F. R. (2026). Undergraduate students' learning outcomes with ChatGPT: A meta-analytic study. *Computers and Education: Artificial Intelligence*, 10, 100536. <https://doi.org/10.1016/j.caeai.2025.100536>
- Monzon, N., & Hays, F. A. (2025). Leveraging generative artificial intelligence to improve motivation and retrieval in higher education learners. *JMIR Medical Education*, 11, e59210.
<https://doi.org/10.2196/59210>

- Mukherjee, A., & Chang, H. H. (2025). Agentic AI: Autonomy, accountability, and the algorithmic society. *arXiv*. <https://doi.org/10.48550/arXiv.2502.00289>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Online Learning Consortium. (2025, November). GenAI use and ethics framework: A pedagogical model for responsible AI integration in K–12 and higher education. <https://onlinelearningconsortium.org/olc-insights/2025/11/genai-use-and-ethics-framework/>
- UNESCO. (2021). Recommendation on the ethics of artificial intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- UNESCO. (2024). AI competency framework for teachers. <https://www.unesco.org/en/articles/ai-competency-framework-teachers>
- UNESCO. (2025). Two thirds of higher education institutions have or are developing guidance on AI use: UNESCO survey. <https://www.unesco.org/en/articles/unesco-survey-two-thirds-higher-education-institutions-have-or-are-developing-guidance-ai-use>
- Wu, C., Zhang, H., & Carroll, J. M. (2024). AI governance in higher education: Case studies of guidance at Big Ten universities. *Future Internet*, 16(10), 354. <https://doi.org/10.3390/fi16100354>
- Yu, L., & Yu, Z. (2023). Qualitative and quantitative analyses of artificial intelligence ethics in education using VOSviewer and CitNetExplorer. *Frontiers in Psychology*, 14, 1061778. <https://doi.org/10.3389/fpsyg.2023.1061778>

Copyright © 2026 by Dr. Jie Ke, Dr. Jun Sun, Dr. Ying Wang. This is an Open-Access Article Distributed under the Terms of the Creative Commons Attribution License (CC BY NC 4.0)

Doi: <https://doi.org/10.52132/Ajrsp.e.2026.90.1>