Review Article

Journal of Advances in Medical Education & Professionalism

Regulation and Ethical Frameworks for AI in Health Professions Education: A Critical Narrative Review

YASAR AHMED[1,2*], MD, FRCP; SIMAA KHAYAL[3], BSc, MSc, PhD

[1]*Department of Medical Oncology, St. Vincent's University Hospital, Dublin, Ireland;* [2]*School of Medicine, University College Dublin, Dublin, Ireland;* [3]*Independent Researcher, Dublin, Ireland*

Introduction: Integrating artificial intelligence (AI)—particularly generative AI and large language models (LLMs)—into health professions education offers substantial pedagogical opportunities, while highlighting critical gaps in existing ethical and regulatory frameworks. This critical narrative review examines how contemporary literature conceptualizes and addresses these challenges.

Methods: Publications from 2020–2025 were identified through searching PubMed, Scopus, and ERIC, focusing on AI, medical and health professions education, ethics, governance, and regulation. The initial search yielded 452 records; following two-stage screening (title/abstract and full-text review), thirty-four sources met the pre-defined inclusion criteria. These sources, including empirical studies, systematic reviews, policy documents, and conceptual papers, were included. Reflexive thematic analysis was used to synthesize the findings, while SANRA (Scale for the Assessment of Narrative Review Articles) informed the weighting of sources, which were categorized into high-, moderate-, and low-weight tiers.

Results: Four themes emerged: 1) a regulatory vacuum marked by policy lag and fragmented institutional responses; 2) reinterpretation of autonomy, beneficence, nonmaleficence, and justice in relation to opaque, datadriven systems, 3) generative AI's impact on scholarship, authorship, and academic integrity; and 4) equity, bias, and professional identity formation in AImediated learning environments. Integrating these themes with crosssector frameworks and the NIST AI Risk Management Framework, the review develops a riskbased governance taxonomy for AI in Health Professions Education, distinguishing low, medium and highrisk educational use cases.

Conclusion: This synthesis indicates that current literature favors evolving beyond reactive policies. Consequently, we propose a theoretical shift toward risk-proportionate governance. We suggest that human-in-the-loop pedagogy should be prioritized as a conceptual framework to guide future research and policy development, rather than presenting it as a validated outcome.

Keywords: Ethics, Artificial intelligence, Medical education

**Corresponding author:*
Yasar Ahmed, MD, FRCP
Department of Oncology,
St Vincent's University Hospital,
Dublin, Ireland
Tel: +35-3871022059
Email: drhammor@gmail.com

Please cite this paper as:
Ahmed Y, Khayal S. Regulation and Ethical Frameworks for AI in Health Professions Education: A Critical Narrative Review. J Adv Med Educ Prof. 2026;14(4):307-319. DOI: 10.30476/jamp.2026.110257.2340.

Received: 20 December 2025
Accepted: 2 June 2026

Introduction

The integration of Artificial Intelligence (AI) into Health Professions Education (HPE) represents a paradigm shift. It offers transformative potential for personalized learning, assessment efficiency, and curriculum development. Moving beyond simple digitization, AI, specifically the recent proliferation of Generative AI (GenAI) and Large Language Models (LLMs), challenges the fundamental epistemic model of medicine of the "physician-as-knower". This model holds that clinical expertise is constituted by individually acquired, examinable, and personally defensible knowledge that the practitioner can explain and justify (1). LLMs capable of generating diagnostic outputs without transparent reasoning pathways reposition the locus of epistemic authority from the trained clinician toward the algorithmic system, with profound implications for how competence is defined, assessed, and governed in HPE.

This technological integration has outpaced the development of robust regulatory and ethical frameworks. Rapid emergence of these technologies complicates the landscape. It introduces novel risks regarding academic integrity, cognitive deskilling, and the erosion of humanistic values in medical training (2, 3).

AI offers opportunities to revolutionize medicine by improving patient outcomes and increasing efficiency. Its application in education necessitates a distinct ethical lens where healthcare, technology, and educational ethics collide (3). Unlike clinical AI, where regulatory oversight is comparatively mature (4), AI in HPE occupies a liminal space between education, healthcare, and data science. This liminal space is a transitional zone of ambiguity where the established norms of clinical supervision do not fully apply, yet the experimental freedom of computer science is constrained by the duty of care. This intersection creates specific dilemmas. For example, should a student trust an AI's diagnosis over their own intuition, potentially leading to automation bias? Or does the use of algorithmic proctoring systems erode the 'culture of trust' essential for professional identity formation? These are not merely technical hurdles but fundamental conflicts between efficiency and educational values. The current literature underscores a critical gap. Clinical AI ethics is a growing field. Specific ethical implications of AI in the educational context where future professionals are formed remain under-regulated, inconsistently addressed (5). The challenge is no longer hypothetical. AI tools are able to pass medical licensing examinations, generating scholarly outputs. This forces a re-evaluation of assessment, academic integrity, and professional standards (5, 6). The regulatory response has been largely reactive and fragmented. Clinical AI is increasingly governed by frameworks like the FDA's software-as-a-medical-device protocols, and educational AI operates in a grey zone. Student data privacy, cognitive sovereignty or the capacity to maintain independent epistemic agency and critical judgment in AI-augmented learning environments (7) and the "hidden curriculum" of algorithmic bias the tacit transmission of discriminatory assumptions embedded in AI training datasets that shapes learner behaviors and clinical heuristics without explicit acknowledgement (8) remain under-regulated (9).

This review examines the latest peer-reviewed literature to synthesize current regulatory challenges, ethical concerns, and proposed governance frameworks for AI in HPE. Authenticated peer-reviewed literature maps the current state of AI regulation. It moves from the tangible realities of current institutional policies to the theoretical applications of bioethics and to the existential considerations of Artificial General Intelligence (AGI). Current regulatory frameworks are insufficient to address the speed of technological advancement. A shift is necessitated from policing academic misconduct to establishing robust, value-based governance that prioritizes human-in-the-loop pedagogy, defined in this review as an educational governance model in which consequential decisions regarding learning progression, competence assessment, and feedback retain qualified human educator oversight and explicit approval, even when AI tools are used to generate or inform those decisions.

Methods

This study adopted a critical narrative review approach. Following Sukhera and Grant and Booth, we selected this methodology not merely to describe the extant literature but to interrogate it through an explicit theoretical lens, specifically the four-principle bioethics framework and cross-sector regulatory frameworks (e.g., EU AI Act, NIST RMF). To apply this lens, we searched PubMed, Scopus, and ERIC databases for articles published between January 1, 2020, and January 1, 2025. The database search was completed in January 2025. Two sources carrying 2026 publication dates were incorporated during the revision stage as head-of-print publications. All other references displaying

a 2026 year denote the date of URL access for web-based resources, consistent with Vancouver citation practice, and not the publication date of those works.

The time frame was chosen to capture the period beginning with the first broadly deployed large-scale AI and generative models in education and healthcare, during which regulatory and ethical debates in HPE accelerated significantly. Earlier AI applications (e.g., rule-based systems and small-scale machine learning) were governed by different technical and policy assumptions, and including them risked diluting the focus on contemporary generative and high-capacity models that dominate current governance discussions.

The search strategy employed combinations of keywords and MeSH terms related to three conceptual pillars:

1. **Technology:** "Artificial Intelligence," "Generative AI," "Large Language Models."
2. **Context:** "Medical Education," "Health Professions Education," "Clinical Training."
3. **Focus:** "Ethics," "Governance," "Regulation," "Policy," "Academic Integrity."

An example of the full Boolean search string as applied in PubMed is as follows: ("Artificial Intelligence" OR "Generative AI" OR "Large Language Models") AND ("Medical Education" OR "Health Professions Education" OR "Clinical Training") AND ("Ethics" OR "Governance" OR "Regulation" OR "Policy" OR "Academic Integrity"). Database-specific search strings with full MeSH term expansions are available from the corresponding author on request. The search was not pre-registered, consistent with narrative review methodology. To ensure the analysis captured both empirical evidence and authoritative guidance, the inclusion criteria prioritized: 1) Empirical studies and systematic reviews specifically addressing AI in HPE contexts; 2) Policy framework documents from major international health and educational bodies, and 3) Scholarly articles proposing specific regulatory or ethical frameworks for AI adoption. We excluded publications that did not explicitly address AI in a medical or educational context, those without a clear discussion of ethical considerations — defined as any explicit treatment of one or more of the following domains: privacy/data governance, algorithmic bias, academic integrity, authorship, professional identity, or equity —, and sources lacking both peer review and identifiable institutional provenance. For operational clarity, peer-reviewed journal articles were included regardless of the study design; official policy and guidance documents from named regulatory or intergovernmental bodies (WHO, EU, GMC, NIST, OECD) were included as authoritative grey literature, consistent with established practice in narrative synthesis. Sources lacking both peer review and named institutional authorship including anonymous web content and informal commentary were excluded. Non-English publications were also excluded.

The initial search yielded 452 records. Following a rigorous screening of the titles and abstracts, and subsequent full-text review, 34 sources were selected for the final synthesis. Following a structured, two-stage selection process (title/abstract screening followed by full-text review), and thirty-four sources meeting the pre-defined inclusion criteria were selected for final synthesis (Figure 1).

Data Analysis

We utilized a reflexive thematic analysis approach (Braun & Clarke)(10). Two authors independently familiarized themselves with the dataset, generating the initial codes (e.g., 'privacy', 'bias', 'authorship'). These codes were collated into candidate themes, which were then reviewed against the policy and empirical datasets to ensure integration. Source quality was assessed using the SANRA scale (Scale for the Assessment of Narrative Review Articles), a validated six-item instrument evaluating justification of the review's importance, statement of concrete aims, description of the literature search, referencing, scientific reasoning, and clarity of presentation. Each included source was independently scored by both authors on a 0–2 scale per item (maximum score: 12). Sources scoring below 6 indicating deficiencies across multiple quality domains were retained in the corpus, but they were designated as *low-weight*, meaning they were used solely to illustrate the range of existing positions rather than to substantiate empirical claims. No source was excluded on SANRA grounds alone, preserving the breadth of perspectives characteristic of a narrative synthesis. Seventeen sources scored ≥ 9 (high weight), eleven scored 6–8 (moderate weight), and six scored below 6 (low weight/illustrative only). This weighting schema is reflected in the epistemic hierarchy described in the Source Weighting and Integration section below.

Coding was conducted manually using a structured codebook developed collaboratively by the two authors in a shared spreadsheet environment (Microsoft Excel). No dedicated qualitative data analysis software was employed, consistent with the manageable corpus size of 34 sources.

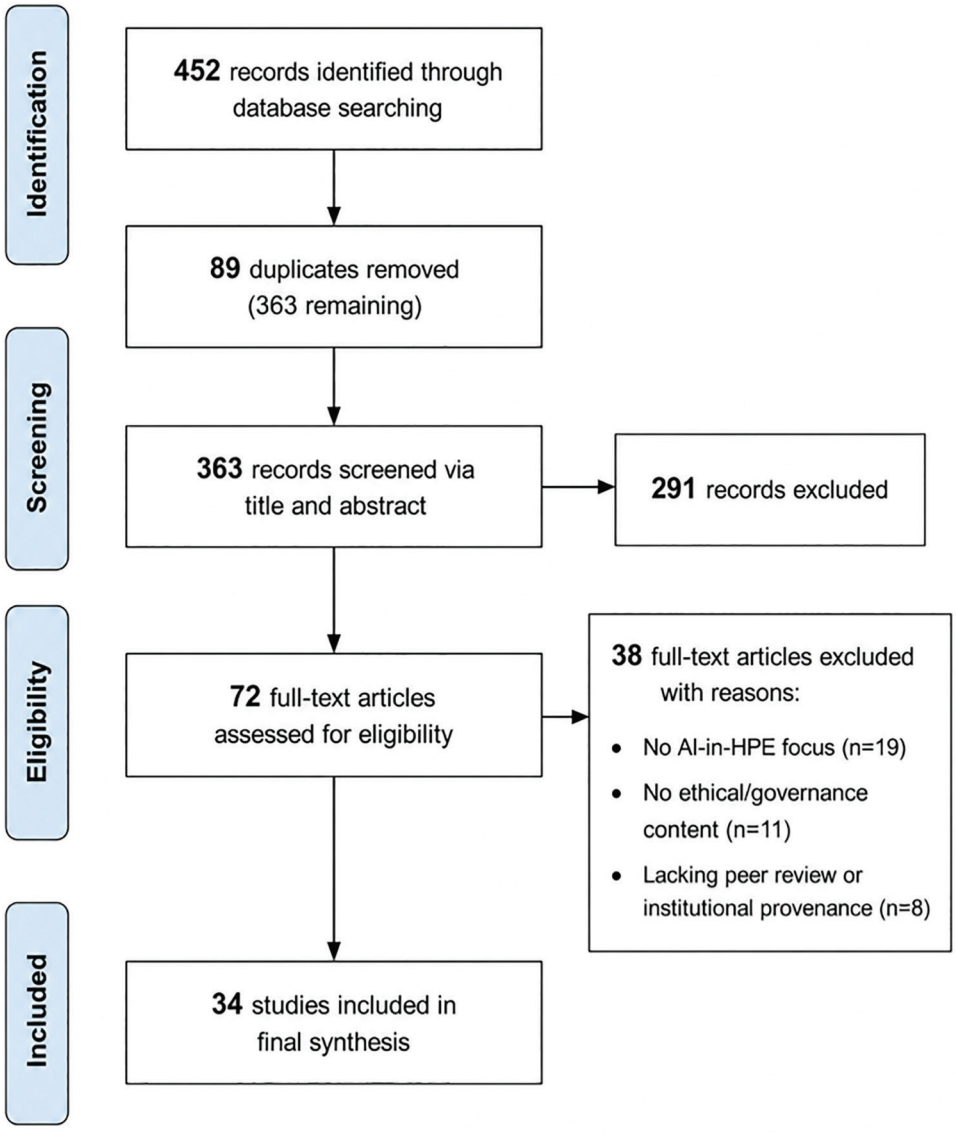

Figure 1. Literature search and selection flow diagram

Inter-coder agreement was established through iterative discussion and consensus rather than a formal reliability coefficient, which is consistent with the reflexive, interpretive orientation of the Braun and Clarke framework, where the goal is theoretical coherence rather than inter-rater reproducibility.

Reflexivity and Limitations

The authors approached this review with the stance that AI integration is inevitable, which may introduce a bias toward 'governance solutions' rather than 'prohibitionist' arguments. Inclusion criteria prioritizing governance-focused literature may have systematically under-sampled studies reporting neutral or successful AI integration outcomes, which represents a potential confirmation bias in the source selection process. Furthermore, the exclusion of non-English sources and the reliance on major databases may omit valuable 'grey literature' or local institutional policies not yet published.

The authorship team's disciplinary background in clinical medicine and health professions education, rather than computer science or regulatory policy, may have amplified sensitivity to pedagogical and ethical risks while affording comparatively less weight to implementation-feasibility arguments advanced in technical literature. Furthermore, the institutional context in which this review was conducted, academic health centers operating within established regulatory environments, may have predisposed the governance framing toward models that presuppose resource infrastructures unavailable in lower-income or resource-constrained settings, a limitation that readers should bear in mind when interpreting the proposed taxonomy.

Table 1. SANRA quality assessment scores and weight assignments for all 34 included sources

#	Author, Year	S1	S2	S3	S4	S5	S6	Total	Weight	Type
1	**Sallam & Sallam, 2025**	**1**	**1**	**0**	**1**	**1**	**1**	**5**	**Low**	NR
2	Busch et al., 2023	2	2	1	2	2	2	11	High	C
3	Patino et al., 2024	2	2	1	2	2	2	11	High	C
4	Masters, 2023 (AMEE 158)	2	2	1	2	1	2	10	High	C
5	Franco D'Souza et al., 2024	2	1	1	2	2	2	10	High	C
6	Tran et al., 2025 (npj governance)	2	2	1	2	2	2	11	High	C
7	Rush et al., 2025	2	2	2	2	2	2	12	High	E
8	Hamilton, 2024	2	2	2	2	2	2	12	High	E
9	Ichikawa et al., 2025	2	2	2	2	2	1	11	High	E
10	Lee et al., 2021	2	2	2	2	2	2	12	High	SR
11	Tran et al., 2025 (npj MKO)	2	1	1	2	2	2	10	High	C
12	Foltynek et al., 2023 (ENAI)	1	1	0	1	1	1	5	Low	P
13	Russell Group, 2023	1	1	0	1	1	1	5	Low	P
14	General Medical Council, 2025	1	1	0	1	1	2	6	Moderate	P
15	Rees et al., 2025	2	2	2	2	2	2	12	High	E
16	TEQSA, 2025	1	1	0	1	1	1	5	Low	P
17	Asiedu et al., 2024	2	2	2	2	2	2	12	High	E
18	Yu & Zhai, 2024	2	2	1	2	2	2	11	High	C
19	Weidener & Fischer, 2024	2	1	1	2	2	2	10	High	C
20	Knopp et al., 2023	2	2	1	2	1	2	10	High	C
21	Katznelson & Gerke, 2021	1	1	1	2	1	1	7	Moderate	C
22	Stokel-Walker, 2023	1	1	0	1	1	1	5	Low	P
23	Masters et al., 2024 (AMEE 172)	2	2	1	2	2	2	11	High	C
24	Almansour & Alfhaid, 2024	1	1	0	1	1	1	5	Low	NR
~~25~~	Lu, 2025	2	1	1	2	1	1	8	Moderate	E
26	Ahmed, 2023	2	1	1	2	1	1	8	Moderate	C
27	Masters et al., 2025 (AMEE 178)	2	1	1	2	1	1	8	Moderate	C
28	Ballantine et al., 2024	2	1	1	2	1	1	8	Moderate	C
29	Tolsgaard et al., 2023 (AMEE 156)	2	1	1	2	1	1	8	Moderate	C
30	Jiang et al., 2025	2	2	2	2	2	2	12	High	SR
31	Herschbach et al., 2025	1	1	1	2	1	1	7	Moderate	E
32	Temper et al., 2025	1	1	1	1	1	1	6	Moderate	C
33	Rai et al., 2025	1	1	1	1	1	1	6	Moderate	C
34	Simoni et al., 2025	2	1	1	2	1	1	8	Moderate	SR

SANRA items: S1=justification of the review's importance; S2=statement of concrete aims; S3=description of the literature search; S4=referencing; S5=scientific reasoning; S6=clarity of presentation. Each item scored 0–2 by both authors independently; maximum total score=12. Weight tiers: High ≥9 (n=17); Moderate 6–8 (n=11); Low/illustrative <6 (n=6). No source was excluded on SANRA grounds alone. Source types: E=empirical study; SR=systematic/scoping review; P=policy/guidance document; C= onceptual/theoretical paper; NR = narrative review.

Source Weighting and Integration

The source types were not treated as epistemically equivalent. Empirical studies were given primary weight for claims about prevalence, institutional practices, and learner behaviours. Policy documents from bodies such as the EU, WHO, GMC, and national regulators were used to characterize the formal governance environment, but not as evidence of actual implementation or effectiveness. Conceptual and ethical papers informed the interpretation of risks and values but were not used to support empirical frequency claims. During thematic synthesis, empirical findings were used to anchor the themes, which were then interpreted through ethical and policy frameworks, making the normative model explicitly derivative of and constrained by available data.

A thematic synthesis approach was employed. Findings fall into three major pillars:

1. The State of Institutional Regulation: Assessing the gap between technological adoption and policy implementation

2. Bioethical Principles in the Algorithmic Age: Reinterpreting autonomy, beneficence, non-maleficence, and justice

3. Generative AI and Scholarship: Addressing the crisis of authorship and truthfulness

Results

1. The Regulatory Vacuum: Institutional Readiness vs. Reality

A critical examination reveals a disconnect between the ubiquity of AI usage by learners and

the preparedness of HPE institutions. Theoretical frameworks abound. Empirical evidence suggests a significant policy lag.

While Rush et al. (2025) provide a rigorous audit of U.S. institutions indicating a policy lag, care must be taken not to generalize this finding globally. International bodies show varying degrees of maturity. Institutions rely on broad, pre-existing clauses regarding cheating or plagiarism to manage AI misuse (11). This approach is critically flawed. It frames AI solely as a tool for dishonesty, ignoring its legitimate role as a learning augmentation and the nuanced intellectual property issues it raises. Specific guidance is lacking. Students and faculty exist in a state of ambiguity where allowable use varies from course to course, creating a hidden curriculum where tech-savvy students gain unfair advantages (12).

Alongside these gaps, several sources report constructive institutional responses, including integration of AI literacy into curricula, revised assessment designs, and faculty development initiatives that leverage AI for feedback and simulation (5, 13-15). These examples demonstrate that AI in HPE is not uniformly experienced as a 'crisis' but also as an opportunity for pedagogical innovation when governance is thoughtfully designed.

Governance vs. Pedagogy: Tran et al. critique the tendency to view AI governance as a purely legalistic endeavor. Governance must be "situated" within pedagogical theory. Regulatory bans on GenAI conflict with the educational goal of preparing students for a digital-first healthcare environment (11). There is concern in the reviewed literature that overly restrictive regulatory postures may reduce transparency in student AI use without eliminating it, potentially precluding educators from cultivating the critical AI literacy needed to appraise algorithmic outputs (16, 17). The European Network for Academic Integrity (ENAI) reinforces this. Regulations should focus on transparency rather than prohibition. Unauthorized use is a breach of integrity. Authorized use is a necessary competence (17).

However, in contrast to the U.S. landscape, a scan of international policy reveals a more proactive, albeit varied, regulatory maturation. In the United Kingdom, the General Medical Council (GMC) and Medical Schools Council have moved toward guidance emphasizing professional judgment over prohibition, aligning with the Russell Group's principles on AI literacy (18, 19). Similarly, in Canada, the Royal College of Physicians and Surgeons have advocated for integrating digital health literacy into existing residency frameworks rather than treating AI as an external threat (20). In Australia, the Tertiary Education Quality and Standards Agency (TEQSA) have pioneered guidelines for Assessment Reform, explicitly advising a shift away from text-based outputs that GenAI can easily mimic, toward process-oriented evaluation (21).

Conversely, the literature highlights a widening regulatory gap in Low- and Middle-Income Countries (LMICs). While high-income nations debate academic integrity, reports from the World Health Organization (WHO) indicate that LMICs face data colonialism and infrastructure deficits (22). Available evidence indicates that regulatory frameworks in these regions are frequently absent, nascent, or transplanted from high-income settings without adaptation to local resource constraints, risking the creation of a two-tiered global medical education system in which AI-augmented training is accessible primarily to the privileged (22, 23).

2. Bioethical Principles in the Algorithmic Age

The four traditional pillars of biomedical ethics—autonomy, beneficence, non-maleficence, and justice— serve as the dominant framework in the literature for evaluating AI. Authors consistently argue these principles warrant radical reinterpretation, especially when applied to the "black box" of machine learning.

Autonomy and the "Black Box" Problem: Patino et al. and Weidener & Fischer highlight a threat. AI poses a danger to professional autonomy. Modern Deep Learning models are often black boxes; their decision-making pathways are opaque even to their creators. In an educational context, if a medical student relies on an AI tutor or diagnostic tool they cannot interrogate, they cede professional autonomy to the algorithm (17, 24). This creates a dependency that undermines the epistemic authority of the future physician. The literature converges on the view that effective governance frameworks for educational settings should prioritize Explainable AI requirements, enabling students to interrogate rather than merely consume algorithmic outputs. Students benefit not only *from* learning with AI, but from developing the capacity to *audit* and critically interrogate algorithmic output. This preserves their capacity for independent judgment (25).

Non-Maleficence: Bias and Hallucination The principle of non-maleficence (do no harm) is most threatened by two technical failures of AI: algorithmic bias and hallucination. Katznelson and Gerke argue that because AI models are trained

on historical healthcare data, they frequently encode systemic racism and sexism. Diagnostic algorithms trained on fair-skinned populations may fail to recognize lesions on darker skin (26). If HPE institutions uncritically integrate these tools into curricula, they risk perpetuating health disparities by training a generation of doctors on biased data. Furthermore, hallucinations where GenAI fabricates convincing but false citations or medical facts pose a direct safety risk. Unlike clinical contexts, ethical breaches in HPE may not cause immediate patient harm but can exert long-term effects on competence, professionalism, and equity. Sallam and Sallam emphasize that without strict verification regulations, the use of LLMs in medical education could lead to the internalization of false medical knowledge, eventually harming patients (2).

Weidener and Fischer propose a principle-based approach that explicitly expands biomedical ethics with public health ethics concepts such as proportionality and common good orientation, reflecting the population-level consequences of AI-enabled education (24). This hybrid framework is particularly relevant for HPE, where educational decisions shape future workforce behaviours at scale.

Justice: The Digital Divide: Busch et al. raise concerns regarding the justice pillar. High-quality AI educational tools (such as personalized adaptive learning platforms) are expensive. There is a tangible risk: creating a two-tiered educational system where well-resourced medical schools produce AI-augmented super-graduates, while resource-poor institutions are left behind (3). Equitable access represents an under-addressed dimension of AI governance in HPE; existing frameworks would benefit from explicit provisions addressing resource disparities between institutions ensuring that the integration of AI into HPE does not exacerbate existing global disparities in medical training.

3. Generative AI, Scholarship, and the Crisis of Truth

The release of ChatGPT in late 2022 precipitated a field-wide reorientation of ethical attention — from data privacy and algorithmic bias toward the integrity of the scientific record itself, encompassing authorship attribution, peer review validity, and knowledge fabrication (27). This shift is directly relevant to HPE, where the integrity of medical scholarship intersects with professional formation and the production of clinical knowledge.

The Erosion of Authorship: Patino et al. discuss the profound implications of AI for medical scholarship and research training. If an AI can generate a literature review, formulate a hypothesis, and write a manuscript, the traditional definition of "authorship" which implies intellectual contribution and accountability is destabilized (9). Current criteria like the ICMJE guidelines struggle to adapt. Regulation in HPE should distinguish between AI as a copyeditor (permissible) and AI as an intellectual generator (impermissible for authorship).

Academic Integrity and Assessment: Generative AI has intensified longstanding concerns regarding academic integrity. ENAI recommendations emphasize that AI-assisted content generation challenges traditional definitions of authorship, originality, and misconduct, requiring recalibration rather than prohibition (17).

In HPE, these concerns are magnified by the stakes of professional competence. Tran et al. argue that over-reliance on generative AI risks undermining authentic assessment, potentially distorting the epistemic relationship between learner and knowledge (11, 16). Without explicit ethical guidance, students may adopt AI as a cognitive surrogate rather than a learning scaffold.

The demonstrated capability of GenAI to pass examinations such as the USMLE has prompted fundamental reconsideration of the validity assumptions underpinning many traditional assessment formats (25, 28). Knopp et al. warn of a future where medical education devolves into a cat-and-mouse game of surveillance, institutions using invasive proctoring AI to catch students using generative AI (28). This adversarial approach is ethically fraught.

Literature advocates for a pivot toward authentic assessment. Evaluations requiring the demonstration of skills like physical exams or patient counseling that AI cannot easily fake. Tran et al. argue that governance should move away from detecting AI (which is technically unreliable) toward designing assessments where AI use is either impossible or explicitly integrated into the task (11).

4. Equity, Bias, and Professional Identity Formation

Algorithmic bias remains a consistently cited ethical risk. AI systems trained on non-representative datasets may reinforce existing inequities in assessment, feedback, and learner progression. Sallam and Sallam identify algorithmic opacity and commercialization as structural risks that disproportionately affect

under-resourced institutions and learners (2, 29).

Beyond distributive justice, authors highlight concerns about professional identity formation. Generative AI's capacity to simulate clinical reasoning and interpersonal interaction raises questions about moral development, empathy, and the cultivation of professional virtues. Tran et al. caution that efficiency gains may inadvertently erode reflective practice if pedagogical intentionality is lacking (30, 31). Notably, few empirical studies examine how AI-mediated education influences identity formation longitudinally, indicating an important research gap.

A balanced synthesis requires acknowledgment of a substantive counternarrative within the reviewed literature. Alongside the governance gaps documented above, a meaningful body of evidence describes AI integration in HPE yielding positive educational outcomes in the absence of regulatory crisis. A study reports that AI-enhanced simulation environments demonstrably improves procedural confidence and clinical reasoning in medical students without triggering the academic integrity failures predicted by prohibitionist frameworks (13). AMEE guides documented structured AI literacy programs, when embedded prospectively within curricula, reduced uncritical AI dependency rather than reinforcing it (28, 32). A recent study describes agentic AI frameworks for general practitioner training that enhances consultation skill acquisition while preserving tutor oversight (14). These findings suggest that the policy lag and ethical vulnerabilities identified in this review are not inevitable outcomes of AI integration per se, but they are contingent on the *absence* of deliberate governance design. Framing HPE's relationship with AI exclusively as crisis management risks generating counterproductive regulatory responses prohibition or excessive restriction that foreclose the pedagogical benefits documented in this literature. The risk-based taxonomy proposed in this review is accordingly designed to enable contextually appropriate adoption, not to function as a barrier to integration (Figure 2).

Discussion

This study used a critical narrative review rather than a systematic or scoping review design. Accordingly, the governance model and risk-based taxonomy presented here should be interpreted as a theoretically informed synthesis of existing evidence and policy frameworks, not as a definitive or exhaustive guideline. The claims made remain conditional on the selected literature and are intended to generate hypotheses and guide future empirical and policy work rather than to close debate. Given the narrative design, the governance recommendations and taxonomy are offered as a conceptual scaffold to stimulate debate and empirical testing, not as a prescriptive policy.

The use of the term "critical" in this review requires methodological clarification. Following Sukhera (33) and Grant and Booth (34) a critical narrative review applies an explicit theoretical or evaluative lens — here, the four-principles bioethics framework and cross-sector regulatory frameworks (EU AI Act; NIST RMF) — to interrogate rather than merely describe the extant literature. This is distinct from critical appraisal-based systematic review methodology and from critical theory as a philosophical orientation. The governance model and taxonomy presented here should, therefore, be read as theoretically informed and normatively positioned, not as prescriptive or exhaustive. While traditional

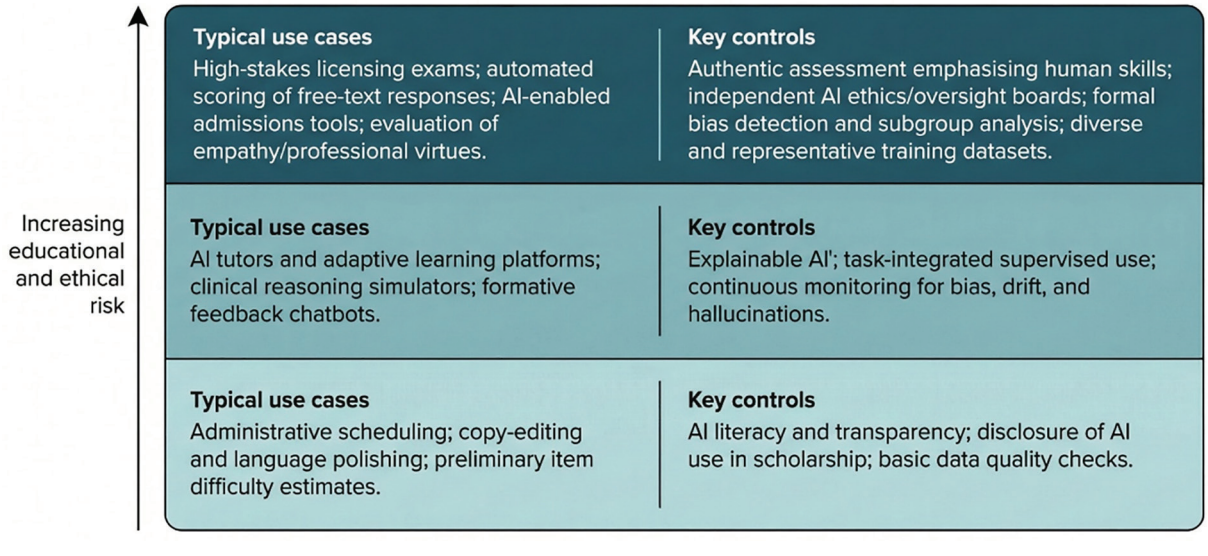

Figure 2. Risk-based taxonomy of AI use in Health Professions Education

systematic reviews are crucial for evidence synthesis in mature fields, the nascent and rapidly evolving landscape of AI in Health Professions Education necessitates flexible, theoretically oriented approach for initial conceptualization and theory-building (35). A critical narrative review, as employed here, allows for the synthesis of diverse literature including empirical studies, policy documents, and theoretical papers to identify emergent ethical and regulatory challenges, synthesize conceptual frameworks, and propose initial governance models (33). This approach is particularly valuable in fields where empirical data is still accumulating, serving as a foundational step to articulate a risk-based taxonomy and guide future systematic research and policy development, rather than offering definitive, empirically validated prescriptions.

The synthesis, therefore, reflects a dual picture; between 2020 and 2025, the period during which large-scale generative models first became widely available, the literature simultaneously documents substantial governance gaps and meaningful experiments in using AI to enhance access, feedback, and personalization in HPE (15). The framing of AI as both threat and opportunity is a product of this mixed evidence base rather than a purely pessimistic stance adopted by the authors (36).

Recent literature frames AI ethics in HPE not as an individual moral responsibility, but as a challenge of governance. AMEE Guides 156 and 158 emphasize institutional readiness, interdisciplinary collaboration, and proactive ethics integration, rather than reactive rulemaking (5, 37). Meaningful regulation in HPE cannot rely solely on institutional policy because learner behavior and assessment practices and national/international legal and standards environments interact continuously. Therefore, the regulatory responses should be risk- and context-sensitive, with proportional obligations aligned to the educational and downstream clinical risks posed by a given AI use case (11). Instead of broad prohibitions, this approach advocates for governance that fosters AI literacy and safe adoption, ensuring future physicians are competent in navigating AI-integrated healthcare environments.

To ensure interoperability and rigor, we propose a Risk-Based Taxonomy for HPE that explicitly maps to major cross-sector frameworks (Figure 2 and Table 2).

• High Risk (Summative Assessment & Admissions): This aligns with the EU AI Act's classification of AI used in education and vocational training as high-risk. For high-risk applications, alignment with the EU AI Act's conformity assessment requirements including bias auditing and documented human oversight represents the most defensible governance position currently available in the literature (11, 38). Similarly, this aligns with the NIST AI Risk Management Framework's Manage function, requiring continuous monitoring of high-impact systems.

• Medium Risk (Simulated Patients & Formative Feedback): This maps to the transparency obligations in the OECD AI Principles. Users (students) should be explicitly informed they are interacting with an AI. Governance here focuses on the "human-in-the-loop" to mitigate hallucination risks (39).

• Low Risk (Brainstorming & Translation): Consistent with the EU AI Act's minimal risk category, governance here focuses on literacy and disclosure rather than restriction (40).

Synthesis of a Governance Model: The literature did not present a single unified framework (41). Consequently, we synthesized the Risk-Based Taxonomy (Table 2) by mapping the specific educational risks identified in the thematic analysis like hallucination and bias (42) against the regulatory tiers of the EU AI Act and NIST Risk Management Framework described in the policy documents. This taxonomy is, therefore, a deductive application of cross-sector regulation to the specific context of HPE. Moreover, this model shifts governance from reactive policing to an educational framework, aligning regulation with learning theories like social constructivism. The RiskBased Taxonomy should be read as an applied conceptual model, generated by mapping the themes from HPE literature onto existing crosssector frameworks.

Building upon the four emergent themes identified in our 'Findings' section, namely the regulatory vacuum, reinterpretation of bioethical principles, crisis of scholarship, and issues of equity and identity formation, this discussion now moves to a critical synthesis. We integrate these observed patterns from the literature with established cross-sector frameworks to propose a risk-based governance taxonomy, offering a conceptual model for navigating the complex ethical and regulatory landscape of AI in HPE.

HPE should evolve from general principal statements toward institutional governance systems, characterized by multi-level alignment, explicit accountability allocation, and lifecycle oversight with documented decision gates. Data governance and evaluation should be linked to educational validity and authenticity (5, 6, 11).

The model shifts governance from reactive policing to an educational framework,

Table 2. Risk-Proportionate Governance and Fairness Protocols in HPE

Risk Category	Concrete HPE Use Cases	Operational Controls & Process Guidance	Fairness & Evaluation Protocols
Low Risk (Administrative/ Preparatory)	AI as a copyeditor; generating preliminary item difficulty estimates; administrative scheduling (31, 38, 43). These were drawn from studies describing administrative applications with minimal impact on learner assessment or identity formation.	Focus on "AI Literacy" and transparency. Prioritise disclosure of AI usage in research and scholarship (24, 29, 30).	Basic data quality checks; ensuring data usage rights are clear before tool adoption (44, 45).
Medium Risk (Formative/ Learning Scaffolds)	AI tutors/adaptive learning platforms; summarizing medical literature; clinical reasoning simulators (45). Reflect literature where hallucination and bias can affect learning but not directly determine progression.	Mandate "Explainable AI" to ensure learners can audit decision-making pathways. Move from "detecting" AI to designing task-integrated use (2, 3).	Identify possible sources of bias (e.g., gender, ethnicity) at the design phase. Continuous monitoring for "drift" or hallucinations (24, 37).
High Risk (Summative/ Identity Formation)	Passing medical licensing exams; automated scoring of free-text responses; evaluating professional virtues or empathy (38). correspond to contexts in which AI outputs can directly affect licensure and professional identity.	Pivot toward "Authentic Assessment" (e.g., physical exams, patient counseling) that AI cannot easily fake. Establish independent "AI Ethics Boards" distinct from IRBs (46).	Formal bias detection using metrics like true positive rates and statistical parity. Mandate diverse and representative training datasets to avoid racial/gender disparities in diagnostic training (47).
Existential Risk (Future Horizon)	AGI systems surpassing human cognition in diagnosis and counseling (48).	Focus on "human-centric" competencies such as complex ethical reasoning and physical presence. Establish lifecycle oversight with documented decision gates (49).	Aggressive protection of non-computational aspects of medicine. Implementation of hybrid principle-based/public health ethics frameworks (49).

AGI, artificial general intelligence; EU, European Union; AI, artificial intelligence; HPE, health professions education; IRB, Institutional Review Board; NIST RMF, National Institute of Standards and Technology AI Risk Management Framework; OECD, Organisation for Economic Co-operation and Development. *Note:* This table represents a conceptual model — not an empirically derived or prescriptive guideline. Risk categories map to cross-sector frameworks as described in the Discussion.

aligning regulation with learning theories like social constructivism to treat AI as a more knowledgeable other rather than just a threat to integrity (11).

Future reviews using systematic or scoping methods will be required to test whether the emphasis on policy lag and ethical risk seen in this critical narrative synthesis is representative of the broader evidence base.

Conclusion

The reviewed literature consistently characterizes the governance of AI in HPE as predominantly reactive with most institutions relying on existing academic misconduct clauses rather than purpose-built AI framework. To ensure the safety and integrity of medical education, institutions may need to consider shifting from academic misconduct policies to robust, value-based governance. We propose that one option is the adoption of risk-based taxonomies, integration of AI literacy into curricula, and establishment of oversight bodies that balance technological affordances with humanistic values. By aligning HPE policy with global frameworks like the EU AI Act and NIST RMF, while addressing the unique vulnerabilities of learners, institutions can move from reactive policing to proactive integration. By treating AI as a more knowledgeable other requiring supervision, rather than a threat to be banned, HPE can prepare the next generation of clinicians for a digitally transformed healthcare landscape.

While the immediate focus of governance should be on the current generation of GenAI and LLMs, the literature suggests that the long-term research agenda in HPE may benefit from anticipating the evolution of AI. Future inquiries should move beyond reactive policy to proactively investigate the pedagogical and ethical implications of increasingly autonomous systems, such as AGI. This includes longitudinal studies on how AI-mediated education influences professional identity formation and the development of moral reasoning, ensuring that future regulatory frameworks are prepared for

the next paradigm shift in medical training.

Priority areas for further work emerging from this synthesis include:

• Shifting policies beyond simple bans on AI toward mandating AI literacy, so future health professionals are competent in the ethical use, auditing, and recognition of the limitations of AI tools.

• Establishing specialized educational ethics bodies to review AI tools for bias, privacy, and pedagogical validity, recognising that the black box nature of many systems suggests the need for greater transparency than traditional educational materials.

• Ensuring that, as AI assumes more cognitive tasks, the ethical framework of HPE actively protects and values the noncomputational dimensions of medicine care, empathy, and physical presence.

Other priorities may emerge as the evidence base continues to develop

Declaration of AI

Artificial intelligence (Google Gemini) was utilized to review this text for grammatical correctness and clarity. The core content, arguments, and conclusions are the original work of the author.

Conflicts of Interest

The authors declare no conflicts of interest. No funding was received for this review. No commercial AI tools, vendors, or educational technology companies had any role in the design, conduct, or reporting of this work.

Authors' Contribution

Y.A conceived and designed the review, formulated the search strategy, conducted the database searches, participated in source screening and quality appraisal using the SANRA scale, synthesized the findings, and drafted the manuscript. S.K participated in study screening, independent SANRA quality scoring, reflexive thematic coding, and contributed to the critical revision of the manuscript. Both authors contributed to developing the risk-based governance taxonomy, reviewed and revised the manuscript for important intellectual content, and approved the final version for submission.

References

1. Bleakley A. Medical Humanities and Medical Education: How the medical humanities can shape better doctors [Internet]. London: Taylor and Francis; 2016 [Cited 2026 May 24]. 1–264 p. Available from: https://www.taylorfrancis.com/books/mono/10.4324/9781315771724/medical-humanities-medical-education-alan-bleakley.
2. Sallam M, Sallam M. Ethical aspects of implementing generative artificial intelligence in medical education: a narrative review. History and Philosophy of Medicine. 2025;7(4):20.
3. Busch F, Adams LC, Bressem KK. Biomedical Ethical Aspects towards the Implementation of Artificial Intelligence in Medical Education. Med Sci Educ. 2023;33(4):1007–12.
4. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon. 2024;10(4):e26297.
5. Masters K. Ethical use of Artificial Intelligence in Health Professions Education: AMEE Guide No. 158. Med Teach. 2023;45(6):574–84.
6. Franco D'Souza R, Mathew M, Mishra V, Surapaneni KM. Twelve tips for addressing ethical concerns in the implementation of artificial intelligence in medical education. Med Educ Online. 2024;29(1):2330250.
7. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach (Dordr). 2018;28(4):689–707.
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science (1979). 2019;366(6464):447–53.
9. Patino GA, Amiel JM, Brown M, Lypson ML, Chan TM. The Promise and Perils of Artificial Intelligence in Health Professions Education Practice and Scholarship. Acad Med. 2024;99(5):477–81.
10. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Quality & Quantity. 2021;56(3):1391–412.
11. Tran M, Balasooriya C, Jonnagaddala J, Leung GKK, Mahboobani N, Ramani S, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. npj Digital Medicine. 2025;8(1):315.
12. Rush E, Byram JN, Garnett CN, DeVaul N, Smith L, Checchi M, et al. An audit of AI-related documents across U.S. medical schools: A framework-based qualitative content analysis. Med Teach. 2025;48(3):493-505.
13. Hamilton A. Artificial Intelligence and Healthcare Simulation: The Shifting Landscape of Medical Education. Cureus. 2024;16(5):e59747.
14. Ichikawa T, Olsen E, Vinod A, Glenn N, Hanna K, Lund GC, et al. Generative Artificial Intelligence in Medical Education—Policies and Training at US Osteopathic Medical Schools: Descriptive Cross-Sectional Survey. JMIR Med Educ. 2025;11:e58766.
15. Lee J, Wu AS, Li D, Kulasegaram K (mahan). Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. Acad Med. 2021;96(11S):S62–70.
16. Tran M, Balasooriya C, Semmler C, Rhee J. Generative artificial intelligence: the 'more knowledgeable other' in a social constructivist framework of medical education. npj Digital Medicine 2025 8:1. 2025;8(1):430.
17. Foltynek T, Bjelobaba S, Glendinning I, Khan ZR,

Santos R, Pavletic P, et al. ENAI Recommendations on the ethical use of Artificial Intelligence in Education. International Journal for Educational Integrity. 2023;19(1):12.
18. Group Russell. Russell Group principles on the use of generative AI tools in education [Internet]. 2023 [Cited 2025 Dec 19]. Available from: https://nationalcentreforai.jiscinvolve.org/wp/2023/05/11/generative-ai-primer/.
19. The General Medical Council [Internet]. 2025 [Cited 2025 Dec 19]. Artificial intelligence and innovative technologies: GMC. Available from: https://www.gmc-uk.org/professional-standards/learning-materials/artificial-intelligence-and-innovative-technologies?utm_source=chatgpt.com.
20. Rees G, Nowell L, Risling T. Shaping the Future of Digital Health Education in Canada: Prioritizing Competencies for Health Care Professionals Using the Quintuple Aim. JMIR Med Educ. 2025;11:e75904.
21. TEQSA. Gen AI strategies for research training: Emerging practice. Australian Government: TEQSA; 2025.
22. Asiedu MN, Dieng A, Haykel I, Rostamzadeh N, Pfohl S, Nagpal C, et al. The Case for Globalizing Fairness: A Mixed Methods Study on Colonialism, AI, and Health in Africa. Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. Africa: EAAMO; 2024.
23. Yu L, Zhai X. Use of artificial intelligence to address health disparities in low- and middle-income countries: a thematic analysis of ethical issues. Public Health. 2024;234:77–83.
24. Weidener L, Fischer M. Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education. JMIR Med Educ. 2024;10(1):e55368.
25. Knopp MI, Warm EJ, Weber D, Kelleher M, Kinnear B, Schumacher DJ, et al. AI-Enabled Medical Education: Threads of Change, Promising Futures, and Risky Realities Across Four Potential Future Worlds. JMIR Med Educ. 2023;9(1):e50373.
26. Katznelson G, Gerke S. The need for health AI ethics in medical school education. Adv Health Sci Educ Theory Pract. 2021;26(4):1447–58.
27. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature. 2023;613(7945):620–1.
28. Masters K, Herrmann-Werner A, Festl-Wietek T, Taylor D. Preparing for Artificial General Intelligence (AGI) in Health Professions Education: AMEE Guide No. 172. Med Teach. 2024;46(10):1258–71.
29. Almansour M, Mohammad Alfhaid F. Generative artificial intelligence and the personalization of health professional education A narrative review. Medicine (United States). 2024;103(31):e38955.
30. Lu Q. Development of generative artificial intelligence in medical education: a bibliometric profiling. Front Educ (Lausanne). 2025;10:1613067.
31. Ahmed Y. Utilization of ChatGPT in Medical Education: Applications and Implications for Curriculum Enhancement. Acta Informatica Medica. 2023;31(4):300.
32. Masters K, MacNeil H, Benjamin J, Carver T, Nemethy K, Valanci-Aroesty S, et al. Artificial Intelligence in Health Professions Education assessment: AMEE Guide No. 178. Med Teach. 2025; 47(9):1410-24.
33. Sukhera J. Narrative Reviews: Flexible, Rigorous, and Practical. J Grad Med Educ. 2022;14(4):414–7.
34. Grant MJ, Booth A. A typology of reviews: An analysis of 14 review types and associated methodologies. Health Info Libr J. 2009;26(2):91–108.
35. Furley P, Goldschmied N. Systematic vs. Narrative Reviews in Sport and Exercise Psychology: Is Either Approach Superior to the Other? Frontiers in Psychology. Front Psychol. 2021;12:685082.
36. Ballantine J, Boyce G, Stoner G. A critical review of AI in accounting education: Threat and opportunity. Critical Perspectives on Accounting. 2024;99:102711.
37. Tolsgaard MG, Pusic MV, Sebok-Syer SS, Gin B, Svendsen MB, Syer MD, et al. The fundamentals of Artificial Intelligence in medical education research: AMEE Guide No. 156. Med Teach. 2023;45(6):565–73.
38. Jiang T, Er S, Heron MJ, Zhu KJ, Yang R. Artificial Intelligence Use in Academic Applicant Screening: A Systematic Review. Plast Reconstr Surg Glob Open. 2025;13(10):e7177.
39. Herschbach L, Festl-Wietek T, Stegemann-Philipps C, Sonanini A, Herrmann B, Erschens R, et al. Evaluation of an AI-Based Chatbot Providing Real-Time Feedback in Communication Training for Mental Health Care Professionals: Proof-of-Concept Observational Study. J Med Internet Res. 2025;27:e82818.
40. Temper M, Tjoa S, David L. Higher Education Act for AI (HEAT-AI): a framework to regulate the usage of AI in higher education institutions. Front Educ (Lausanne). 2025;10:1505370.
41. Purdy RJ. Toward a Certification Framework for AI Governance in Education: Design Principles for a Sector-Appropriate Standard [Internet]. 2026 Feb [Cited 2026 Apr 29]. Available from: https://papers.ssrn.com/abstract=6193658 doi:10.2139/SSRN.6193658.
42. Rai S, Verma K, Yadav V. Generative AI in Medical Pharmacology: Balancing Educational Benefits and Hallucination Risks Running Title: AI Hallucinations in Pharmacology Teaching. International Journal of Science and Research. 2025;14(4):1158–69.
43. Simoni J, Urtubia-Fernandez J, Mengual E, Simoni DA, Royo M, Egaña-Yin D, et al. Artificial intelligence in undergraduate medical education: an updated scoping review. BMC Med Educ. 2025;25(1):1609.
44. Sasseville M, Yousefi F, Ouellet S, Naye F, Stefan T, Carnovale V, et al. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. Healthcare (Switzerland). 2025;13(12):1447.
45. Khakpaki A. Advancements in artificial intelligence transforming medical education: a comprehensive overview. Med Educ Online. 2025;30(1):2542807.
46. Chan A, Rahimi-Ardabilli H, Rogers WA, Coiera E. The real-world impact of artificial intelligence ethics frameworks across a decade in healthcare: a scoping review. Journal of the American Medical Informatics Association. 2025;32(11):1767–77.
47. Bouhouita-Guermech S, Gogognon P, Bélisle-Pipon JC. Specific challenges posed by artificial intelligence in research ethics. Front Artif Intell. 2023;6:1149082.

48. Sass R. Defining dangerous AI: existential risk, power-intelligence, and the limits of AGI. AI and Ethics. 2025;5(5):5557–73.
49. Jha D, Durak G, Sharma V, Keles E, Cicek V, Zhang Z, et al. A Conceptual Framework for Applying Ethical Principles of AI to Medical Practice. Bioengineering. 2025;12(2):180.