

Chapter 1

The Illusion of Security: AI-Powered Cybersecurity, Data Privacy, and Human-Centered Design Failures in Education

Asegul Hulus

 <http://orcid.org/0000-0001-8797-450X>

Independent Researcher, Cyprus

Manuel B Garcia

 <http://orcid.org/0000-0003-2615-422X>

FEU Institute of Technology, Manila, Philippines

C. Kishor Kumar Reddy

Stanley College of Engineering and Technology for Women, India

Shugufta Fatima

 <http://orcid.org/0009-0005-4396-4947>

Stanley College Of Engineering and Technology For Women, India

Ari Happonen

 <http://orcid.org/0000-0003-0744-1776>

Lappeenranta-Lahti University of Technology, Lappeenranta, Finland

ABSTRACT

This chapter critically examines the assumption that artificial intelligence (AI)-powered cybersecurity inherently strengthens data protection within educational environments. Drawing on a critical review of interdisciplinary literature, it argues that the rapid adoption of AI-driven educational technologies has outpaced the development of effective governance, resulting in significant challenges relating

DOI: 10.4018/979-8-2600-2518-5.ch001

Copyright © 2027, IGI Global Scientific Publishing. Copying or distributing in print or electronic forms without written permission of IGI Global Scientific Publishing is prohibited. Use of this chapter to train generative artificial intelligence (AI) technologies is expressly prohibited. The publisher reserves all rights to license its use for generative AI training and machine learning model development.

to data privacy, transparency, provenance, informed consent, human-centered design, and cybersecurity. Rather than viewing these issues as isolated concerns, the chapter demonstrates how they interact to create an illusion of security, whereby systems marketed as protective may simultaneously introduce new vulnerabilities and accountability gaps. It concludes that addressing these challenges requires an integrated approach that combines cybersecurity, human-computer interaction, AI governance, and educational policy to develop more transparent, ethical, and user-centered AI-mediated learning environments.

1. INTRODUCTION: SCOPE OF THE VOLUME

The incorporation of artificial intelligence (AI) within educational environments has resulted in the unification of previously separate fields: digital pedagogy, cybersecurity, human-computer interaction, and user experience design (Islam et al., 2026). This convergence is neither incidental nor purely technical. This indicates a significant shift in learner-institution engagement, institutional data practices, and the establishment of trust amidst complex, often inscrutable systems (Hummel, 2025).

AI-powered tools have become embedded across the pedagogical lifecycle, from adaptive learning platforms to automated assessment systems, conversational tutoring agents, remote proctoring software, and administrative decision-support tools (Andrija Bernik et al., 2025). Their expeditious implementation has preceded the establishment of commensurate protective protocols, necessitating that students, educators, and institutions manage a landscape where the aspiration for superior learning achievements is frequently juxtaposed with uninvestigated repercussions for privacy, security, and autonomy (Islam et al., 2026). This asymmetry between adoption speed and safeguard development is not unique to education, but its consequences are arguably more acute here, given that the population most affected, students, frequently has limited bargaining power, limited technical literacy regarding data practices, and limited alternative means of accessing the educational services in question (Andrija Bernik et al., 2025).

The scale of this adoption bears emphasis. Institutions at all levels, from primary and secondary schools to universities and continuing professional education programs, have integrated AI powered platforms into core instructional and administrative functions within a compressed timeframe (Nasni Naseri & Abdullah, 2024).

This compression is significant due to the frequent curtailment or complete omission of standard institutional processes for evaluating new technologies. These processes, such as pilot programs, procurement reviews, and staff training, are often circumvented owing to competitive and budgetary imperatives for rapid adoption (Alotaibi, 2026). This results in an educational technology ecosystem character-

ized by deployment rates exceeding critical evaluation rates, leading to a structural situation where potential risks are identified, if at all, solely after a tool has been incorporated into daily institutional practice (Islam et al., 2026).

This volume, and this introductory chapter in particular, positions itself within that unresolved tension. Rather than treating AI-powered cybersecurity as a neutral or self-evidently protective apparatus, this chapter interrogates the premise that securing data privacy in education is a solved or solvable problem simply by virtue of deploying more AI. This premise, while rarely stated explicitly by vendors or institutions, is nonetheless embedded implicitly in marketing language that positions AI-powered security tools as a comprehensive solution to a complex, multidimensional problem. The critical task of this chapter, and of the volume it introduces, is to unsettle that implicit premise by examining where the apparatus positioned as protective in fact contributes to the harms it claims to prevent.

The scope of this volume is accordingly broad by necessity: it spans the technical (cybersecurity threats and defenses), the interactional (human-computer interaction and user experience), the pedagogical (digital learning design), and the ethical and governance dimensions that bind these domains together. This breadth reflects a deliberate editorial position. Much of the existing literature addressing AI in education treats cybersecurity, privacy, and design as separable concerns, each the province of a distinct research community operating with its own methods, venues, and vocabulary. This volume proceeds from the position that such separation, however useful for disciplinary specialization, obscures the ways in which these domains interact in practice and, in doing so, obscures the compounding effect of failures that occur at their intersections.

The intended readership of this volume reflects that breadth. It is written for institutional policymakers responsible for procurement and data governance decisions, for human-computer interaction and user experience designers building or evaluating educational tools, for cybersecurity practitioners operating within resource-constrained educational information technology environments, and for educators and administrators who mediate between these systems and the students who use them daily.

Each of these audiences approaches the problem from a different vantage point, and part of the volume's purpose is to make each vantage point legible to the others. A cybersecurity practitioner may be well versed in threat modeling but unfamiliar with the human-computer interaction literature on informed consent; a curriculum designer may understand pedagogical objectives thoroughly while having little visibility into the data flows underlying the tools deployed to achieve them. This volume treats that cross-disciplinary translation as itself a contribution, not merely a byproduct of covering a broad topic.

The regulatory environment surrounding these developments further complicates the picture this volume addresses. Data protection frameworks vary considerably across jurisdictions in the degree to which they anticipate and regulate AI-specific data practices such as the NIS2 Directive (European Parliament & Council of the European Union, 2022), with some regions maintaining comprehensive statutory protections for student data while others rely on sector-specific guidance that has not been substantially updated to account for the categories of data collection described throughout this chapter. Educational institutions operating internationally, or serving increasingly mobile and international student populations, must accordingly navigate a fragmented regulatory landscape in which compliance with the requirements of one jurisdiction offers no assurance of compliance with another, a condition that further complicates institutional efforts to establish consistent data governance practices across the full range of AI-powered tools in use. This regulatory fragmentation is not the central focus of this volume, but it forms an important part of the backdrop against which the arguments advanced in the chapters that follow should be read, since many of the design and governance failures identified in this chapter persist in part because no single regulatory authority has clear jurisdiction to compel their correction.

The remainder of this chapter proceeds as follows. It first situates its own methodological approach as a general critical review, guided by a specific research question. It then examines, in turn, the data practices, trust deficits, provenance failures, consent ambiguities, and escalating threat landscape that characterize AI-powered education today. It closes by returning to the research question directly, before articulating the rationale and contribution of the volume as a whole.

2. METHODOLOGICAL NOTE

This chapter utilizes a broad critical review methodology, deviating from a systematic review or a formally protocolized approach like PRISMA (Page et al., 2021). It does not employ reflexive thematic analysis, nor does it seek to exhaustively map the totality of literature addressing AI, cybersecurity, and education. Instead, existing scholarship is drawn upon selectively and purposefully to substantiate and supplement a critical argument regarding the relationship between AI-powered cybersecurity and data privacy failures in educational settings (Braun & Clarke, 2006).

This methodological choice reflects the function of the chapter within the volume. As an introductory chapter rather than a standalone empirical study, its purpose is to frame the terrain that subsequent contributions will examine in greater depth, rather than to report a replicable review process. The critical review format permits synthesis across disparate bodies of literature, including human-computer

interaction, cybersecurity studies, educational technology, and information ethics, that a narrower systematic protocol would be less suited to accommodate (Snyder, 2019). Where a systematic review prioritizes exhaustiveness and reproducibility of search strategy, a critical review of this kind prioritizes the construction of a coherent argument across domains that rarely cite one another despite addressing overlapping populations and systems (Page et al., 2021).

This is a meaningful methodological distinction rather than a mere formality. Systematic review protocols, including those guided by Preferred Reporting Items for Systematic Reviews and Meta-Analyses, are designed for questions that can be operationalized into a bounded search strategy with clearly specified inclusion and exclusion criteria, typically within a single, relatively coherent body of literature (Snyder, 2019). The argument advanced in this chapter, by contrast, requires drawing connections across literatures that are not indexed under shared terminology, do not cite a common set of foundational sources, and are frequently published in venues that do not overlap. A systematic search strategy calibrated to any one of these literatures would necessarily exclude relevant material from the others, undermining rather than serving the chapter's synthetic purpose.

The chapter is guided by the following research question:

How does the AI-powered cybersecurity apparatus in education, marketed as protective, itself contribute to the erosion of data privacy, provenance, consent, and human-centered design in AI-mediated learning environments?

This question is deliberately framed to resist a purely technical answer. It does not ask merely whether AI-powered cybersecurity tools are effective at their stated function, a question that could in principle be answered through conventional security evaluation metrics such as detection rates or response times. Instead, it asks a broader question about the systemic relationship between the deployment of these tools and the erosion of the very protections they are marketed to provide, a question that necessarily draws on human-computer interaction, governance, and ethical analysis alongside technical security literature (Islam et al., 2026).

This question is addressed cumulatively across the sections that follow, each of which examines a distinct dimension of the erosion in question, before being answered directly in a dedicated section preceding the volume's rationale and contribution. Section 3 examines the scope and mechanisms of data collection. Section 4 examines the resulting trust and transparency deficit. Section 5 examines the specific failure of provenance and origin-tracking. Section 6 examines the unresolved status of consent and proprietary rights. Section 7 examines the cybersecurity threat landscape that this data-rich, trust-deficient environment produces. Section 8 then synthesizes these threads into a direct answer to the research question, and Section 9 situates that answer within the broader rationale and contribution of the volume.

3. THE DATA STUDENTS DON'T KNOW THEY'RE GIVING AWAY

The proliferation of AI-powered platforms within educational settings has been accompanied by a substantial expansion in the volume, granularity, and sensitivity of student data collected, frequently without students possessing a complete understanding of what is gathered or how it is subsequently used (Wang et al., 2024). This data spans personal identifiers, academic performance records, biometric markers, and institutional information, extending into textual submissions, behavioral interaction patterns, voice recordings, and images (Islam et al., 2026). Each of these categories carries distinct implications for privacy, yet they are frequently aggregated within the same platforms and subjected to the same, often minimal, disclosure practices.

The specific mechanisms through which this collection occurs are worth naming concretely. Remote proctoring software, deployed widely during and after periods of expanded online assessment, routinely captures webcam footage, eye-movement and gaze data, keystroke patterns, and in some implementations, scans of a student's physical environment, all under the framing of academic integrity enforcement (Malhotra & Chhabra, 2026). Writing assistance tools retain not only final submissions but iterative draft histories, capturing the evolution of a student's thinking in a form that extends well beyond what a conventional assignment submission would reveal (Wang, 2024). Adaptive learning platforms log click-level behavioral data, response latency, and navigation patterns, generating a granular behavioral profile of each student that persists across sessions and, in many cases, across academic terms (Arizmendi et al., 2022). Conversational tutoring agents, meanwhile, retain the full content of student queries, which frequently extend beyond academic content into disclosures of stress, confusion, or personal circumstance offered in the course of seeking help, data of a qualitatively different sensitivity than a completed assignment (Barzanji & Loitsch, 2025). None of these practices are necessarily concealed in a legal sense, since they are typically disclosed somewhere within platform terms of service, yet the practical effect on student awareness is functionally similar to concealment.

The consequences of this aggregation extend beyond the immediate collection event. Information collected through routine educational engagements may be reutilized for the creation of behavioral profiles, employed in the development of predictive models for student outcomes or conduct, or preserved indefinitely without a well-defined policy for its removal (Arizmendi et al., 2022). Surveillance in this context need not be overt, it can operate through the routine, ambient logging of interaction data that students reasonably assume is transient rather than permanent (Islam et al., 2026). The ongoing practice of ambient logging provides institutional and commercial entities with a comprehensive view of student conduct. This level of insight, which would typically necessitate explicit and detailed justification in other

settings, is often accepted as a standard aspect of educational technology platforms without significant scrutiny (Malhotra & Chhabra, 2026).

3.1. Interface Design as the Mechanism of Concealment

A substantial amount of this data acquisition is facilitated not by outright deception, but by the strategic design of interfaces and interactions that obscure the process of data harvesting from user awareness (Rossi et al., 2024). Consent mechanisms within AI-driven educational platforms often manifest as protracted, legally complex agreements, which students are structurally disincentivized from fully reviewing, a phenomenon extensively documented in human-computer interaction research concerning informed consent (Wiberg, 2026). When interfaces incorporate deceptive design practices, known as dark patterns, which guide users toward divulging information or agreeing to terms through obstacles, predetermined options, or unclear language, the resulting consent is not truly informed (Rossi et al., 2024).

According to Qiang and Liu (2025) common manifestations of this pattern in educational technology include default-enabled data-sharing settings that require active navigation to disable, consent interfaces that bundle unrelated permissions into a single acceptance action, and onboarding flows that present privacy disclosures only after a student has already invested time in account setup, thereby increasing the psychological cost of declining . A related pattern involves the framing of consent as a binary, all-or-nothing choice between full platform access and no access at all, foreclosing the more granular consent options that would allow a student to participate in a course while declining specific categories of data collection unrelated to core instructional function. This binary framing is particularly consequential in educational contexts because declining is frequently not a genuine option: a student required to use a specific platform as a condition of course participation has no practical alternative to acceptance, regardless of how the consent interface is designed (Dadson et al., 2025).

This is not merely a peripheral usability concern. It constitutes a central mechanism by which the data practices described above are able to proceed with minimal resistance or scrutiny from the very population most affected by them. The opacity of data-collection interfaces therefore functions as a precondition for the erosion of privacy described throughout this chapter, rather than an incidental byproduct of otherwise sound data governance (Qiang and Liu, 2025).

3.2. The Compounding Effect of Data Aggregation Across Platforms

A further dimension of this problem, distinct from the operation of any single platform in isolation, concerns the compounding effect of data aggregation across the multiple AI-powered tools a single student typically encounters over the course of an academic program (Vieriu & Petrea, 2025). A student enrolled in a contemporary degree program may interact with a proctoring platform, a writing assistance tool, an adaptive learning system, and a conversational tutoring agent within a single academic term, each operated by a different commercial vendor with its own data retention and processing practices (Klimova & Pikhart, 2025)

While each individual platform's data collection may appear proportionate when evaluated in isolation, the aggregate profile constructed across these overlapping systems is considerably more revealing than any single data source alone, a pattern well established in the broader privacy literature concerning the mosaic effect of combined data sources (Vieriu & Petrea, 2025). No institutional mechanism typically exists to evaluate this aggregate exposure, since data governance review, where it occurs at all, tends to proceed on a platform-by-platform basis rather than assessing the cumulative footprint a student accumulates across the full suite of tools an institution deploys (Klimova & Pikhart, 2025). This gap between individual platform review and aggregate exposure assessment represents a structural blind spot in current institutional data governance practice.

4. TRUST, TRANSPARENCY, AND THE HUMAN-CENTERED DESIGN GAP

Trust between students and the AI-powered systems they are required or encouraged to use is not a static condition but an ongoing negotiation shaped by the degree to which those systems make their operations legible to the people affected by them. Within AI-powered educational technology, this negotiation is frequently undermined from the outset. Students are often unable to determine what data a given platform collects, how that data informs the outputs they receive, or what recourse exists when a system's behavior appears inconsistent or unjust (Martín-Moncunill & Alonso Martínez, 2025; Passlack et al., 2025).

Transparency deficits of this kind are compounded by the black-box character of many underlying AI models, whose decision processes remain opaque even to the institutions deploying them, let alone to the students subject to their outputs. This opacity has concrete consequences in practice. A student who receives an AI-generated grade, an academic integrity flag, or a risk score used for early-intervention

advising is frequently given no accessible explanation of the reasoning underlying that determination, only the determination itself. Where an appeals process exists, it typically requires the student to contest an outcome without access to the evidentiary basis on which it was reached, placing the burden of proof on the party with the least information (Dadson et al., 2025; Wang et al., 2024).

Where explainability mechanisms exist, they are frequently retrofitted rather than designed in from the outset, and their effectiveness in genuinely informing user understanding, as opposed to merely satisfying a procedural requirement, remains contested within the human-computer interaction literature (Martín-Moncunill & Alonso Martínez, 2025; Passlack et al., 2025).. An explanation interface that presents a numerical confidence score or a generic statement of contributing factors, without translating that information into terms a non-specialist student can act upon, satisfies the letter of transparency while failing its underlying purpose (Wiberg, 2026). This distinction, between transparency as a formal property of a system and transparency as an experienced, actionable quality for the end user, recurs throughout the remainder of this chapter and constitutes one of its central analytical contributions.

4.1. The Asymmetry of Institutional and Student Technical Literacy

A further dimension of the trust deficit concerns the asymmetry in technical literacy between the institutions procuring AI-powered platforms and the students required to use them. Institutional procurement decisions are typically made by administrators who, while responsible for evaluating vendor claims regarding security and data handling, may themselves lack the specialized technical expertise necessary to critically assess those claims, relying instead on vendor-supplied documentation and marketing materials (Hulus, 2026). Students, positioned even further from the procurement process, have essentially no visibility into it at all, encountering the resulting platform only as a fixed condition of course participation, with no meaningful input into its selection or configuration (Alotaibi, 2026; Nasni Naseri & Abdullah, 2024).

This asymmetry produces a governance structure in which the population with the least information and the least power, students, bears a disproportionate share of the resulting risk, while the population with greater information and negotiating leverage, institutional administrators, is insulated from the direct consequences of an inadequate evaluation. Addressing this asymmetry would require not merely improved disclosure to students after a procurement decision has been made, but meaningful student and educator input into the procurement process itself, a practice that remains rare within current institutional norms (Dadson et al., 2025).

4.2. Accessibility and Human-Centered Design as Underaddressed Dimensions

The human-centered design gap in AI-powered education extends beyond transparency into accessibility more broadly. Students with disabilities, students using assistive technologies, and students from linguistic or cultural backgrounds not well represented in the training data underlying many AI systems frequently encounter interfaces and outputs that were not designed with their needs in mind. This is a design failure with material consequences, not merely an inconvenience, since it can determine whether a student is able to participate meaningfully in a learning activity at all (Wiberg, 2026).

Concrete instances of this failure include automated proctoring systems whose facial detection algorithms perform inconsistently across different skin tones and lighting conditions, generating disproportionate integrity flags for some student populations relative to others, and natural language processing systems whose outputs are calibrated primarily to dominant dialects, producing degraded performance for students whose speech or writing patterns diverge from that norm (O'Malley et al., 2024). Voice-based interaction systems similarly present accessibility barriers for students with speech differences, while visually dense analytics dashboards intended to communicate performance data to students may be poorly suited to students using screen readers or other assistive technologies (Wiberg, 2026).

Human-centered design, understood as an approach that positions the needs, capabilities, and contexts of users at the center of the design process, offers a corrective framework against which current AI-powered educational tools can be evaluated and largely found wanting (Wang et al., 2024). The persistence of this gap, despite the maturity of human-centered design as a discipline, suggests that its absence in AI-powered education is not attributable to a lack of available methodology, but to insufficient prioritization of user-centered concerns relative to functional or commercial objectives during development. Where human-centered design methodology is applied in the development of educational technology, it is more frequently directed toward optimizing engagement metrics or reducing support costs than toward surfacing and addressing the accessibility and equity concerns described above, a divergence between the availability of a methodology and its actual application in practice (Wiberg, 2026).

When accessibility failures of this kind are identified, whether through student complaint, formal accommodation request, or external audit, the remediation pathway available to affected students is frequently slow and burdensome relative to the immediate academic consequences they face (Wang et al., 2024). A student disadvantaged by a proctoring system's inconsistent facial detection, for instance, may need to pursue a formal accommodation process running in parallel with the

academic term in which the disadvantage occurred, meaning that even a successful remediation outcome frequently arrives too late to address the specific instance of harm that prompted it. This temporal mismatch between the pace of harm and the pace of institutional remediation is itself a human-centered design failure, since a genuinely user-centered accommodation process would be designed to operate on a timescale commensurate with the academic stakes involved, rather than treating accessibility remediation as a background administrative process disconnected from the immediate context in which the harm occurs (O'Malley et al., 2024).

5. THE PROVENANCE PROBLEM

AI-generated content occupies an increasingly ambiguous position within educational settings, in large part because no standardized mechanism exists for establishing the origin, timing, or platform of a given output. This absence of reliable provenance information leaves institutions without dependable means of ascertaining whether a piece of student work was generated wholly, partially, or not at all by an AI system, and if so, through which tool and at what point in the drafting process (Garzón et al., 2025).

The consequences of this gap extend directly into questions of academic integrity. Detection tools marketed as capable of identifying AI-generated text have demonstrated inconsistent reliability, producing both false positives that wrongly implicate students and false negatives that fail to flag genuinely generated content (Ali et al., 2024). The consequences of a false positive are particularly severe in the educational context, given the disciplinary and reputational weight typically attached to an academic integrity finding, while the consequences of a false negative undermine the very academic integrity function the detection tool was procured to serve. In the absence of standardized timestamping or origin-tracking protocols, institutions are left to adjudicate integrity cases using tools whose evidentiary basis is itself contested, a circumstance that undermines confidence in the fairness and consistency of academic integrity processes more broadly.

This unreliability is compounded by the rapid evolution of the underlying generative systems themselves. Detection tools are typically trained to recognize patterns characteristic of a particular generation of AI models, and their accuracy degrades as newer models produce output with different statistical characteristics, requiring continual retraining that is not always kept current with the pace of underlying model development (Garzón et al., 2025). This creates a persistent lag between the capabilities of generative systems available to students and the capabilities of detection systems available to institutions, a lag that structurally favors whichever party has more timely access to the newest tools.

5.1. Consequences for Academic Integrity and Institutional Accountability

Institutional accountability is implicated by this provenance gap in a further respect. Where an institution cannot reliably establish how a given AI-powered tool arrived at a particular output, whether an assessment, a recommendation, or a disciplinary flag, it becomes correspondingly difficult for that institution to take responsibility for the consequences of that output, or to offer students a meaningful avenue for contesting it (Yan et al., 2025). This diffusion of accountability, in which neither the platform vendor nor the deploying institution can fully account for a given outcome, represents a governance failure with direct implications for student due process.

The due process implications extend beyond the individual case. Where institutional integrity policies rely on detection tools whose reliability cannot be independently verified, the resulting body of disciplinary decisions accumulates without a clear evidentiary foundation, creating institutional exposure that may only become apparent when a sufficiently large number of contested cases prompts broader scrutiny of the underlying detection methodology (Ali et al., 2024). Some institutions have responded to this uncertainty by declining to rely on automated detection as a sole basis for disciplinary action, instead treating a positive detection result as one input among several requiring corroboration, a more cautious approach that nonetheless does not resolve the underlying provenance gap, only mitigates its most severe consequences (Garzón et al., 2025). Even this more cautious corroborative approach depends on the availability of human evaluators with the time and training to conduct a meaningful secondary review, a resource-intensive requirement that not all institutions are equally positioned to meet, meaning that the practical rigor of academic integrity adjudication in the absence of reliable provenance information may vary considerably across institutions with differing levels of administrative capacity (Ali et al., 2024).

5.2. Interface Transparency

The provenance problem is not solely a backend or technical failure; it is also an interface failure. Even where some provenance information is technically available, whether embedded metadata, generation logs, or model version identifiers, this information is rarely surfaced to students or educators in a form that is accessible, comprehensible, or actionable (Yan et al., 2025). A system that technically retains provenance data but does not present it meaningfully to its users cannot be said to have solved the provenance problem from a human-centered design standpoint, even if it has done so at a purely technical level (Ali et al., 2024). This distinction between

technical availability and interactional accessibility of provenance information is a recurring pattern throughout this chapter's broader argument regarding the gap between AI-powered security claims and their user-facing reality (Yan et al., 2025).

Proposed technical solutions to the provenance problem, including cryptographic watermarking of generated content and standardized metadata schemas for recording generation history, remain in relatively early stages of adoption and face significant practical obstacles, including the ease with which watermarks can be stripped or altered and the absence of industry-wide agreement on a common metadata standard (Garzón et al., 2025). Even where such solutions mature technically, their value to students and institutions will depend entirely on whether the resulting provenance information is surfaced through interfaces genuinely designed for comprehension by non-specialist users, rather than treated as a compliance artifact accessible only to those with the technical background to interpret it.

6. CONSENT AND PROPRIETARY RIGHTS IN AI-ASSISTED WORK

The question of consent in AI-mediated education is complicated by the involvement of multiple parties whose interests do not necessarily align. Where a student generates work with the assistance of an AI system, the resulting output may implicate proprietary interests belonging not to the student or the institution, but to the commercial entity operating the underlying model (Dadson et al., 2025). The terms governing many AI platforms grant the operating company rights to retain, process, and in some cases further train models on submitted content, a condition to which students frequently agree without full awareness, given the interface and disclosure patterns discussed earlier in this chapter.

This arrangement raises a distinct possibility that student intellectual effort, rendered as prompts, drafts, or iterative refinements submitted to an AI system, may unintentionally contribute to the commercial value or training data of that system, without corresponding compensation, acknowledgment, or even awareness on the part of the student (Zdravkova, 2026). Existing intellectual property frameworks were not designed with this configuration in mind, and the resulting ambiguity leaves both students and institutions with limited clarity regarding who holds rights to AI-assisted academic work. This ambiguity is particularly acute for graduate students and early-career researchers, whose academic work product may carry independent commercial or publication value, and for whom the unintentional transfer of proprietary interest through routine platform use carries consequences extending well beyond a single course assignment (Dadson et al., 2025).

6.1. Institutional Versus Individual Consent

A further complication arises from the asymmetry between individual and institutional consent. Many AI-powered platforms are procured at the institutional level, with terms of service negotiated or accepted by an administrator on behalf of an entire cohort, department, or student body (Zdravkova, 2026). The individual student, whose data is subsequently collected under this arrangement, is rarely a party to that negotiation and frequently has no meaningful opportunity to decline participation short of withdrawing from the course or program in which the tool is used (Dadson et al., 2025).

This institutional consent, while procedurally valid in a contractual sense, does not substitute for the individually informed consent that data protection frameworks in other sectors typically require, producing a structural gap between what is legally authorized and what is genuinely consented to by the individuals whose data is at stake (Zdravkova, 2026). This gap is further complicated where minors are involved, since institutional consent frameworks for younger students typically route through parental or guardian consent mechanisms that carry their own separate set of disclosure and comprehension challenges, layering an additional intermediary between the platform's data practices and the individual whose data is actually being processed.

6.2. Third-Party Materials and Unresolved Consent Boundaries

A related but distinct concern arises where students submit to AI systems materials that are not solely their own, including peer-authored work, institutional documents, and proprietary course content. In such cases, the student submitting the material is not the sole rights holder, yet the consent obtained by the platform, where obtained at all, typically extends only to the submitting individual rather than to the originators of the embedded material (Go et al., 2026).

This creates a permissible scope problem: the boundaries of what a student may legitimately submit to a third-party AI system remain largely undefined within institutional policy, leaving students to make consequential decisions about the data and intellectual property of others without clear guidance (Wieczorek et al., 2025). Group projects present a particularly clear instance of this problem, since a single student submitting shared work to an AI platform for assistance effectively makes a consent decision on behalf of collaborators who may be entirely unaware that the submission occurred (Sakka et al., 2026). Instructional materials present a related but distinct instance, where a student submitting a proprietary reading, dataset, or assessment instrument to an AI system for assistance may inadvertently expose institutional or instructor intellectual property to a third-party commercial platform, potentially in violation of licensing terms the student has no visibility into.

The absence of institutional frameworks addressing this gap represents a further instance of the human-centered design failure identified throughout this chapter, insofar as students are left to navigate a consequential decision space without adequate support, information, or institutional scaffolding (Sakka et al., 2026). Few institutions have developed clear, accessible policy guidance addressing what categories of third-party material are and are not appropriate to submit to AI platforms, leaving individual students to make case-by-case judgments with limited institutional support (Go et al., 2026)

Where institutional policy does exist, it is frequently framed in terms of academic integrity, addressing what a student may or may not submit to an AI system as part of completing an assignment, rather than in terms of data protection or third-party consent, addressing what a student may or may not disclose about others in the course of doing so (Sakka et al., 2026). This framing gap means that a student may be in full compliance with an institution's academic integrity policy while simultaneously exposing peer-authored or institutional material to a commercial platform in a manner the institution has never explicitly evaluated or authorized (Go et al., 2026). Closing this gap would require institutions to treat third-party data exposure through AI platforms as a distinct policy question from academic integrity, developed in consultation with data protection and legal expertise rather than left to academic integrity offices whose institutional mandate and expertise lie elsewhere.

7. THE ESCALATING CYBERSECURITY THREAT LANDSCAPE IN EDUCATION

The education sector occupies a persistent and disproportionate position within the global cybersecurity threat landscape, enduring sustained exposure to malware, phishing campaigns, ransomware, and data breaches. This vulnerability is not incidental. Educational institutions typically manage large volumes of sensitive personal and financial data while operating with security infrastructure and staffing that frequently lag behind sectors such as finance or healthcare, producing a threat environment in which attackers face comparatively low resistance relative to the value of the data available (Islam et al., 2026; Garzón et al., 2025).

This structural vulnerability is compounded by the decentralized governance typical of many educational institutions, particularly larger universities, in which individual departments, research centers, and even individual faculty members may procure and deploy AI-powered tools independently of centralized information technology oversight (Alotaibi, 2026). This decentralization, often motivated by legitimate pedagogical or research flexibility, has the secondary effect of multiplying the number of potential attack surfaces an institution must defend, frequently

without a complete institutional inventory of which tools are in active use across the organization.

7.1. Malware, Phishing, Ransomware, and Breaches as a Persistent Pattern

Ransomware attacks against educational institutions have grown in both frequency and severity, disrupting instructional continuity and, in some cases, compromising sensitive student and staff records on a large scale (Islam et al., 2026). The financial cost of such incidents extends well beyond any ransom payment itself, encompassing incident response, system remediation, regulatory reporting obligations, and reputational harm that can affect enrollment and institutional partnerships for years following an incident. Beyond the financial dimension, ransomware incidents in educational settings carry a distinct instructional cost, since the disruption of core systems, including learning management platforms, grading systems, and communication infrastructure, can halt instructional activity entirely for extended periods, with cascading effects on academic calendars and student progress that are difficult to fully remediate after the fact (Go et al., 2026).

Phishing campaigns targeting educational communities have similarly evolved in sophistication, exploiting the high volume of routine institutional communication, such as enrollment notices, financial aid correspondence, and administrative announcements, to disguise malicious intent within plausible, expected contexts (Islam et al., 2026). The recurrence of these attack types across successive years suggests a structural vulnerability rather than an isolated set of incidents. Data breaches affecting educational institutions frequently expose not only the standard categories of personal and financial information common across sectors, but also the additional categories of sensitive data discussed in Section 3 of this chapter, including biometric and behavioral data collected through AI-powered platforms, expanding the scope of harm beyond what a conventional data breach in a non-educational context would typically involve (Go et al., 2026).

The persistence of this threat pattern is further reinforced by a chronic staffing and resourcing gap within educational information technology security functions relative to the scale of the infrastructure they are responsible for defending. Many educational institutions, particularly those operating with constrained public funding, maintain information security teams that are small relative to the number of systems, platforms, and endpoints under their purview, a mismatch that widens further as the number of AI-powered tools deployed across an institution grows without a corresponding increase in security staffing (Islam et al., 2026). This resourcing gap is not merely a matter of headcount but of specialized expertise, since effective defense against the evolving threat vectors described in this section increasingly

requires familiarity with AI-specific attack methods that fall outside the traditional training and certification pathways available to educational information technology security professionals (Go et al., 2026).

7.2. AI-Driven Social Engineering as an Underreported Threat Vector

A more recent and comparatively undertheorized threat vector involves the use of AI-powered tools to conduct social engineering at scale. Such tools are capable of convincingly mimicking legitimate institutional services, offering financial, administrative, or academic assistance as pretext for illicitly obtaining credentials or account access. Unlike earlier generations of phishing, which often relied on generic or poorly localized messaging that alert users could identify, AI-generated social engineering content can be tailored with a degree of linguistic and contextual precision that substantially raises the difficulty of detection (Yan et al., 2025).

This precision extends to impersonation of specific institutional communication styles, administrative offices, or even individual staff members, generated from publicly available institutional content such as official communications, websites, and social media, allowing attackers to construct highly convincing pretexts with minimal manual effort (Go et al., 2026). Voice synthesis technology introduces a further escalation, enabling social engineering attacks conducted through voice calls that convincingly mimic a real staff member or administrator, a threat vector for which most educational institutions have not yet developed corresponding verification protocols. Despite this escalation, this risk category remains insufficiently theorized and underreported within educational cybersecurity literature, leaving institutions with limited guidance on how to anticipate or counter it (Hulus, 2026).

7.3. Interface Exploitation and “Vibe Hacking”

The effectiveness of AI-driven social engineering cannot be fully explained through technical means alone. It depends substantially on interface design and the psychological and behavioral dispositions of users, a pattern that has been informally termed *vibe hacking*, referring to attacks that succeed primarily by exploiting a target's affective trust in a familiar-seeming interface rather than by defeating a technical safeguard (Wiberg, 2026). This positions interface exploitation as a genuinely human-computer interaction concern, not merely a downstream consequence of technical vulnerability, and reinforces the chapter's broader argument that cy-

bersecurity failures in education are frequently failures of design as much as they are failures of infrastructure.

The psychological mechanisms underlying this vulnerability are consistent with established findings in cyberpsychology regarding trust formation in digital interactions, including the tendency of users to extend trust based on surface-level cues of familiarity and legitimacy, such as institutional branding, formatting conventions, and communication tone, rather than verifying the underlying authenticity of a communication through more effortful means (Rossi et al., 2024). Students, who are frequently navigating a high volume of legitimate institutional communication as a routine feature of academic life, may be particularly susceptible to this dynamic, since the cognitive effort required to distinguish a sophisticated AI-generated impersonation from genuine institutional correspondence is considerably higher than the effort required to process ordinary, expected communication (Wiberg, 2026).

7.4. Evidence from Live Threat Intelligence

Live cyber threat intelligence dashboards, including those maintained by Check Point, have repeatedly identified the education sector as a primary global target for cyberattacks, lending empirical weight to the pattern described throughout this section (Check Point Research, 2025). Comparable findings from other threat intelligence sources reinforce this ranking across successive reporting periods, suggesting that the sector's exposure is a durable structural condition rather than a transient anomaly tied to any single reporting cycle (Islam et al., 2026). The persistence of this ranking across reporting periods underscores the urgency of integrating cybersecurity awareness and data protection more directly into educational frameworks, rather than treating them as a peripheral concern managed solely by institutional information technology departments (Check Point Research, 2025)

The consistency of this ranking across multiple independent sources and reporting periods is itself analytically significant. It suggests that the vulnerability of the education sector is not attributable to any single, correctable lapse, but to durable structural conditions, including the decentralized governance, resource constraints, and rapid, under-scrutinized technology adoption described throughout this chapter, that will continue to produce elevated risk absent a more fundamental reorientation of how educational institutions approach cybersecurity and data governance (Islam et al., 2026).

8. ANSWERING THE RESEARCH QUESTION

This chapter has examined how the AI-powered cybersecurity apparatus in education, marketed as protective, itself contributes to the erosion of data privacy, provenance, consent, and human-centered design in AI-mediated learning environments. The evidence presented across the preceding sections supports an affirmative and multidimensional answer.

With respect to data privacy, the expansion of personal, academic, biometric, and behavioral data collection has proceeded largely through interfaces designed to minimize friction rather than to maximize genuine informed consent, allowing data harvesting to occur under conditions that students neither fully understand nor meaningfully control (Dadson et al., 2025; Rossi et al., 2024; Wang et al., 2024). This is compounded by the aggregation effect described in Section 3, whereby the cumulative exposure a student accumulates across multiple platforms considerably exceeds what any single platform's disclosure practices would suggest, and by the absence of institutional mechanisms capable of assessing that aggregate exposure (Klimova & Pikhart, 2025; Vieriu & Petrea, 2025; Arizmendi et al., 2022).

With respect to trust and transparency, the black-box character of many AI systems, compounded by accessibility and human-centered design gaps, has left students without the interpretive resources necessary to hold these systems accountable (Wiberg, 2026; Martín-Moncunill & Alonso Martínez, 2025; Passlack et al., 2025). This deficit is reinforced by the asymmetry in technical literacy between institutional procurement processes and the students ultimately subject to procured tools, a structural condition that concentrates decision-making authority in the population least exposed to the resulting risk (Dadson et al., 2025; Passlack et al., 2025; Ali et al., 2024).

With respect to provenance, persistent uncertainty surrounding the origin and extent of AI involvement in academic work, together with the limited accessibility of available provenance information, has undermined institutions' capacity to adjudicate academic integrity fairly and to take responsibility for system outputs (Zdravkova, 2026; Wiberg, 2026). This failure is not merely technical but interactional, since the value of provenance information depends on whether it reaches students and educators in a genuinely comprehensible and actionable form (Wiberg, 2026).

With respect to consent and proprietary rights, students routinely engage in AI-assisted work without sufficient clarity regarding data use, consent boundaries, and responsibility for submitted material, while institutional-level consent frequently substitutes for individually informed consent in ways that leave students with little meaningful choice (Sakka et al., 2026; Dadson et al., 2025; Wiecezorek et al., 2025). This is compounded by the unresolved status of third-party material submitted to AI systems, a gap that leaves students making consequential decisions about the

intellectual property and data of others without adequate institutional guidance (Go et al., 2026; Sakka et al., 2026; Wieczorek et al., 2025).

With respect to the threat landscape, the same AI capabilities marketed as strengthening institutional security have simultaneously enabled more sophisticated forms of attack, including social engineering that exploits interface design and user psychology as readily as it exploits technical vulnerability (Islam et al., 2026; Wiberg, 2026; Rossi et al., 2024). The persistence of cybersecurity threats across educational contexts suggests that this is a durable structural condition rather than a transient or easily correctable vulnerability (Islam et al., 2026).

Taken together, these findings indicate that the erosion identified in the research question is not attributable to a single point of failure, but to a recurring pattern across otherwise distinct domains: AI-powered systems in education are consistently designed, deployed, and marketed in ways that privilege functional and commercial objectives over the transparency, accountability, and user-centered design that genuine data security would require (Dadson et al., 2025; Yan et al., 2025; Ali et al., 2024). The apparatus positioned as protective is, on the evidence reviewed here, frequently implicated in the very harms it claims to guard against. This is not to suggest that AI-powered cybersecurity tools are without any protective value, but rather that their protective value has been substantially overstated relative to the risks their broader deployment introduces, a gap between marketed promise and delivered protection that constitutes the illusion of security named in this chapter's title (Islam et al., 2026; Ali et al., 2024; Wang et al., 2024). This conclusion motivates the remaining task of this chapter, which is to articulate why an integrated response, one that draws simultaneously on human-computer interaction, cybersecurity, and educational governance, is necessary, and how this volume contributes to that response.

9. RATIONALE AND CONTRIBUTION OF THIS VOLUME

The preceding sections have demonstrated that data privacy, provenance, consent, human-centered design, and cybersecurity in education are not separable concerns to be addressed by distinct specialist communities operating in isolation from one another. Rather, they constitute interlocking dimensions of a single, underexamined problem: the deployment of AI-powered systems in educational settings under a presumption of security and benefit that the evidence reviewed in this chapter does not fully substantiate (Islam et al., 2026; Yan et al., 2025; Ali et al., 2024).

This interlocking character is precisely why an integrated response, one that draws simultaneously on human-computer interaction, cybersecurity, AI governance, and educational scholarship, is necessary rather than optional (Garzón et al., 2025; Yan et al., 2025; Ali et al., 2024). A cybersecurity intervention that does not account

for interface design and user psychology will fail to anticipate how attacks such as AI-driven social engineering actually succeed (Wiberg, 2026; Rossi et al., 2024). A data governance policy that does not account for human-centered design may produce consent mechanisms that are technically compliant but practically meaningless (Wiberg, 2026; Dadson et al., 2025; Rossi et al., 2024). Likewise, provenance information that is not surfaced through accessible interfaces may address a technical requirement while leaving the interactional problem, and the accountability gap that follows from it, untouched (Zdravkova, 2026; Wiberg, 2026). No single disciplinary lens, applied alone, is adequate to the problem this chapter has described.

This diagnosis carries practical implications for each of the readership groups identified in Section 1. For institutional policymakers, it suggests that procurement decisions regarding AI-powered platforms cannot be adequately evaluated through security checklists alone, but require integrated assessment of interface design, accessibility, and aggregate data exposure across an institution's full technology portfolio (Garzón et al., 2025; Dadson et al., 2025; Ali et al., 2024). For human-computer interaction and user experience designers, it suggests an expanded design mandate that treats consent comprehension, provenance transparency, and accessibility as core design requirements rather than secondary compliance features layered onto an otherwise complete system (Wiberg, 2026; Rossi et al., 2024). For cybersecurity practitioners, it suggests that effective defense against emerging threats such as AI-driven social engineering requires collaboration with human-computer interaction and cyberpsychology expertise that has not traditionally been part of educational information technology security practice (Islam et al., 2026; Wiberg, 2026). For educators and administrators operating at the point of contact between these systems and students, it suggests a need for institutional advocacy and literacy that extends beyond simply implementing tools handed down through procurement processes in which they had limited input (Go et al., 2026; Hulus, 2026; Passlack et al., 2025).

This volume responds to that necessity by convening contributions that span these domains, addressing the convergence of AI, digital pedagogy, cybersecurity, human-computer interaction, and user experience introduced at the outset of this chapter. Contributions to this volume are expected to range from empirical investigations of specific platforms and populations, to policy-oriented analyses of institutional governance frameworks, to design-oriented proposals for more transparent and accountable AI-powered educational tools. Where this introductory chapter has offered a critical review establishing the scope, mechanisms, and stakes of the problem, the chapters that follow extend this foundation into more specific empirical, technical, and policy-oriented territory.

Their collective contribution is to move the conversation surrounding AI-powered cybersecurity in education beyond the presumption that securing data privacy is a solved problem, and toward a more rigorous, human-centered, and accountable

framework for AI-mediated learning environments going forward. The chapter has argued throughout that the apparatus of AI-powered cybersecurity in education, however well intentioned in its individual components, has not delivered the comprehensive protection its marketing and institutional adoption presume. Addressing that gap will require sustained, cross-disciplinary attention of precisely the kind this volume seeks to convene, and represents the central task to which the remaining chapters of this volume are dedicated.

REFERENCES

- Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K., & Malik, F. T. (2024). The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change*, 199(123076), 123076–123076. DOI: 10.1016/j.techfore.2023.123076
- Alotaibi, N. B. (2026). Adoption of AI tools in education: Examining influencing factors and their impact on students’ critical thinking skills. *Computers in Human Behavior Reports*, 21, 100990. DOI: 10.1016/j.chbr.2026.100990
- Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2022). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, 55(6), 3026–3054. Advance online publication. DOI: 10.3758/s13428-022-01939-9 PMID: 36018483
- Barzanji, C., & Loitsch, C. (2025). Exploring conversational agents for novice programmers: A scoping review. *Discover Artificial Intelligence*, 5(1), 271. Advance online publication. DOI: 10.1007/s44163-025-00521-4
- Bernik, A., Radošević, D., & Čep, A. (2025). AI-Powered learning: Revolutionizing education and automated code evaluation. *Information (Basel)*, 16(11), 1015–1015. DOI: 10.3390/info16111015
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. DOI: 10.1191/1478088706qp063oa
- Check Point Research. (2025). Back-to-school sees cyber attacks surge in education. In *Check Point Blog*. CheckPoint. <https://blog.checkpoint.com/research/cyber-attacks-surge-against-education-sector-ahead-of-back-to-school-season/>
- Dadson, Y. A., Mohammed, Z., Hussein, B., Armah, Y. E., & Kumi-Yeboah, A. (2025). Ethics and privacy in AI education: Prospects and challenges in higher education. In Adarkwah, M. A., Amponsah, S., Huang, R., & Thomas, M. (Eds.), *Artificial Intelligence and Human Agency in Education: Volume One* (pp. 215–242). Springer Nature Link., DOI: 10.1007/978-981-96-7937-9_10

European Parliament & Council of the European Union. (2022, December 14). Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive). *Official Journal of the European Union*, L, 333, 80–152. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022L2555>

Garzón, J., Patiño, E., & Marulanda, C. (2025). Systematic Review of Artificial Intelligence in Education: Trends, Benefits, and Challenges. *Multimodal Technologies and Interaction*, 9(8), 84–84. DOI: 10.3390/mti9080084

Go, B., Min, S., & Jeon, M. (2026). Making AI in education responsible: Classroom-ready ethical guidelines for teachers and students. *AI and Ethics*, 6(3), 337. Advance online publication. DOI: 10.1007/s43681-026-01176-2

Hulus, A. (2026). A systematic response to ethical blind spots in AI education: Cross-cultural insights and the role of digital literacy. *Smart Learning Environments*, 13(1), 11. Advance online publication. DOI: 10.1186/s40561-026-00437-1

Hummel, S. (2025). Ethical and responsible AI in education: Situated ethics for Democratic learning. *Education Sciences*, 15(12), 1594–1594. DOI: 10.3390/educsci15121594

Islam, R., Yildirim, G., Rahman, A., & Kamal, M. S. (2026). Artificial intelligence in cybersecurity and education: A bibliometric review of emerging trends and research frontiers. *Sustainable Futures : An Applied Journal of Technology, Environment and Society*, 11, 101807. DOI: 10.1016/j.sftr.2026.101807

Klimova, B., & Pikhart, M. (2025). Exploring the effects of artificial intelligence on student and academic well-being in higher education: A mini-review. *Frontiers in Psychology*, 16(16), 1498132. Advance online publication. DOI: 10.3389/fpsyg.2025.1498132 PMID: 39963679

Malhotra, M., & Chhabra, I. (2026). Ensuring academic integrity through automated online exam proctoring a decade long systematic review. *Discover Education*, 5(1), 207. Advance online publication. DOI: 10.1007/s44217-026-01224-3

Martín-Moncunill, D., & Alonso Martínez, D. (2025). Students' trust in AI and their verification strategies: A case study at Camilo José Cela University. *Education Sciences*, 15(10), 1307. DOI: 10.3390/educsci15101307

Nasni Naseri, R. N., & Abdullah, M. S. (2024). Understanding AI technology adoption in educational settings: A review of theoretical frameworks and their applications. *Information Management and Business Review*, 16(3(I)), 174–181. DOI: 10.22610/imbr.v16i3(I).3963

O'Malley, A., Veenhuizen, M., & Ahmed, A. (2024). Ensuring Appropriate Representation in Artificial Intelligence-Generated Medical Imagery: Protocol for a Methodological Approach to Address Skin Tone Bias. *JMIR AI*, 3, e58275. DOI: 10.2196/58275 PMID: 39602221

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., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal*, 372(71), n71. Advance online publication. DOI: 10.1136/bmj.n71 PMID: 33782057

Passlack, N., Hammerschmidt, T., & Posegga, O. (2025). With great power comes great responsibility: What shapes AI literacy for responsible interactions of knowledge workers with AI? *Information Systems Frontiers*, 28(1), 11–46. Advance online publication. DOI: 10.1007/s10796-025-10648-5

Qiang, J., & Liu, Y. (2025). AI-based security and privacy enhancements for student data protection in mobile educational networks. *Discover Applied Sciences*, 8(2), 175. Advance online publication. DOI: 10.1007/s42452-025-08144-1

Rossi, A., Carli, R., Botes, M. W., Fernandez, A., Sergeeva, A., & Sánchez Chamorro, L. (2024). Who is vulnerable to deceptive design patterns? A transdisciplinary perspective on the multi-dimensional nature of digital vulnerability. *Computer Law & Security Review*, 55, 106031. DOI: 10.1016/j.clsr.2024.106031

Sakka, S., Liagkou, V., Ferreira, A., & Stylios, C. (2026). Digital boundaries and consent in the metaverse: A comparative review of privacy risks. *Journal of Cybersecurity and Privacy*, 6(1), 24. DOI: 10.3390/jcp6010024

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104(1), 333–339. ScienceDirect. <https://doi.org/https://doi.org/10.1016/j.jbusres.2019.07.039>

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, 15(3), 343. DOI: 10.3390/educsci15030343

Wang, C. (2024). Exploring students' generative AI-Assisted writing processes: Perceptions and experiences from native and nonnative English speakers. *Technology, Knowledge and Learning*, 30(3), 1825–1846. Advance online publication. DOI: 10.1007/s10758-024-09744-3

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252, 124167. DOI: 10.1016/j.eswa.2024.124167

Wiberg, C. (2026). Enhancing cybersecurity through user experience (UX) design: Bridging usability and protection. *Lecture Notes in Computer Science*, 16337, 391–404. DOI: 10.1007/978-3-032-12801-0_26

Wieczorek, M., Hosseini, M., & Gordijn, B. (2025). Unpacking the ethics of using AI in primary and secondary education: A systematic literature review. *AI and Ethics*, 5(5), 4693–4711. Advance online publication. DOI: 10.1007/s43681-025-00770-0 PMID: 40959538

Yan, Y., Liu, H., & Chau, T. (2025). A systematic review of AI ethics in education. *Journal of Global Information Management*, 33(1), 1–50. DOI: 10.4018/JGIM.386381

Zdravkova, K. (2026). Generative Artificial Intelligence and the ambiguity of academic integrity in higher education. *Education Sciences*, 16(7), 1120. DOI: 10.3390/educsci16071120

KEY TERMS AND DEFINITIONS

AI Governance: The policies, regulatory frameworks, ethical principles, and institutional practices that guide the responsible design, deployment, oversight, and use of AI systems.

AI-Mediated Learning Environments: Learning environments in which AI technologies actively influence instructional delivery, assessment, learner support, or institutional decision-making.

AI-Powered Cybersecurity: The application of AI techniques to detect, prevent, and respond to cybersecurity threats, while recognising that these systems may also introduce new privacy, transparency, and governance challenges.

Artificial Intelligence (AI): Computational systems capable of performing tasks that typically require human intelligence, including learning, reasoning, prediction, content generation, and decision support within educational environments.

Data Privacy: The protection of personal, behavioural, biometric, and academic information through appropriate collection, processing, storage, and governance practices.

Educational Technology (EdTech): Digital platforms and technologies used to support teaching, learning, assessment, administration, and student engagement in educational settings.

Human-Centered Design (HCD): A design approach that prioritises users' needs, abilities, contexts, and experiences throughout the development process to improve usability, accessibility, transparency, and trust.

Informed Consent: A process whereby individuals understand and voluntarily agree to how their data will be collected, processed, shared, and retained before using an AI-powered system.

Provenance: Information describing the origin, history, authorship, and generation process of digital content, enabling accountability and supporting fair academic integrity decisions.

Trust and Transparency: The extent to which AI systems communicate their data practices, decision-making processes, and limitations in ways that users can understand, evaluate, and challenge.

