sustainability

MDPI

Article

Deepfake-Style AI Tutors in Higher Education: A Mixed-Methods Review and Governance Framework for Sustainable Digital Education

Hanan Sharif [1,*], Amara Atif [2] and Arfan Ali Nagra [1]

[1] Faculty of Computer Science, Lahore Garrison University, Lahore 5400, Pakistan; arfanalinagra@lgu.edu.pk
[2] School of Computer Science, University of Technology Sydney, Sydney 2007, Australia; amara.atif@uts.edu.au
* Correspondence: hanankhan386@gmail.com or hanansharif@lgu.edu.pk

Abstract

Deepfake-style AI tutors are emerging in online education, offering personalized and multilingual instruction while introducing risks to integrity, privacy, and trust. This study aims to understand their pedagogical potential and governance needs for responsible integration. A PRISMA-guided, systematic review of 42 peer-reviewed studies (2015–early 2025) was conducted from 362 screened records, complemented by semi-structured questionnaires with 12 assistant professors (mean experience = 7 years). Thematic analysis using deductive codes achieved strong inter-coder reliability ($\kappa = 0.81$). Four major themes were identified: personalization and engagement, detection challenges and integrity risks, governance and policy gaps, and ethical and societal implications. The results indicate that while deepfake AI tutors enhance engagement, adaptability, and scalability, they also pose risks of impersonation, assessment fraud, and algorithmic bias. Current detection approaches based on pixel-level artifacts, frequency features, and physiological signals remain imperfect. To mitigate these challenges, a four-pillar governance framework is proposed, encompassing Transparency and Disclosure, Data Governance and Privacy, Integrity and Detection, and Ethical Oversight and Accountability, supported by a policy checklist, responsibility matrix, and risk-tier model. Deepfake AI tutors hold promise for expanding access to education, but fairness-aware detection, robust safeguards, and AI literacy initiatives are essential to sustain trust and ensure equitable adoption. These findings not only strengthen the ethical and governance foundations for generative AI in higher education but also contribute to the broader agenda of sustainable digital education. By promoting transparency, fairness, and equitable access, the proposed framework advances the long-term sustainability of learning ecosystems and aligns with the United Nations Sustainable Development Goal 4 (Quality Education) through responsible innovation and institutional resilience.

Keywords: deepfake AI tutors; synthetic media in education; online education governance; academic integrity; AI ethics in education; detection of deepfakes; privacy and fairness in AI; AI literacy; sustainable education; digital sustainability; SDG 4 quality education

Academic Editor: Fernando Moreira

Received: 5 September 2025
Revised: 28 October 2025
Accepted: 30 October 2025
Published: 3 November 2025

Citation: Sharif, H.; Atif, A.; Nagra, A.A. Deepfake-Style AI Tutors in Higher Education: A Mixed-Methods Review and Governance Framework for Sustainable Digital Education. *Sustainability* **2025**, *17*, 9793. https://doi.org/10.3390/su17219793

1. Introduction

Advances in generative artificial intelligence (AI) have enabled synthetic audio-visual agents that convincingly mimic human expression and speech. In education, such capabilities underpin what are sometimes termed "deepfake AI tutors" [1,2]. These avatars are powered by deep learning technologies and deliver instruction in real time or via prerecorded content.

In this paper, deepfake AI tutors are synthetic, avatar-based instructional agents produced with deep learning that either (i) replicate the likeness or voice of a real person (e.g., face-swap, voice cloning) or (ii) embody entirely synthetic personas with high-fidelity, real-time verbal and nonverbal interactivity. They differ from (a) traditional talking-head or text-only virtual assistants that lack identity mimicry and (b) agentic AI systems (multi-agent/planning models) whose primary risks arise from autonomy and goal-directed behavior rather than identity replication. The distinctive risk profile of deepfake tutor centers on mimicry and identity provenance during instructional delivery [3].

Proponents emphasize the potential for expanded access, scalability, and engagement through personalized learning experiences [4]. However, critics caution that such systems may introduce risks of impersonation, misinformation, and erosion of trust in educators [5–8]. Parallel research on learning analytics has similarly highlighted the role of data-driven personalization in shaping engagement and feedback [9], demonstrating how technological innovations can enhance equity while introducing new governance challenges [10].

Recent research highlights increasing adoption and evolving attitudes toward generative AI in academic contexts [11]. A peer-reviewed study involving surveys of 823 higher-education students, teachers, and researchers found that trust in generative AI was the strongest predictor of adoption, driven more by user experience than demographics. While not specific to deepfake, this finding is particularly relevant for synthetic tutors. Similarly, recent studies indicate that generative AI can enhance engagement and efficiency, while introducing risks of academic dishonesty [12,13].

Accordingly, the discussion of detection methods, governance controls, and assessment integrity in later sections refers specifically to deepfake AI tutors rather than AI in education in general.

1.1. Motivation and Historical Trajectory

Generative artificial intelligence and deepfake technologies have advanced significantly over the past decade. Deep autoencoders and other early deep learning methods showed that face-swapping was possible. The later development of Generative Adversarial Networks (GANs) enabled high-fidelity synthesis of audio-visual content with striking realism [14,15]. These innovations have since paved the way for educational applications, where synthetic tutors can deliver multilingual instruction around the clock, adapt dynamically to individual learning styles, and simulate human social presence [16,17]. The historical trajectory highlights both the promise and peril of synthetic media in learning environments, from the first widely publicized deepfake videos in 2017 to the rapid adoption of large language models like GPT-4o in 2024 [18]. This technological convergence marks a significant evolution from early AI tutoring experiments such as the Programmed Logic for Automated Teaching Operations (PLATO) system of the 1960s and the Intelligent Tutoring System (ITS) of the 1990s, which primarily relied on text-based prompts and rule-based logic. In contrast, today's synthetic tutors deliver realism through high-fidelity audio-visual generation, scalability through continuous multilingual instruction, and interactivity through features that simulate authentic social presence. Alongside this technological evolution, researchers have increasingly turned to human factors, examining how learners psychologically and emotionally engage with synthetic tutors.

The pedagogical and governance dimensions of deepfake-style AI tutors are closely connected to sustainability in education. As digital transformation accelerates, sustainability now extends beyond environmental concerns to include social, ethical, and institutional continuity in technology-enhanced learning. In alignment with the United Nations Sustainable Development Goals, particularly SDG 4 (Quality Education), SDG 9 (Industry, Inno-

vation and Infrastructure), and SDG 16 (Peace, Justice and Strong Institutions), this study advances the idea of educational sustainability by emphasizing responsible innovation, equitable access, and long-term governance mechanisms for trustworthy AI integration in higher education.

Recent psychological research further suggests that the perceived empathy and anthropomorphic features of generative AI agents can positively influence students' sociocultural adaptation and engagement [19]. These findings emphasize that the design of synthetic tutors should incorporate empathy cues and transparent disclosure practices to enhance trust and acceptance among learners [20].

1.2. Regulatory Landscape and Ethical Debates

Building upon this historical overview, the regulatory landscape surrounding deepfake AI tutors has evolved rapidly. The deployment of deepfake AI tutors intersects with evolving frameworks in AI governance. Key initiatives include the European Union's AI Act (Regulation 2024/1689), which came into force on 1 August 2024 [21]. The regulation adopts a risk-based model that classifies AI systems, including educational applications, as high-risk when they are subject to human oversight requirements and compliance checks.

At the global level, UNESCO's Guidance on Generative AI in Education and Research (2023) emphasizes human-centered implementation, proposing ethical strategies to ensure equity, accountability, and capacity building in educational contexts [22]. Beyond governmental regulation, professional organizations such as the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE) have issued codes of ethics that stress fairness, transparency, and accountability in the design and use of AI-based educational tools [23].

Scholars, such as Chesney and Citron [24], are actively debating whether deepfake tutors should be classified as high-risk under the EU AI Act, given their potential impact on learners' rights. They highlight concerns around student identity protection, misuse of cloned likenesses, and ensuring that learners are not deceived about instructors' authenticity and qualifications. These issues illustrate the importance of ethically informed design alongside technical innovation.

Beyond formal regulation, policy briefs and professional bodies are also responding to the rapid adoption of generative AI. For instance, the National Association of Foreign Student Advisers (NAFSA) has noted the widespread weekly use of AI tools by international educators, recommending robust best practices around transparency, bias mitigation, data privacy, and preventing hallucinations [25]. These considerations further motivate a governance framework specially tailored to deepfake tutors.

1.3. Contribution and Scope

Prior reviews have surveyed deepfake technologies broadly or focused primarily on legal implications without a specific lens on education [2]. Notably, Roe et al. [26] reviewed 182 publications on detection, misuse, and possible benefits but found no studies explicitly addressing higher education. Similarly, thematic analyses of academic narratives challenge simplistic "pros vs. cons" views of deepfakes [27]. To our knowledge, no prior study has combined empirical literature synthesis with grounded expert perspectives to propose a deployable governance framework for deepfake AI tutors in educational settings.

Our contributions are threefold:

1. **A PRISMA-compliant systematic literature review** of peer-reviewed sources (2015–early 2025), specifically focused on deepfake tutors in educational contexts—a gap not addressed by existing scoping or technical reviews.

2. **A thematic synthesis** from semi-structured questionnaires with assistant professors across disciplines, offering practice-grounded insights into both pedagogical potential and ethical safeguards.
3. **Development of actionable governance artifacts**, including a policy checklist, responsibility matrix, risk-tier matrix, cross-modal detection matrix, and a phased adoption roadmap—to guide institutions toward responsible deployment.

Prior reviews rarely integrated technical detection, pedagogical outcomes, and policy analysis, but our study bridges these domains while incorporating emerging evidence on trust, sociocultural adaptation, and fairness [28]. These contributions collectively establish the foundation for the review and the proposed governance framework described in subsequent sections.

1.4. Research Questions

To address the identified gaps and guide our study, we formulated the following research questions:

RQ1: How can deepfake-style AI tutors enhance personalization and learner engagement in higher education?

RQ2: What challenges do they pose to academic integrity and detection mechanisms?

RQ3: What governance and ethical frameworks are needed to ensure responsible deployment?

These questions directly align with our study's aims, ensuring that both the literature review and expert semi-structured interviews administered in writing via a standardized questionnaire remain focused on evaluating benefits, risks, and governance strategies for deepfake AI tutors. Each research question corresponds to one of the core themes emerging from our analysis (engagement benefits, integrity risks, governance needs, and ethical implications), providing a clear roadmap for the subsequent sections of the paper.

2. Literature Review

Building on the introduction, this section reviews existing studies addressing deepfake tutors, related pedagogical agents, and governance perspectives. The existing literature covers reported pedagogical benefits, ethical and legal risks, and advances in technical detection. In this section, we synthesize and organize prior research into thematic subsections to contextualize our review.

2.1. Empirical Evidence on Deepfake Tutors

Research on ITS and pedagogical agents demonstrates that personalized feedback, adaptive pacing, and multimodal interaction enhance both learning outcomes and learner motivation [29]. Herrero et al. [30] reported that affective ITS employing real-time learner state detection improved cognitive and emotional engagement. Similarly, Wambsganss et al. [31] found that conversational tutoring systems with adaptive interactivity increased learner satisfaction and knowledge retention.

Empirical studies focused specifically on deepfake tutors remain limited. Initial experiments suggest that realistic avatars can improve attention, retention, and satisfaction in language and STEM education [31]. However, these studies also identify potential drawbacks. Highly realistic tutors may evoke discomfort through the uncanny valley effect, and learners may over-attribute authority to avatars, thereby diminishing critical thinking.

In view of the limited direct evidence, insights from broader research on generative AI provide valuable parallels. Large-scale surveys indicate that trust in generative AI is a key predictor of adoption [12]. Theoretical work further distinguishes between trust based on perceived humanness and trust grounded in system reliability, both of which

influence learner engagement and educational outcomes [32]. These results indicate that trust, familiarity, and anthropomorphic design significantly influence the acceptance of synthetic tutors, while also highlighting the necessity for additional research specifically targeting deepfake applications in education [33].

2.2. Detection Technologies and the Arms Race

Detecting deepfakes, including potential misuse in educational settings, has progressed from focusing on low-level visual artifacts to leveraging frequency and physiological signals. Early methods used convolutional neural networks (CNNs) trained on datasets like FaceForensics++ to detect blending, warping, and compression inconsistencies [34,35]. Researchers subsequently extracted noise residuals and phase spectrum features to enhance cross-domain generalization [36].

Recent methods combine frequency-domain analysis and local binary patterns with biometric signals. One example is remote photoplethysmography (rPPG), which estimates heart rate from video [37]. Other approaches integrate multimodal cues, such as audio-visual synchronization or voice-biometric matching [38,39]. Hybrid detection models, including CNN-LSTM frameworks, exploit both spatial and temporal inconsistencies for enhanced performance, while spatio-temporal convolutional networks further improve detection across video frames [40].

As generative pipelines advance (e.g., diffusion models, neural rendering, audio-driven avatars), no single detection feature remains robust across all sources or conditions. This arms race is important for deepfake AI tutors because it has a direct effect on identity provenance, impersonation risks, and the protection of student identity during teaching and testing. Reliable, layered detection therefore remains essential in educational settings, with lighter checks appropriate for routine instruction and stronger safeguards necessary in high-stakes examinations [38].

Detection is technical, but it also has social effects on deepfake AI tutors. In education, fairness and generalizability are critical: detectors may underperform for some skin tones, accents, or non-Western demographics, creating disparate false positives/negatives that can harm students (e.g., mistaken identity flags in tutorials or exams). Recent work explores synthetic data augmentation and multi-task/fairness-aware objectives to reduce bias and improve cross-domain performance [41]. For tutor deployments, this implies local bias evaluation, human-in-the-loop review for adverse decisions, and transparent appeal pathways so integrity safeguards do not undermine equity [42,43].

2.3. Ethical, Legal, and Social Dimensions

Scholars in education and ethics emphasize that deepfake tutors raise complex issues beyond technical detection. Key concerns include informed consent, disclosure of synthetic identity, manipulation of trust, and protection of training data. These questions are not abstract. For a university deploying deepfake tutors, they translate into concrete responsibilities around how student interactions are recorded, monitored, and safeguarded [44,45]. Legal scholars further highlight gaps in copyright and right-of-publicity laws when avatars mimic real instructors without consent [46]. This issue has led to debates over whether new protections are required, including a "digital performance right". Such a right would grant individuals control over the use of their likeness, voice, and identity in synthetic media [47], much like existing protections for creative works. While legal debates focus on ownership and identity rights, educational ethicists direct attention to issues of access and equity.

Here, the discussion shifts from who controls identity to who benefits from the technology. Some argue that synthetic tutors could democratize access to high-quality instruction,

particularly in underserved regions. Others caution that commercial platforms may exploit student data for profiling or surveillance, raising concerns about fairness and student autonomy [2,48]. Cultural context adds another layer. Cross-cultural studies show that student reactions to synthetic instructors vary widely. In East Asia, animated avatars are commonly accepted, whereas in many Western contexts, students report discomfort when real instructors' likenesses are simulated. For example, experimental comparisons of AI-generated and human-made teaching videos reveal that acceptance is often culturally dependent [49,50].

Empirical research on student perceptions remains limited. A recent survey [51] shows that 40% of international educators use generative AI weekly, while about 70% of U.S. teenagers have experimented with these tools. Respondents highlight ethical concerns, including transparency, bias, data privacy, hallucinations, and call for clearer citation and data governance frameworks.

Together, these perspectives suggest that ethical, legal, cultural, and equity debates cannot remain abstract. For universities and schools considering deepfake tutors, the real challenge lies in balancing the potential benefits of access and personalization against risks of surveillance, bias, and identity misuse in everyday learning environments.

2.4. Technical Gaps and Research Needs

Despite growing interest in deepfake AI tutors, evidence remains fragmented. Much of the literature benchmarks stand-alone detectors on curated datasets that do not reflect classroom realities (e.g., variable lighting, compression, screen-capture artifacts). By contrast, tutor deployments require identity-centric safeguards: (i) provenance/disclosure of synthetic media, (ii) liveness and identity verification (face/voice consistency checks to prevent impersonation), and (iii) multi-modal detection only where risk is high (e.g., exams), with privacy-preserving settings and clear due process for students. Empirical gaps persist on learning outcomes, integrity impacts, and how repeated exposure affects trust and classroom dynamics; few studies integrate ethical, technical, and educational requirements into a single governance approach. Beyond technical safeguards, effective governance of deepfake tutors also depends on how well educators and students are prepared to understand and critically engage with AI technologies.

Bridging this gap also requires attention to AI literacy, which shapes how educators and students can critically interpret and respond to synthetic instruction. A recent integrative review of AI literacy across K–12 and higher education identifies three key dimensions: functional, critical, and indirectly beneficial literacy. It also highlights a growing shift toward generative AI tools and prompt engineering [52,53]. While not specific to deepfake tutors, these findings underline that responsible deployment of such tutors will depend on parallel literacy initiatives that help stakeholders understand disclosures, recognize risks, and engage constructively with AI-mediated learning. This view aligns with evidence that AI literacy supports academic performance and well-being in AI-enriched environments [54].

Overall, the literature reveals fragmented evidence spread across technical, ethical, and educational domains. While progress has been made in detection and personalization, few studies integrate these perspectives into a unified governance framework. This gap directly motivates the mixed-methods approach and the governance model proposed in this study.

3. Methodology

3.1. Research Design

We adopted a mixed-methods design that combined a PRISMA-guided Systematic Literature Review (SLR) [55] and a semi-structured written questionnaire for expert input. The questionnaire was designed to elicit detailed, comparable responses from participants. The SLR identified and synthesized peer-reviewed literature from high-impact venues, while the expert input provided practice-grounded insights into risks and safeguards. Data from both components were analyzed using thematic and structured synthesis to ensure methodological triangulation.

3.2. Systematic Literature Review (PRISMA)

3.2.1. Search Strategy

The SLR adhered to the PRISMA 2020 guidelines. We searched IEEE, Scopus, Web of Science, and Google Scholar using the following Boolean string:

(deepfake* OR 'face swap*' OR 'face reenact*' OR 'voice clone*' OR 'lip sync*' OR 'media synthesis' OR 'synthetic media' OR 'virtual human*' OR 'digital human*' OR 'avatar*' OR 'talking head*' OR 'embodied conversational agent*') AND (tutor* OR instructor* OR teacher* OR 'pedagogical agent*' OR 'education' OR 'e-learning' OR 'assessment')

Searches were limited to peer-reviewed, English-language articles published between 2015 and 2025 (inclusive), with 2025 covering publications available up to the date of the search (July, 2025). This initial search yielded 362 records (IEEE Xplore: 97; Scopus: 108; Web of Science: 84; Google Scholar: 73, based on the top 200 results sorted by relevance). No records were retrieved from trial registers or gray literature sources.

3.2.2. Screening and Eligibility

After deduplication (n = 112), 250 unique records remained. Two independent reviewers screened titles and abstracts, excluding 130 for irrelevance. The full-text screening of 120 reports applied the following criteria:

Inclusion: Empirical studies (quantitative, qualitative, or mixed-methods) evaluating or discussing deepfake or synthetic tutors in educational contexts (K–12, higher education, professional training).

Exclusion: Non-educational applications, absence of empirical data, generic chatbots without deepfake capabilities, opinion pieces, patents, and non-English publications.

A total of 42 studies met the inclusion criteria. Screening disagreements were resolved through discussion or adjudication by a third reviewer.

3.2.3. Data Extraction and Charting

From each included study, we extracted bibliographic information, study design, educational context, tutor type, outcome measures, and key findings. Data were charted in structured tables to allow systematic comparison across detection approaches, governance implications, and reported pedagogical effects. The extraction process was piloted on five sample papers and refined iteratively to ensure consistency and comprehensiveness.

3.2.4. PRISMA Flow Diagram

As summarized in Figure 1, 362 records were retrieved. After removing 112 duplicates, 250 unique records remained. Following screening and eligibility checks, 42 studies met the inclusion criteria for synthesis.

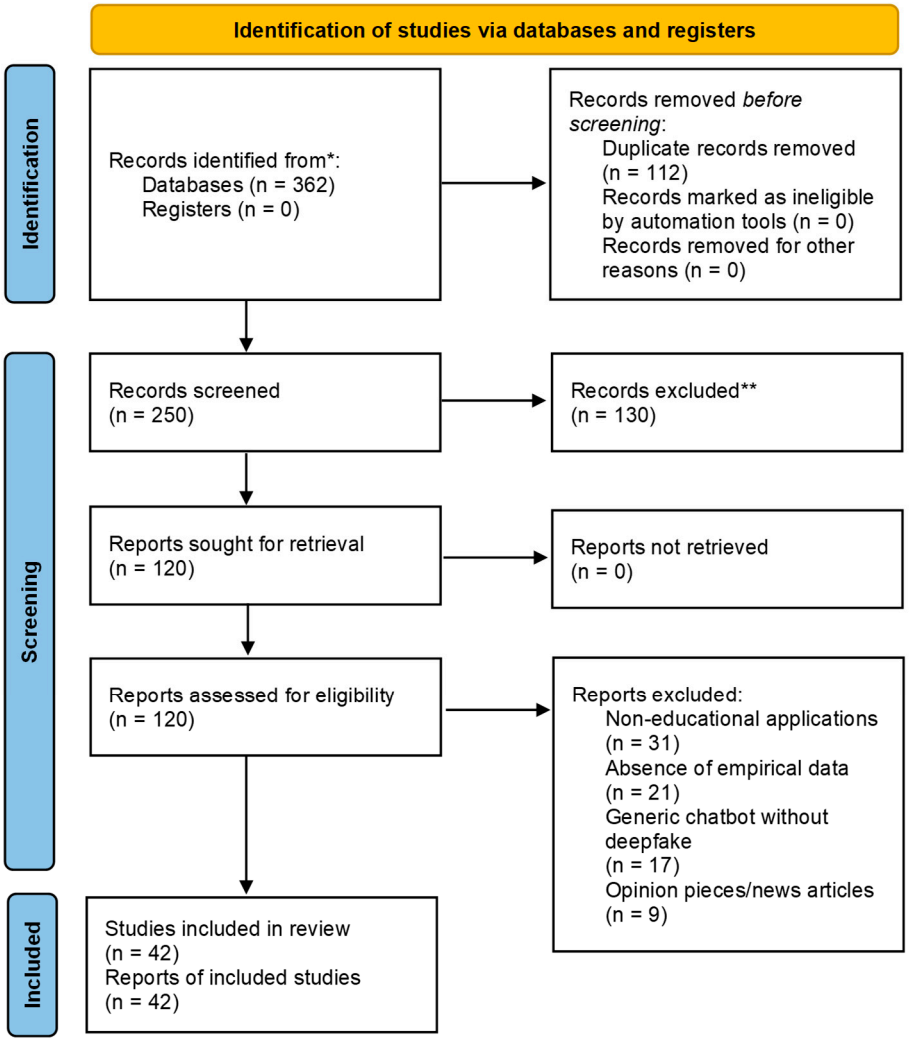

Figure 1. PRISMA 2020 flow diagram illustrating identification, screening, eligibility, and inclusion steps. * Number of records identified from each database or register searched (rather than the total number across all databases/registers). ** Records excluded by a human.

3.2.5. Quality Appraisal

The methodological quality of the included studies was systematically appraised using the Mixed Methods Appraisal Tool (MMAT, 2018) [56]. This tool evaluates qualitative, quantitative, and mixed-methods research against five quality indicators. Scores were expressed as a proportion of criteria met. For example, a score of 3/5 indicates that three criteria were met satisfactorily, while two showed weaknesses. Out of the 42 studies reviewed, 26 were rated as high-quality ($\geq$4/5), reflecting strong methodological soundness with only minor limitations. Twelve studies were rated moderate (3/5), indicating acceptable but incomplete adherence to quality standards. Four studies scored below 3/5 and were categorized as low quality, showing notable weaknesses in study design, reporting, or both.

Inter-rater agreement between the two independent reviewers was high (Cohen's κ of 0.81), confirming strong reliability in the evaluation process. In synthesizing results, greater interpretive weight was placed on high-quality studies, whereas findings from low-quality studies were included but considered cautiously, serving primarily to highlight possible trends rather than to support firm conclusions. This rigorous appraisal ensured that the synthesis was grounded in methodologically sound evidence, directly supporting our aim of developing a governance framework for deepfake tutors.

3.3. Expert Questionnaires and Thematic Analysis

3.3.1. Participants and Sampling

Twelve assistant professors participated in this study through semi-structured written questionnaires. Participants represented four academic disciplines (computer science, software engineering, education, and ethics). Of these, 75% were male and 25% female, with a mean teaching and research experience of approximately seven years. All participants were affiliated with Lahore Garrison University and had prior familiarity with AI-driven educational technologies.

The decision to recruit twelve participants was guided by qualitative research conventions: this sample size was sufficient to achieve thematic saturation while remaining manageable for in-depth analysis. A purposive sampling strategy was employed to guarantee pertinent expertise and disciplinary diversity among early-career academics proficient in educational technology. Summary demographics are provided here for transparency, while Appendix A presents additional details for reference.

3.3.2. Data Collection

A standardized semi-structured written questionnaire (pro forma) containing open-ended questions (Appendix B) was distributed electronically. Participants provided detailed written responses that were subsequently exported to Microsoft Excel for analysis. All identifying information was removed during data cleaning, and participants were assigned anonymized codes (AP#1–AP#12). Ethical approval was obtained, and informed consent was documented prior to participation.

The questionnaire instrument was developed based on themes identified in the literature and refined through expert input to ensure content validity. It also included an initial question about each participant's prior experience with or exposure to AI-driven tutors, thereby gauging AI familiarity at the outset. This design choice provided context for interpreting responses; for instance, it distinguished comments from those who had hands-on experience versus those speaking hypothetically, and helped ensure that all participants shared a baseline understanding of the topic before answering further questions.

3.3.3. Data Analysis

We employed a thematic analysis following Braun and Clarke's [57] six-phase approach using a hybrid deductive–inductive process:

1. **Familiarization:** Researchers read all questionnaire responses multiple times.
2. **Coding:** Initial codes were generated using a literature-informed codebook, supplemented by open coding to capture new insights.
3. **Theme identification:** Related codes were grouped into candidate themes.
4. **Theme review:** The coding team refined and validated themes against the data.
5. **Theme naming:** Each theme was defined and labeled to reflect its essence.
6. **Reporting:** Compelling excerpts were selected to illustrate each theme.

Two researchers coded data independently, achieving high inter-coder reliability (Cohen's $\kappa = 0.81$). Discrepancies were resolved through discussion, ensuring consistency. NVivo 14 was used to organize transcripts, visualize patterns, and maintain an audit trail.

3.3.4. Triangulation

Themes from the semi-structured written questionnaires were compared against SLR findings to identify convergences and divergences. Where themes aligned, they reinforced the evidence base; where they diverged, differences highlighted contextual nuances or research gaps. This triangulation enhanced the robustness of conclusions by integrating practice-based expert perspectives with empirical evidence from peer-reviewed literature.

3.4. Operational Definitions and Taxonomy

An operational taxonomy of deepfake tutor types and modes of operation was created using information from both the SLR and expert questionnaires. This taxonomy classifies deepfake tutors across three key dimensions: tutor construction, delivery mode, and contextual controls. Together, these categories provide a framework for analyzing risks, benefits, and governance strategies. While our taxonomy classifies tutors primarily by identity and delivery mode, other recent frameworks emphasize agentic dimensions such as autonomy, multi-agent collaboration, and tool use [3]. These approaches complement our focus by highlighting the diversity of AI architectures in education. Figure 2 presents the integrated conceptual framework developed from both the systematic literature review and expert questionnaire findings. It illustrates how the construction and delivery of deepfake AI tutors influence the need for specific contextual controls, which in turn inform the broader governance pillars of transparency, privacy, integrity, and ethical accountability.

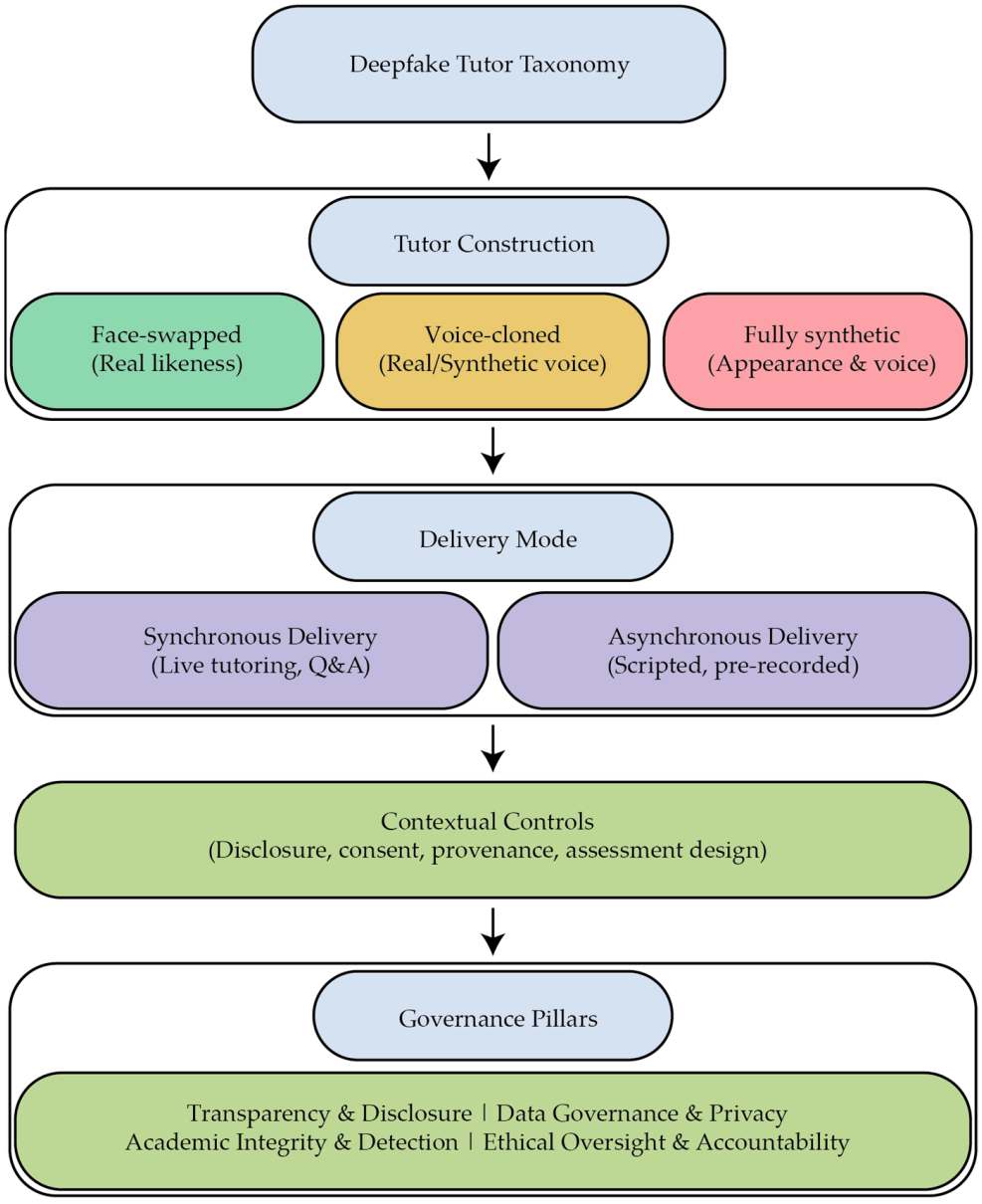

Figure 2. Integrated conceptual framework illustrating the relationship between tutor construction, delivery mode, contextual controls, and governance pillars for deepfake AI tutors.

3.4.1. Tutor Construction

Three forms of tutor identity are distinguished. Face-swapped tutors replicate the visual likeness of a real individual, typically an educator, while retaining another's or a synthetic voice. Voice-cloned tutors reproduce a real or synthetic voice, with varying degrees of fidelity, but may be paired with either a synthetic or generic avatar. Finally, fully synthetic tutors combine AI-generated visual and vocal features, creating avatars with no direct human counterpart. Each type raises distinct issues of consent, authenticity, and intellectual property.

3.4.2. Delivery Mode

Deepfake tutors can operate in synchronous contexts (e.g., live tutoring, question-and-answer sessions) or asynchronous contexts (e.g., scripted or pre-recorded lessons). Synchronous delivery amplifies interaction and adaptability but heightens risks of impersonation or real-time misuse. Asynchronous delivery offers scalability and reusability but raises concerns about provenance, storage, and replay manipulation.

3.4.3. Contextual Controls

The taxonomy also incorporates contextual control safeguards that influence how tutors are perceived and governed. These include disclosure of synthetic identity, informed consent, verification of provenance, and assessment design that minimizes opportunities for impersonation or academic dishonesty. Such controls are critical to ensuring transparency, trust, and accountability in deployment.

By integrating tutor type, delivery mode, and contextual safeguards, this taxonomy provides a structured lens for both researchers and institutions. This taxonomy not only categorizes technical configurations, but also pinpoints the areas that require governance interventions to strike a balance between innovation and ethical responsibility.

3.4.4. Detection Classes and Technical Constraints

This study concentrates on the pedagogical and governance aspects of deepfake AI tutors, while the technical detection framework supports integrity assurance in educational environments. Table 1 summarizes representative detection classes, their benchmark datasets, common evaluation metrics, and known limitations. These categories reflect how advances in general deepfake detection research can be translated to educational contexts, helping institutions select appropriate authentication and integrity safeguards.

Table 1. Detection classes relevant to deepfake AI tutors, with representative datasets, common evaluation metrics, and typical limitations affecting deployment in educational settings.

Detection Class	Representative Datasets	Common Metrics	Key Limitations
Pixel-level CNN/Visual artifacts	FaceForensics++, DeepfakeTIMIT, DFDC	AUC 0.85–0.98; F1	Vulnerable to adversarial attacks and improved generators
Noise/Texture features	Celeb-DF v2, WildDeepfake	Accuracy; F1; EER	Dataset-specific cues; poor cross-domain generalization
Frequency-domain features	FaceForensics++, Celeb-DF	AUC; F1	Frequency-aware generators can spoof cues
Physiological signals (rPPG)	DFDC-Phys, Celeb-DF	AUC; EER	High-quality fakes mimic plausible physiological patterns
Audio-visual sync and Biometrics	DFDC, Audio-Visual DeepFake	EER; precision/recall	Coordinated spoofing; privacy and fairness concerns

Addendum: This summary of deepfake detection methods is presented in the context of AI tutors, highlighting how general advances in deepfake detection research translate into the educational domain. In particular, the table emphasizes which detection approaches are most relevant for authenticating tutor videos and where their weaknesses might pose challenges in an academic environment.

4. Results

Four dominant themes emerged from the analysis: (1) Personalization and Learner Engagement; (2) Detection Limitations and Academic Integrity Risks; (3) Governance and Policy Gaps; and (4) Ethical and Societal Implications. While these themes arose inductively from the coding process, they align closely with the study's research questions (RQ1–RQ3). The fourth theme extends the ethical dimension of RQ3 by capturing participants' broader reflections on trust, transparency, and societal acceptance of synthetic tutors. The themes and their alignment with the research questions are summarized in Table 2.

Table 2. Themes derived from expert questionnaires (n = 12), including concise definitions, frequency across participants, and illustrative quotations.

Theme	Definition	Frequency (n = 12)	Illustrative Quote
Personalization and Engagement	Deepfake tutors support adaptive pacing, multimodal feedback and multilingual delivery.	11	"The avatar keeps students engaged, especially in large online cohorts". AP#3
Detection Challenges & Integrity Risks	High-quality forgeries evade detectors; risk of impersonation and exam fraud.	10	"We need biometrics for high-stakes exams—detectors alone won't suffice". AP#7
Governance and Policy Gaps	Institutions lack clear policies on disclosure, consent and provenance.	9	"Policy clarity and accountability are essential before any deployment". AP#1
Ethical and Societal Implications	Concerns about trust, privacy and bias in synthetic tutors.	8	"Students should know when an instructor is synthetic—transparency is non-negotiable". AP#4

Table 3 summarizes the misuse vectors identified and the corresponding controls, along with feasibility and institutional burden ratings. The patterns show that most high-feasibility controls, like requiring disclosure and verifying identity, put a moderate burden on institutions. On the other hand, technically intensive safeguards, like multimodal detection pipelines, are still resource-heavy for universities with limited infrastructure. Similar governance trade-offs have been discussed by Chesney and Citron (2024) [24] and in UNESCO's Guidance on Generative AI in Education (2023) [58], both emphasizing proportionality between risk and implementation cost. These findings reinforce the need for adaptive governance models that prioritize high-impact, low-burden measures and justify the phased framework proposed in Section 6 of this study.

Table 3. Misuse Vectors in Deepfake Tutor Deployments and Corresponding Controls (Feasibility and Burden).

Misuse Vector	Controls	Feasibility	Institutional Burden
Identity Impersonation	Biometric authentication; liveness checks; verified device binding	High	Medium
Assessment Fraud (in tutor-mediated contexts)	Hybrid exams; oral defenses; multi-modal evidence (drafts, logs) when AI tutors are used for formative or summative assessments	Medium	High
Content Manipulation	Provenance watermarking; cryptographic hashes; version control	Medium	Low
Data Privacy Breach	Data minimization; access controls; secure storage and audits	High	Medium

4.1. Integration of Questionnaires and SLR Findings

Themes derived from semi-structured interviews administered in writing via a standardized questionnaire strongly overlapped with SLR-identified domains, particularly regarding learner engagement benefits, detection challenges, and policy gaps. However, expert respondents placed greater emphasis on immediate, institution-ready safeguards such as biometric verification and hybrid exam models, an aspect less frequently covered in the literature.

4.2. Systematic Review Findings

The PRISMA-guided review identified 42 unique peer-reviewed studies. Because many papers addressed more than one topic, we coded each study to one or more domains; the domain counts below are therefore non-mutually exclusive and exceed 42.

- Engagement and Learning Benefits (17 studies).
- Detection and Technical Development (26 studies; some overlapping).
- Misuse and Privacy Risks (12 studies).
- Governance and Policy (8 studies).

Studies on benefits consistently reported that synthetic tutors could enhance motivation, support multilingual instruction, and deliver scalable feedback. However, learning gains varied; several quasi-experiments showed no significant difference compared to conventional video lectures when controlling for content and duration.

Detection studies mainly evaluated algorithms on benchmark datasets such as FaceForensics++ and Celeb-DF. Although these studies improved detection performance, their dependence on curated datasets constrained ecological validity, especially in live classroom environments where issues like inadequate lighting and screen capture distortions frequently occur. Misuse research documented identity impersonation in assessments and phishing attempts, while privacy studies explored the legal status of cloned faces and voices. Governance-oriented papers proposed disclosure frameworks but lacked empirical field tests in educational environments.

4.2.1. Tutor Type Distribution

Among the studies, generic synthetic personas accounted for the majority (60%). These tutors typically appeared as neutral avatars with no direct human likeness. Identity-bearing clones, such as celebrities or instructor avatars, represented 25% of the sample and raised ethical concerns due to issues of consent and authenticity. Hybrid systems (15%), which combined AI-generated voices with human-curated visuals, reflected an emerging category that sought to balance realism with reduced impersonation risks.

4.2.2. Learning Context Distribution

The distribution of learning contexts was relatively balanced across sectors: higher education (~33%), language learning platforms (~33%), and professional training (~34%). This spread suggests that interest in deepfake tutors is not confined to a single educational domain but rather reflects a broad experimentation with synthetic instruction across multiple levels of learning.

5. Governance Framework for Deepfake AI Tutors

Building on the risk themes and mitigation gaps identified in the results, this section proposes a structured framework to guide the safe and ethical adoption of deepfake AI tutors. The framework is organized into four core pillars, each with minimum (foundational) and advanced (mature) controls. Together, these pillars provide both immediate safeguards and a roadmap for progressive institutional maturity. Operationalization strate-

gies, technical detection considerations, and pilot deployment procedures are detailed to enable phased implementation in educational environments.

5.1. Four-Pillar Structure

The proposed governance framework consists of four interlocking pillars:

- **Transparency and Disclosure**

Learners and educators must be informed whenever synthetic tutors are deployed. Disclosure reduces risks of deception and helps maintain trust in instructional relationships.

- **Data Governance and Privacy**

Robust data practices protect both student and institutional information while ensuring compliance with evolving legal standards across jurisdictions. In the European Union, the AI Act (2024) [59] classifies many educational applications as high-risk, requiring transparency and oversight. In the United States, FERPA governs the privacy of student records, while in Australia the Privacy Act 1988 (amended 2022) [60] sets national standards for data handling. Together, these frameworks emphasize the need for institutions to embed privacy-by-design principles into the deployment of deepfake tutors.

- **Academic Integrity and Detection**

Reliable detection mechanisms, ideally multimodal, are needed to prevent the misuse of deepfake AI tutors, including impersonating real instructors, generating fraudulent assessment support, or presenting unverified identities. In the tutor context, detection safeguards ensure that synthetic instructional agents cannot be exploited to undermine academic integrity. This pillar directly addresses the risks identified in the literature and by our expert participants.

- **Ethical Oversight and Accountability**

Clear governance structures assign responsibility for oversight, monitoring, and incident response. Accountability mechanisms ensure that institutions can manage ethical risks alongside technical ones.

Table 4 presents the policy checklist for each pillar, including both foundational and advanced controls.

Table 4. Policy checklist by pillar.

Pillar	Minimum Controls (Foundational)	Advanced Controls (Mature)
Transparency and Disclosure	On-screen labels; staff and student consent; syllabus notices	Dynamic disclosure toggles; user opt-outs; transparency logs
Data Governance and Privacy	Data minimization; DPO oversight; encryption; retention limits	Differential privacy; zero-trust storage; third-party audits; liveness detection; fairness audits; explicit retention windows; appeals
Academic Integrity and Detection	Biometric verification of examinees; tutor-detection pipeline; manual review of flagged cases	Cross-modal provenance; continuous monitoring; red teaming
Ethical Oversight and Accountability	AI ethics committee; incident response SOP	External audits; public reports; KPIs and dashboards

These four governance pillars were derived from recurring risk domains identified across both the systematic literature review and expert feedback. Each pillar addresses a distinct but interrelated objective essential to the responsible deployment of deepfake AI tutors. The Transparency pillar preserves clarity and authenticity by ensuring that

synthetic instructional agents are identified and traceable. The privacy pillar safeguards the data, consent, and identity rights of both educators and learners. The Integrity pillar focuses on maintaining the authenticity and fairness of instructional and assessment processes through robust detection and verification mechanisms. Finally, the Accountability pillar enforces ethical oversight, delineating institutional responsibilities and promoting auditability. Together, these four pillars provide a comprehensive foundation for balancing innovation with ethical and regulatory compliance in higher education environments.

5.2. Cross-Modal Detection Matrix and Failure Modes

Effective governance depends on deploying detection tools that address multiple attack vectors. Table 5 shows the strengths and weaknesses of five major types of detectors. Visual artifact detectors remain the most widely studied, efficiently identifying spatial inconsistencies, but they are highly sensitive to compression, low-light conditions, and novel generation methods. Noise- and texture-based approaches extend this capacity by capturing subtle residual patterns, yet they often overfit to specific datasets and generalize poorly across domains. Frequency-domain methods offer an alternative viewpoint by identifying spectral anomalies; however, they are susceptible to frequency-aware generators and may experience significant degradation under substantial compression. Physiological-signal approaches, such as remote photoplethysmography (rPPG), use heart rate and blink detection as markers of authenticity; while promising, these methods demand high-quality video and raise additional privacy and fairness concerns. Finally, provenance and watermarking solutions use cryptography to link media to its source, giving strong guarantees of where it came from. However, they need to be used by everyone in the ecosystem and can sometimes be bypassed by re-encoding.

Table 5. Cross-modal detection matrix for deepfake tutors.

Detector Family	Key Strengths	Common Failure Modes
Visual artifact	Detects spatial artifacts; well-studied; efficient	Sensitive to compression, low-light conditions, and novel generators
Noise/texture	Captures subtle noise patterns and textures	Dataset-specific; poor cross-domain generalization
Frequency-domain	Identifies spectral anomalies	Vulnerable to frequency-aware generators and heavy compression
Physiological (rPPG)	Uses heart-rate and blink signals for authenticity	Requires high-quality video; privacy and fairness concerns
Provenance/watermark	Cryptographically binds media to source	Requires ecosystem adoption; can be removed by re-encoding

Addendum: This matrix highlights the complementary strengths and failure modes of different detector families, underscoring that no single detection technique is foolproof. Effective safeguarding of AI tutors may require combining multiple detection approaches. Notably, detection stringency should be calibrated to context, with less intensive measures for low-stakes learning scenarios, whereas high-stakes uses (e.g., exams) demand more robust, multi-layered detection strategies.

These families illustrate that no single technique can provide reliable coverage across all contexts. For deepfake AI tutors, risks span from impersonating legitimate instructors to assisting with fraudulent assessment submissions. Detection cannot depend on just one signal to deal with these different threats. Instead, layered, cross-modal pipelines are needed—for example, combining visual artifact detection (to identify face-swaps), audio/voice verification (to flag cloned speech), and interaction or provenance logs (to detect assessment fraud). By mapping detector types to misuse vectors, institutions can align safeguards directly with the risks outlined in Tables 2 and 3. Accordingly, the pilot configurations outlined in the next section incorporate multiple detector types to evaluate their combined effectiveness under real-world classroom conditions.

5.3. Pilot Configuration and Key Performance Indicators (KPIs)

To validate the framework, we recommend a phased pilot in one or two low-stakes courses employing synthetic personas without identity-bearing clones. A six-month pilot provides sufficient time to observe learner adaptation and institutional workflows while limiting exposure to high-risk misuse. During this period, detection pipelines, liveness checks, and disclosure mechanisms should be fully operational.

Table 6 outlines sample KPIs and their target thresholds. Illustrative metrics include the percentage of learners correctly identifying disclosure notices, detection accuracy across multiple modalities, average system response time, and learner trust scores captured via post-course surveys. These indicators provide both technical and human-centered benchmarks to assess readiness for broader deployment.

Table 6. Sample KPIs and target thresholds for pilot deployments.

Metric	Description	Target Threshold
Detector TPR	Proportion of deepfakes correctly detected	$\geq$90%
Liveness FRR	False-reject rate in liveness detection	$\leq$5%
Time to detection	Average time to flag suspicious content (seconds)	$\leq$60 s
User satisfaction	Students reporting satisfaction (%)	$\geq$80%
Disclosure compliance	Courses with visible disclosure labels	100%
Incident resolution	Incidents resolved within 24 h (%)	$\geq$95%

It is important to note that the choice and stringency of deepfake detection measures should align with the stakes of the educational context. In low-stakes tutor scenarios, for example, supplemental practice modules or ungraded learning activities, minor inaccuracies in detection may be tolerable, and simpler detection techniques (or even periodic manual monitoring) could suffice. By contrast, in high-stakes contexts such as proctored examinations, credit-bearing coursework, or scenarios where tutor identity verification is critical, a much more robust detection approach is required. In these high-risk situations, institutions may need to combine multiple detection methods (e.g., real-time video authenticity checks, biometric verification, and secure user authentication) to minimize the chance of a deepfake tutor being misused for impersonation or cheating. This principle of matching detection rigor to context is reflected in our risk-tier framework: lower-risk educational applications might rely on basic disclosure and trust mechanisms, whereas higher-risk applications demand multi-modal detection solutions and strict oversight. Ultimately, effective governance will require a calibrated approach where detection strategies are scaled to the potential impact of a deepfake tutor's failure in each use case.

5.4. Failure-Mode Playbooks

Institutions should prepare incident-response protocols for detection failures or identity breaches. For example, if a liveness evaluation fails, a secondary verification (e.g., ID card validation) should be triggered, or the system should switch to a supervised offline exam. Similarly, if the unauthorized use of a synthetic tutor is detected in an assessment context (e.g., impersonation or fraudulent draft generation), the exam should be paused, flagged for manual review, and supplemented with alternate items. In this setting, "misuse" refers specifically to the deployment of AI tutors in restricted contexts such as summative examinations, where independent performance is required. Detection would typically be flagged through automated monitoring systems (e.g., anomaly detectors) and then verified by proctors or institutional IT staff.

These playbooks should be embedded in broader institutional cybersecurity and academic integrity procedures, with clear escalation paths, notification protocols, and post-incident reviews to update policies and training.

5.5. RACI Matrix for Pilot Deployment

To ensure clear accountability during pilot implementation, a RACI matrix (Responsible, Accountable, Consulted, Informed) is used to define roles for each key task. This approach outlines who will carry out the task, who will exercise decision-making authority, who will seek expertise, and who must remain informed. Each task is paired with measurable KPIs to track effectiveness. For example, defining the disclosure policy assigns responsibility to the program lead, accountability to the dean or provost, consultation with legal and ethics experts, and information sharing with faculty and students, with the KPI being the percentage of courses displaying disclosure labels. In another case, biometric integrity checks might mean that IT staff are in charge, the Chief Information Security Officer is in charge, accessibility experts are consulted, and students are given information, with the KPI being that the system stays up during tests. Similarly, data governance tasks could assign responsibility to the Data Protection Officer, accountability to senior leadership, consultation with legal advisors, and information sharing with faculty, measured through compliance audits.

This structured role assignment and outcome measurement ensure pilot activities are coordinated, transparent, and aligned with institutional objectives. Table 7 details responsibilities for pilot deployment, ensuring that each task has designated responsible, accountable, consulted, and informed stakeholders.

Table 7. RACI for pilot deployment with example KPIs.

Task	Responsible	Accountable	Consulted	Informed	KPI Example
Define disclosure policy	Program Lead	Dean/Provost	Legal, Ethics Committee	Students, Faculty	Courses with disclosure labels (%)
Data protection controls	IT Security	CIO	DPO, Legal	Faculty, Students	High-stakes exams with verified identity; mean liveness failure rate
Assessment redesign	Assessment Office	Dean	Faculty, QA	Students	Assessments using hybrid exams or multi-modal evidence (%)
Detector pipeline setup	AI/IT Team	CIO	Vendors, Researchers	Faculty	Mean time-to-detection (TTD); detector accuracy
Pilot evaluation and reporting	Research Team	Dean/Provost	Ethics Committee	Stakeholders	Pilot courses completing evaluation (%); stakeholder satisfaction

6. Implications for Policy and Practice

6.1. Policy and Regulatory Guidance

Based on our review, we recommend that policymakers develop clear disclosure mandates, provenance standards, and minimum-security baselines for deepfake AI tutors, particularly in assessment contexts. High-stakes applications, such as credit-bearing exams or identity-bearing tutor use, should be formally designated as higher-risk deployments, thereby triggering stricter controls. Existing regulatory precedents provide useful anchors: the EU AI Act (2024) [59] already classifies many educational AI applications as high-risk; the U.S. Family Educational Rights and Privacy Act (FERPA) govern the protection of student records; and Australia's Privacy Act (amended 2022) sets national standards for data handling.

Institutional policies are also advised to include visible disclosure labels, require informed consent from staff and students, and align with privacy-by-design principles.

Engagement with standardization bodies will be important to harmonize provenance watermarking and interoperability requirements across platforms.

6.2. Institutional Implementation

While policymakers set guardrails, institutions must translate them into practice. A phased deployment strategy should begin with low-stakes instructional contexts and synthetic personas that do not replicate real individuals. Foundational controls should include disclosure labels, provenance watermarking, and AI literacy training for both educators and students. Literacy initiatives are particularly important, as they help learners critically evaluate synthetic interactions and avoid over-attributing authority to avatars. Progression to higher-risk contexts should occur only after pilot evaluations demonstrate acceptable KPI performance (Table 7) and incident rates remain within predefined thresholds. For high-stakes assessments, biometric verification, liveness detection, multi-modal evidence collection, and hybrid exam formats are recommended.

To operationalize these recommendations, deployment contexts were categorized according to their risk level and authenticity demands. Table 8 shows the risk-tier matrix that matches deployment contexts with the right governance controls. The matrix differentiates between instructional (low-stakes) and assessment (high-stakes) scenarios and between synthetic personas and identity-bearing tutors. This tiered approach enables institutions to apply proportionate safeguards that reflect contextual risk and authenticity requirements.

Table 8. Risk-tier matrix for deployment contexts and recommended controls.

Context	Synthetic Persona (No Real Identity)	Identity-Bearing (Cloned Tutor)
Instructional/low-stakes	Disclosure labels; provenance watermarking; AI literacy training	Disclosure labels; provenance; staff consent; fairness auditing
Assessment/high-stakes	Disclosure; multi-modal evidence; hybrid exam design	Biometrics (face and voice); liveness detection; hybrid exams; multi-proctoring

Lower-risk instructional deployments primarily rely on transparency, disclosure, and literacy measures, whereas high-stakes assessment contexts require stronger technical safeguards such as biometric verification, liveness detection, and multi-proctoring. This proportional structure ensures that institutional policies balance practicality with integrity protection and align with the proposed phased governance strategy.

We recognize that not all institutions will have abundant resources to implement advanced technical solutions. Thus, our governance recommendations are designed to be scalable to resource-constrained settings. For example, universities with limited funding or technical expertise can prioritize low-cost measures: emphasizing transparency and disclosure policies, leveraging open-source deepfake detection tools, and adopting simpler identity verification steps (like secure login procedures) before resorting to expensive biometric systems. The framework encourages collaboration and knowledge-sharing (e.g., consortia of institutions pooling resources to develop or access detection tools) as an adaptation strategy. Even schools with limited resources can start using AI tutors responsibly by adjusting the depth of implementation to what they have. They can start with the most practical safeguards and then add more features over time.

6.3. Privacy and Security Assurance

In addition to risk-based governance, privacy threats were systematically mapped using the LINDDUN framework. This framework provides a structured taxonomy of potential data-protection issues in deepfake-enabled tutoring systems and guides the development of targeted mitigation strategies, summarized in Table 9. A LINDDUN-based

privacy threat analysis highlights potential vulnerabilities such as linkability, identifiability, and unauthorized disclosure, and pairs them with targeted mitigations. For example, pseudonymous identifiers and differential privacy can mitigate linkability and identifiability risks, while encryption and access control can address disclosure threats. Detectability concerns, such as inferring student presence from system logs, require minimization of metadata retention. Explicit audit trails combined with opt-out safeguards can counter non-repudiation risks, while layered consent processes and clear disclosure interfaces are necessary to address unawareness risks.

Table 9. LINDDUN privacy threat categories and example controls.

Threat Category	Example Threat in Deepfake Tutors	Mitigation
Linkability	Linking student identities across sessions	Use pseudonymous tokens; minimize data retention
Identifiability	Re-identifying anonymized interactions	Apply differential privacy; limit logging
Non-repudiation	Inability to deny session involvement	Maintain opt-out options; log consent
Detectability	Observing whether a deepfake tutor is used	Encrypt metadata; obfuscate session identifiers
Disclosure	Unauthorized exposure of sensitive data	Enforce encryption; access controls; audits
Unawareness	Students unaware of data collection	Transparent consent forms; clear disclosures
Non-compliance	Breach of data protection laws	Regular audits; DPO oversight; privacy training

For educational deepfake tutors, these threats are especially important when biometric signals (like liveness detection or rPPG) or recorded interactions are stored, which makes it more likely that they will be used in ways that are not allowed or for profiling. Embedding privacy-by-design into learning management systems through default pseudonymization, explicit opt-in for data retention, and restricted data-sharing agreements helps ensure compliance. Continuous monitoring, third-party audits, and institutional privacy reviews should be institutionalized to maintain alignment with evolving data protection regulations such as the EU AI Act, FERPA in the U.S., and Australia's Privacy Act.

By systematically mapping vulnerabilities to mitigations, privacy and security assurance become not only a compliance requirement but also a foundation for building learner trust in synthetic tutoring environments.

The LINDDUN analysis highlights that privacy vulnerabilities in deepfake tutors extend beyond data leakage to include linkability, re-identification, and unawareness. Applying privacy-by-design controls such as pseudonymization, differential privacy, and transparent consent mechanisms can substantially reduce these risks. This systematic mapping reinforces the data governance and Privacy pillar of the framework by translating theoretical privacy principles into actionable institutional controls.

6.4. Capacity Building and Stakeholder Preparedness

Human expertise remains as important as technical controls. Faculty and administrative staff should be explicitly informed about institutional AI tutor deployments and equipped through structured training on deepfake detection, provenance verification, and incident escalation procedures. For example, academic integrity offices could run workshops on identifying synthetic media, while IT staff develop playbooks for detection-system failures. We should also equip students with media literacy skills to identify synthetic content, evaluate the credibility of sources, and report any suspicious interactions. Embedding such training into orientation modules or general education curricula ensures broad coverage.

Institutional support units, including teaching and learning centers, IT security teams, and legal offices, must coordinate governance implementation, policy updates, and compliance monitoring. Appointing a dedicated AI governance lead or cross-unit task force can help ensure accountability and continuity. Cross-departmental workshops and peer-

learning communities provide forums to surface discipline-specific concerns, such as differing sensitivities in STEM, humanities, or professional training contexts.

At a broader scale, capacity building should align with emerging regulatory expectations. For instance, faculty training on AI disclosure could parallel EU AI Act requirements for transparency, while student literacy initiatives could meet accreditation standards in Australia and the U.S. International collaboration through professional associations can further support shared preparedness across jurisdictions.

6.5. Stepwise Adoption Roadmap

The Stepwise Adoption Roadmap (Table 10) provides a practical path from initial pilots to scaled deployment.

Table 10. Stepwise adoption roadmap for deepfake tutors.

Phase (Timeline)	Key Activities	Milestones and Criteria
Phase 1 (0–6 months)	Establish ethics committee; draft policies; procure detectors; pilot synthetic persona tutors; AI literacy training	Policies published; disclosure labels live; KPI thresholds met; stakeholder approval
Phase 2 (6–12 months)	Integrate biometrics and liveness; scale pilots to assessments; conduct audits; engage regulators	Audits passed; low incident rates; positive learning and fairness metrics; external engagement
Phase 3 (beyond 12 months)	Embed deepfake tutor governance into regular operations: continuous monitoring, annual policy updates, refresher training, cross-institutional benchmarking, and international collaboration on interoperability standards	Governance embedded in operations; annual reviews completed; ongoing training delivered; benchmarking reports published; participation in international standard-setting

In Phase 1 (0–6 months), institutions should establish an AI ethics committee, draft disclosure and governance policies, pilot low-risk applications using synthetic personas, and run AI literacy workshops for both faculty and students. To identify and manage early-stage risks before expansion, continuous monitoring of KPI compliance is essential.

In Phase 2 (6–12 months), deployments can be extended into higher-stakes contexts such as credit-bearing courses or formative assessments. At this stage, biometric and liveness verification should be integrated, provenance mechanisms scaled across platforms, and independent audits conducted. Engagement with external regulators, peer institutions, and professional associations will help align institutional practices with evolving global standards such as the EU AI Act, FERPA in the U.S., and Australia's Privacy Act.

In Phase 3 (beyond 12 months), institutions should embed deepfake tutor governance into regular operations. This includes continuous monitoring, annual policy updates, refresher training, cross-institutional benchmarking, and international collaboration on interoperability standards. By treating governance as an ongoing process rather than a one-off deployment, institutions can adapt to technological advances and regulatory changes over time.

By following this sequenced approach, institutions can balance innovation with security, trust, and fairness. The responsible deployment of deepfake AI tutors in education will depend significantly on a layered governance approach that incorporates both technical detection methods and ethical oversight mechanisms.

It should be noted that the proposed governance framework is conceived as a "living" framework; in other words, it is meant to be dynamic and continuously updated. As deepfake AI tutor technology and its use cases evolve, the framework's guidelines and

safeguards can be revised in response to new challenges and evidence. We have included illustrative Key Performance Indicators (KPIs) for each governance pillar to help institutions monitor implementation success; however, these KPIs are currently conceptual benchmarks rather than empirically validated measures. To date, the framework and its KPIs have not been trialed in a real deployment. As such, any KPI targets (e.g., acceptable detection true-positive rates or compliance percentages) serve as starting points. Ongoing pilots and future empirical studies will be essential to test these KPIs, refine their thresholds, and adjust the framework's recommendations based on what is proven effective in practice. In essence, the governance model is intended to remain evidence-informed and iterative, accommodating improvements as stakeholder feedback and new data become available.

7. Discussion

The findings feature the ongoing tension between pedagogical gains and integrity risks when deploying deepfake AI tutors in higher education. Personalization and multilingual delivery can improve engagement and accessibility, yet risks such as identity deception and tutor-enabled assessment fraud (e.g., impersonation or unauthorized assistance during exams) threaten institutional trust. Technical detection, while advancing, remains insufficient on its own. Institutions require layered controls that combine disclosure, provenance, and biometric verification. These results are consistent with prior studies noting the evolution of deepfake generation from spatial artifacts to frequency and physiological cues and the demonstrated capacity of generators to adapt to detection systems over time [38].

Detection research demonstrates a rapidly intensifying competition in technology. Early face-swap and autoencoder models produced visible artifacts that convolutional neural networks could identify. The advent of generative adversarial networks (GANs) and style-based architectures improved realism, prompting detectors to incorporate noise signatures, texture features, and frequency-domain analyses [61]. Recent diffusion and neural rendering models can now produce not only photo-realistic faces but also coherent lip movements and plausible physiological signals, challenging even advanced detection pipelines. This progression highlights the importance of multi-modal detection strategies and robust provenance mechanisms rather than reliance on any single detection method.

By combining detection research from different fields, we stress the importance of making sure that the technical capabilities of deepfake AI tutors match the rules and regulations that govern them. Our cross-modal detection matrix (Table 5) shows that different detector types offer complementary strengths and weaknesses. For instance, physiological detectors can help stop visual spoofing, but they also raise privacy and fairness issues. This is because bias in rPPG and liveness systems has been shown to have a bigger effect on people with darker skin tones and non-native accents [62]. Provenance mechanisms can circumvent the detection arms race, yet require ecosystem-level adoption of watermarking or cryptographic standards. The risk-tier matrix (Table 8) operationalizes these principles by assigning more stringent controls to high-stakes contexts and identity-bearing tutors.

Our results should also be interpreted in light of emerging evidence on user trust, sociocultural adaptation, and fairness in generative AI. Large-scale surveys in higher education show that trust in AI tools is a significant predictor of behavioral intention, most strongly influenced by usage frequency and self-perceived proficiency rather than demographic variables [63–65]. This aligns with our questionnaire's findings, suggesting that exposure and targeted training will be critical for acceptance of deepfake tutors, illustrating the importance of institutional AI literacy initiatives. In parallel, emerging work on agentic AI systems in education highlights that governance challenges will soon extend beyond identity mimicry to include autonomous behaviors, planning, and multi-

agent coordination [3]. This suggests that governance frameworks for deepfake tutors must be adaptable to evolving AI paradigms. Cross-cultural studies further highlight that acceptance varies across contexts, with animated avatars more readily accepted in East Asia compared to the discomfort reported in Western settings, implications that governance frameworks must account for.

These insights show that sustainability in digital education is achieved not only through technological efficiency but also through ethical durability. Systems and practices must maintain trust, fairness, and accountability over time. The governance model proposed in this study supports sustainable education by ensuring that AI tutors promote inclusive access while protecting academic integrity and institutional credibility. This balance between innovation and governance provides a foundation for long-term sustainability in higher education's AI-driven transformation.

However, fairness remains a pressing challenge. Taken together, these findings reinforce the call for layered governance that integrates technological, pedagogical, and ethical interventions to ensure responsible adoption of deepfake tutors. This requires not only technical safeguards but also policy clarity, faculty training, and institutional leadership to sustain trust. While our review and insights from questionnaires provide an integrative framework, future research must address current limitations, including small-scale empirical studies, short exposure durations, and limited perspectives beyond assistant professors. Our findings highlight trade-offs between innovation and governance. While stronger controls (e.g., biometric verification, multi-modal detectors) bolster integrity, they entail cost, complexity, and privacy considerations. Conversely, minimal governance can expedite adoption but risks eroding trust and exposing vulnerabilities. The proposed framework supports calibrated decision-making by matching control strength to risk tiers and institutional capability. While the framework offers useful insights, its applicability should be interpreted cautiously given the single-institution expert sample and English-only literature scope.

8. Limitations and Future Research

This study has several limitations that should be acknowledged when interpreting the findings. The systematic review and expert input were conducted exclusively in English and primarily represent perspectives from a single institution in Pakistan. These factors may limit the global and cross-cultural generalizability of the findings. The expert sample consisted of twelve assistant professors from a single institution, excluding perspectives from students, administrators, and policymakers. The qualitative design, based on written rather than oral interviews, may have constrained the depth of responses, limited opportunities for probing or clarification, and introduced response variability across participants. Moreover, the study reflects the state of generative AI up to early 2025, a rapidly evolving technological landscape that will require ongoing reassessment as detection methods and governance practices mature. Sociocultural attitudes toward deepfake tutors may also differ across regions, suggesting the need for localized adaptation of governance controls. Finally, raw qualitative data could not be publicly shared due to confidentiality constraints, though derived materials are available upon request. Future research ought to enhance stakeholder diversity, integrate longitudinal and cross-cultural methodologies, and empirically assess the proposed governance framework via pilot implementations to determine its robustness, equity, and relevance across diverse educational contexts.

9. Conclusions

This study reviewed 42 peer-reviewed publications and collected expert input from 12 assistant professors to explore the potential, risks, and governance needs of deepfake-

style AI tutors in higher education. The analysis revealed four central themes: personalization and engagement benefits, detection challenges and integrity risks, governance and policy gaps, and ethical and societal implications. Expert feedback highlighted the necessity of hybrid assessment models and fairness-aware safeguards, while the literature review highlighted the limited real-world validation of detection systems and underrepresentation of non-English perspectives.

Building on these findings, the study proposed a structured governance framework comprising four pillars: transparency and disclosure, data governance and privacy, integrity and detection, and ethical oversight and accountability. Accompanied by policy checklists, detection matrices, and institutional readiness measures, this framework provides practical guidance for implementing deepfake tutor systems responsibly.

As generative AI technologies continue to evolve, collaboration among educators, technologists, and policymakers will be critical. Transparent reporting of pilot results and open data sharing can promote continuous learning and adaptive governance across cultural and regulatory contexts. Ultimately, embedding these safeguards into institutional practice will allow higher education to benefit from the advantages of deepfake tutors while protecting integrity, privacy, and academic trust.

The proposed governance framework also reflects the principles of sustainability in education. By embedding transparency, fairness, and accountability into AI-driven instructional systems, it promotes enduring institutional trust and equitable opportunities for learners. In this way, the framework supports SDG 4 (Quality Education) and complements SDG 9 (Innovation and Infrastructure) and SDG 16 (Strong Institutions) by linking technological advancement with ethical resilience and responsible governance.

Author Contributions: Conceptualization, H.S., A.A. and A.A.N.; Methodology, H.S., A.A. and A.A.N.; Software, H.S., A.A. and A.A.N.; Validation, H.S., A.A. and A.A.N.; Formal analysis, H.S., A.A. and A.A.N.; Investigation, H.S., A.A. and A.A.N.; Resources, H.S., A.A. and A.A.N.; Data curation, H.S., A.A. and A.A.N.; Writing—original draft, H.S.; Writing—review & editing, A.A. and A.A.N.; Visualization, H.S., A.A. and A.A.N.; Supervision, A.A. and A.A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki (2013 revision) and approved by the Ethical Institutional Review Board, Lahore Garrison University (Approval Code: 2025-EIRB-0605-1; Approval Date: 6 May 2025).

Informed Consent Statement: All participants received an information sheet outlining the study's purpose and procedures, and they provided written informed consent prior to completing the questionnaire. Participation was voluntary and could be withdrawn at any time without penalty. Responses were pseudonymized (AP#1 to AP#12), and all identifying details were removed prior to analysis and reporting.

Data Availability Statement: Due to the sensitive nature of the qualitative questionnaire data and institutional confidentiality requirements, raw responses and signed consent forms cannot be made publicly available. These materials are safely stored on encrypted drives that only the research team can access. They will be kept for five years, as required by institutional policy. To support transparency, derived supporting materials, including the blank questionnaire, thematic coding framework, and summary tables, can be provided upon reasonable request to the corresponding author, subject to approval by the IRB and institutional data protection office. Aggregated findings are fully reported within the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Participant Demographics and Representative Coded Excerpts

Table A1. Participant Demographics.

Attribute	Category	*n*	%
Total Participants	Assistant Professors (all from Lahore Garrison University, Pakistan)	12	100%
Gender	Male	9	75%
	Female	3	25%
Region	South Asia (Pakistan)	12	100%
Discipline	Computer Science	3	25%
	Software Engineering	4	33%
	Education	3	25%
	Ethics	2	17%
Experience	Mean years of teaching/research	7	-
AI familiarity	At least some familiarity with AI-driven educational technologies	12	100%

The following anonymized excerpts illustrate participant perspectives. Full transcripts are not shareable due to confidentiality, but these representative quotes are included with permission:

Theme 1: Personalization and Engagement

"The avatar kept students engaged, especially in large online cohorts—it felt like having a real instructor present". (AP#3)

"Because it could switch languages on the fly, my international students didn't have to wait for translations". (AP#9)

Theme 2: Detection Challenges and Integrity Risks

"We need biometrics for high-stakes exams—detectors alone won't suffice". (AP#7)

"Even with state-of-the-art detectors, some fakes slip through, especially when students manipulate their webcams". (AP#11)

Theme 3: Governance and Policy Gaps

"Policy clarity and accountability are essential before any deployment—there's too much gray area right now". (AP#1)

"Our university has no guidelines on synthetic media, so everyone decides for themselves whether and how to use it". (AP#5)

Theme 4: Ethical and Societal Implications

"Students should know when an instructor is synthetic—transparency is non-negotiable". (AP#4)

"I worry about the long-term effects on trust: will students believe anything their instructors say if they know it could be generated?" (AP#8)

Appendix B. Questionnaire Guide (Summary)

The semi-structured interviews administered in writing via a standardized questionnaire comprised open-ended questions organized into five topics. While the complete responses cannot be shared due to confidentiality, the following represents the structure of the questionnaire:

- **Definitions and experiences:** How do you define a deepfake AI tutor? Have you encountered or used such systems in your teaching?
- **Perceived benefits:** In what ways could deepfake tutors enhance learning (e.g., personalization, accessibility, multilingual delivery)?
- **Risks and concerns:** What risks do you foresee (e.g., academic fraud, privacy, bias)? Which contexts raise the most concern?
- **Current policies and controls:** Are there existing institutional policies governing AI or deepfake use in education? What controls are in place?

- **Desired safeguards:** What safeguards or guidelines would you like to see implemented to ensure responsible deployment of deepfake tutors?

References

1. Sun, L.; Zhou, L. Does Generative Artificial Intelligence Improve the Academic Achievement of College Students? A Meta-Analysis. *J. Educ. Comput. Res.* **2024**, *62*, 1676–1713. [CrossRef]
2. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
3. Kamalov, F.; Calonge, D.S.; Smail, L.; Azizov, D.; Thadani, D.R.; Kwong, T.; Atif, A. Evolution of Ai in Education: Agentic Workflows. *arXiv* **2025**, arXiv:2504.20082. [CrossRef]
4. Wang, H.; Dang, A.; Wu, Z.; Mac, S. Generative AI in Higher Education: Seeing ChatGPT through Universities' Policies, Resources, and Guidelines. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100326. [CrossRef]
5. Rana, N.K. Generative AI and Academic Research: A Review of the Policies from Selected HEIs. *High. Educ. Future* **2025**, *12*, 97–113. [CrossRef]
6. Floridi, L.; Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
7. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. Opinion Paper: "So What If ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [CrossRef]
8. Vaccari, C.; Chadwick, A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Soc. Media Soc.* **2020**, *6*, 2056305120903408. [CrossRef]
9. Lim, L.-A.; Atif, A.; Heggart, K.; Sutton, N. In Search of Alignment between Learning Analytics and Learning Design: A Multiple Case Study in a Higher Education Institution. *Educ. Sci.* **2023**, *13*, 1114. [CrossRef]
10. Sharif, H.; Atif, A. The Evolving Classroom: How Learning Analytics Is Shaping the Future of Education and Feedback Mechanisms. *Educ. Sci.* **2024**, *14*, 176. [CrossRef]
11. Pinski, M.; Benlian, A. AI Literacy for Users—A Comprehensive Review and Future Research Directions of Learning Methods, Components, and Effects. *Comput. Human. Behav. Artif. Hum.* **2024**, *2*, 100062. [CrossRef]
12. Đerić, E.; Frank, D.; Milković, M. Trust in Generative AI Tools: A Comparative Study of Higher Education Students, Teachers, and Researchers. *Information* **2025**, *16*, 622. [CrossRef]
13. Dwivedi, Y.K.; Hughes, L.; Baabdullah, A.M.; Ribeiro-Navarrete, S.; Giannakis, M.; Al-Debei, M.M.; Dennehy, D.; Metri, B.; Buhalis, D.; Cheung, C.M.K.; et al. Metaverse beyond the Hype: Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy. *Int. J. Inf. Manag.* **2022**, *66*, 102542. [CrossRef]
14. Gupta, G.; Raja, K.; Gupta, M.; Jan, T.; Whiteside, S.T.; Prasad, M. A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. *Electronics* **2023**, *13*, 95. [CrossRef]
15. Henriksen, D.; Creely, E.; Gruber, N.; Leahy, S. Social-Emotional Learning and Generative AI: A Critical Literature Review and Framework for Teacher Education. *J. Teach. Educ.* **2025**, *76*, 312–328. [CrossRef]
16. Babaei, R.; Cheng, S.; Duan, R.; Zhao, S. Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis. *J. Sens. Actuator Netw.* **2025**, *14*, 17. [CrossRef]
17. Veletsianos, G.; Houlden, S. Radical Flexibility and Relationality as Responses to Education in Times of Crisis. *Postdigital Sci. Educ.* **2020**, *2*, 849–862. [CrossRef]
18. Gallifant, J.; Fiske, A.; Levites Strekalova, Y.A.; Osorio-Valencia, J.S.; Parke, R.; Mwavu, R.; Martinez, N.; Gichoya, J.W.; Ghassemi, M.; Demner-Fushman, D.; et al. Peer Review of GPT-4 Technical Report and Systems Card. *PLoS Digit. Health* **2024**, *3*, e0000417. [CrossRef]
19. Ma, H.; You, Q.; Jin, Z.; Liu, X.; Chen, Z. Exploring the Role of Generative AI in International Students' Sociocultural Adaptation: A Cognitive-Affective Model. *Front. Artif. Intell.* **2025**, *8*, 1615113. [CrossRef]
20. Waytz, A.; Heafner, J.; Epley, N. The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle. *J. Exp. Soc. Psychol.* **2014**, *52*, 113–117. [CrossRef]
21. Gstrein, O.J.; Haleem, N.; Zwitter, A. General-Purpose AI Regulation and the European Union AI Act. *Internet Policy Rev.* **2024**, *13*, 1–26. [CrossRef]
22. Miao, F. Holmes Wayne. In *Guidance for Generative AI in Education and Research*; UNESCO: Paris, France, 2023; ISBN 9789231006128.
23. Jobin, A.; Ienca, M.; Vayena, E. The Global Landscape of AI Ethics Guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
24. Chesney, R.; Citron, D.K. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electron. J.* **2019**, *107*, 1753. [CrossRef]

25. McFarland, M.; Hibbard, J. Are International Educators Ready for a Generative AI Future? Available online: https://www.nafsa.org/professional-resources/research-and-trends/trends-insights/are-international-educators-ready-generative-ai-future (accessed on 14 August 2025).
26. Roe, J.; Perkins, M.; Furze, L. Deepfakes and Higher Education: A Research Agenda and Scoping Review of Synthetic Media. *J. Univ. Teach. Learn. Pract.* **2024**, *21*, 1–22. [CrossRef]
27. Twomey, J.; Ching, D.; Peter Aylett, M.; Quayle, M.; Linehan, C.; Murphy, G. What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology. *IEEE Trans. Technol. Soc.* **2025**, *6*, 64–79. [CrossRef]
28. Roe, J.; Perkins, M.; Somoray, K.; Miller, D.; Furze, L. To Deepfake or Not to Deepfake: Higher Education Stakeholders' Perceptions and Intentions towards Synthetic Media. *arXiv* **2025**, arXiv:2502.18066. [CrossRef]
29. Létourneau, A.; Deslandes Martineau, M.; Charland, P.; Karran, J.A.; Boasen, J.; Léger, P.M. A Systematic Review of AI-Driven Intelligent Tutoring Systems (ITS) in K-12 Education. *NPJ Sci. Learn.* **2025**, *10*, 29. [CrossRef]
30. Fernández-Herrero, J. Evaluating Recent Advances in Affective Intelligent Tutoring Systems: A Scoping Review of Educational Impacts and Future Prospects. *Educ. Sci.* **2024**, *14*, 839. [CrossRef]
31. Wambsganss, T.; Benke, I.; Maedche, A.; Koedinger, K.; Käser, T. Evaluating the Impact of Learner Control and Interactivity in Conversational Tutoring Systems for Persuasive Writing. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 791–822. [CrossRef]
32. Pitts, G.; Motamedi, S. Understanding Human-AI Trust in Education. *arXiv* **2025**, arXiv:2506.09160. [CrossRef]
33. de Visser, E.J.; Monfort, S.S.; McKendrick, R.; Smith, M.A.B.; McKnight, P.E.; Krueger, F.; Parasuraman, R. Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *J. Exp. Psychol. Appl.* **2016**, *22*, 331–349. [CrossRef]
34. Li, Y.; Lyu, S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. *arXiv* **2019**, arXiv:1811.00656. [CrossRef]
35. Verdoliva, L. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process* **2020**, *14*, 910–932. [CrossRef]
36. Zhai, T.; Lu, K.; Li, J.; Wang, Y.; Zhang, W.; Yu, P.; Xia, Z. Learning Spatial-frequency Interaction for Generalizable Deepfake Detection. *IET Image Process* **2024**, *18*, 4666–4679. [CrossRef]
37. Hernandez-Ortega, J.; Tolosana, R.; Fierrez, J.; Morales, A. DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-Frame. In *Handbook of Digital Face Manipulation and Detection. Advances in Computer Vision and Pattern Recognition*; Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C., Eds.; Springer: Cham, Switzerland, 2022; pp. 255–273.
38. Heidari, A.; Jafari Navimipour, N.; Dag, H.; Unal, M. Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review. *WIREs Data Min. Knowl. Discov.* **2024**, *14*, e1520. [CrossRef]
39. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection. *Inf. Fusion.* **2020**, *64*, 131–148. [CrossRef]
40. de Lima, O.; Franklin, S.; Basu, S.; Karwoski, B.; George, A. Deepfake Detection Using Spatiotemporal Convolutional Networks. *arXiv* **2020**, arXiv:2006.14749. [CrossRef]
41. Ezeakunne, U.; Eze, C.; Liu, X. Data-Driven Fairness Generalization for Deepfake Detection. *arXiv* **2024**, arXiv:2412.16428.
42. Lin, L.; He, X.; Ju, Y.; Wang, X.; Ding, F.; Hu, S. Preserving Fairness Generalization in Deepfake Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
43. Drozdowski, P.; Rathgeb, C.; Dantcheva, A.; Damer, N.; Busch, C. Demographic Bias in Biometrics: A Survey on an Emerging Challenge. *IEEE Trans. Technol. Soc.* **2020**, *1*, 89–103. [CrossRef]
44. Raina, A.; Mann, G.; Professor, A. Exploring the Ethics of Deepfake Technology in Media: Implications for Trust and Information Integrity. *J. Inform. Educ. Res.* **2024**, *4*. [CrossRef]
45. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The Ethics of Algorithms: Mapping the Debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [CrossRef]
46. Meskys, E. Regulating Deep-Fakes: Legal and Ethical Considerations. *J. Intellect. Prop. Law Pract.* **2020**, *15*, 24–31. [CrossRef]
47. Mahashreshty Vishweshwar, S.; Vishweshwar, M. Implications of Deepfake Technology on Individual Privacy and Implications of Deepfake Technology on Individual Privacy and Security Security Recommended Citation Recommended Citation. 2023. Available online: https://repository.stcloudstate.edu/msia_etds/ (accessed on 12 August 2025).
48. Chan, C.K.Y.; Hu, W. Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 43. [CrossRef]
49. Netland, T.; von Dzengelevski, O.; Tesch, K.; Kwasnitschka, D. Comparing Human-Made and AI-Generated Teaching Videos: An Experimental Study on Learning Effects. *Comput. Educ.* **2025**, *224*, 105164. [CrossRef]
50. Xu, T.; Liu, Y.; Jin, Y.; Qu, Y.; Bai, J.; Zhang, W.; Zhou, Y. From Recorded to AI-generated Instructional Videos: A Comparison of Learning Performance and Experience. *Br. J. Educ. Technol.* **2025**, *56*, 1463–1487. [CrossRef]
51. Matsiola, M.; Lappas, G.; Yannacopoulou, A. Generative AI in Education: Assessing Usability, Ethical Implications, and Communication Effectiveness. *Societies* **2024**, *14*, 267. [CrossRef]

52. Gu, X.; Ericson, B.J. AI Literacy in K-12 and Higher Education in the Wake of Generative AI: An Integrative Review. In Proceedings of the 2025 ACM Conference on International Computing Education Research V.1, Charlottesville, VA, USA, 3–6 August 2025; ACM: New York, NY, USA, 2015; pp. 125–140.
53. Ng, D.T.K.; Leung, J.K.L.; Chu, S.K.W.; Qiao, M.S. Conceptualizing AI Literacy: An Exploratory Review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100041. [CrossRef]
54. Shi, J.; Liu, W.; Hu, K. Exploring How AI Literacy and Self-Regulated Learning Relate to Student Writing Performance and Well-Being in Generative AI-Supported Higher Education. *Behav. Sci.* **2025**, *15*, 705. [CrossRef]
55. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
56. Available online: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf (accessed on 6 August 2025).
57. Braun, V.; Clarke, V. Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
58. UNESCO. *Guidance for Generative AI in Education and Research*; UNESCO: Paris, France, 2023. [CrossRef]
59. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 16 August 2025).
60. Available online: https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/bd/bd2223a/23bd030#:~:text=The%20purpose%20of%20the%20Privacy%20Legislation%20Amendment,and%20Media%20Authority%20Act%202005%20(ACMA%20Act) (accessed on 16 August 2025).
61. Abbas, F.; Taeihagh, A. Unmasking Deepfakes: A Systematic Review of Deepfake Detection and Generation Techniques Using Artificial Intelligence. *Expert. Syst. Appl.* **2024**, *252*, 124260. [CrossRef]
62. Dasari, A.; Prakash, S.K.A.; Jeni, L.A.; Tucker, C.S. Evaluation of Biases in Remote Photoplethysmography Methods. *NPJ Digit. Med.* **2021**, *4*, 91. [CrossRef] [PubMed]
63. Ayyoub, A.M.; Khlaif, Z.N.; Shamali, M.; Abu Eideh, B.; Assali, A.; Hattab, M.K.; Barham, K.A.; Bsharat, T.R.K. Advancing Higher Education with GenAI: Factors Influencing Educator AI Literacy. *Front. Educ.* **2025**, *10*, 721. [CrossRef]
64. Zhang, Y.; Reusch, P. Trust in and Adoption of Generative AI in University Education: Opportunities, Challenges, and Implications. In Proceedings of the 2025 IEEE Global Engineering Education Conference (EDUCON), London, UK, 22–25 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–10.
65. Luo, J. How Does GenAI Affect Trust in Teacher-Student Relationships? Insights from Students' Assessment Experiences. *Teach. High. Educ.* **2025**, *30*, 991–1006. [CrossRef]