TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Rashad Gafarli 233918IVCM

ROBOT ASSISTANTS IN HIGHER EDUCATION: A STUDY OF ETHICAL AND CYBERSECURITY CHALLENGES

Master's Thesis

Supervisor: Fuad Budagov
MBA

Tallinn 2026

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Rashad Gafarli 233918IVCM

ROBOTASSISTENDID KÕRGHARIDUSES: EETILISTE JA KÜBERTURBE-ALASTE PROBLEEMIDE UURING

Magistritöö

Juhendaja: Fuad Budagov
MBA

Tallinn 2026

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Rashad Gafarli

04.01.2026

Abstract

As robot assistants become increasingly integrated into higher education environments, they offer promising advancements in personalized learning, administrative efficiency, and remote accessibility. However, their adoption also raises critical ethical and cybersecurity concerns that must be addressed to ensure responsible and secure implementation. This study explores the dual dimensions of ethical and cybersecurity challenges associated with the use of robot assistants in higher education. Through an interdisciplinary review of current literature and case examples, this thesis identifies key issues such as data privacy, AI bias, student surveillance, autonomy in educational decision-making, and system vulnerabilities. It further examines regulatory gaps, institutional preparedness, and the potential risks posed by AI-driven robotic systems handling sensitive academic and biometric data. The findings underscore the urgent need for robust ethical guidelines, privacy protections, and cybersecurity frameworks tailored to educational contexts. By evaluating these risks and proposing best practice recommendations, this study aims to contribute to the safe, equitable, and ethical integration of robotic assistants in higher learning institutions.

The thesis is written in English and is 82 pages long, including 8 chapters, 9 figures and 18 tables.

List of Abbreviations and Terms

2FA	Two-Factor Authentication
3D	Three-Dimensional
ABB	Asea Brown Boveri
ACA	Attendance Check Application
AES	Advanced Encryption Standard
AI	Artificial Intelligence
API	Application Programming Interface
ARI	Autonomous Robot Interface
AV	Autonomous Vehicle
CETYS	Centro de Enseñanza Técnica y Superior
CMDB	Configuration Management Database
COAST	Cybersecurity Operations and Strategy Team
CVE	Common Vulnerabilities and Exposures
DCAE	Deep Cognitive Architecture for Evolution
DDS	Data Distribution Service
EU	European Union
EU AI Act	European Union Artificial Intelligence Act
FCC	Federal Communications Commission
FERPA	Family Educational Rights and Privacy Act
GDPR	General Data Protection Regulation
GPS	Global Positioning System
HE	Higher Education
IAM	Identity and Access Management
IEC	International Electrotechnical Commission
IEEE Xplore	Institute of Electrical and Electronics Engineers Xplore
IoT	Internet of Things
ISO/SAE	International Organization for Standardization/Society of Automotive Engineers
K–12	Kindergarten through 12th Grade
LBR	Lightweight Robot
LED	Light Emitting Diode
LIDAR	Light Detection and Ranging
MITM	Man-In-The-Middle

MQTT	Message Queuing Telemetry Transport
NAO	Nao Robot
OS	Operating System
OTA	Over-The-Air
PIAs	Privacy Impact Assessments
PII	Personally Identifiable Information
PIN	Personal Identification Number
PRO-CARED	Professional Care Robot Development
RBAC	Role-Based Access Control
RGB	Red Green Blue
RQ	Research Question
ROS	Robot Operating System
RVD	Robot Visual Detection
SAGES	Smart Autonomous Global Engineering Systems
SIFROBOT (SIF-SOF)	Smart Intelligent Framework for Robotic Systems
SLAM	Simultaneous Localization and Mapping
SLR	Systematic Literature Review
SMS	Systematic mapping study
SROS	Secure Robot Operating System
SSID	Service Set Identifier
SSH	Secure Shell
STEM	Science, Technology, Engineering, and Mathematics
TDNN	Time-Delay Neural Network
THR	Task-based Human-Robot Interaction
VLAN	Virtual Local Area Network
VPN	Virtual Private Network
WebRTC	Web Real-Time Communication

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

The rapid digital transformation of higher education (HE) driven by breakthroughs in artificial intelligence (AI), robotics, and remote learning has redefined both teaching and administrative practices. Robot assistants have emerged as powerful tools to enhance engagement, streamline support tasks, and foster inclusivity. Their use in automating classroom attendance, facilitating navigation, delivering announcements, and supporting real-time interaction enables educators to shift their focus toward deeper student engagement and instructional quality.

Recent literature emphasizes the pedagogical promise of educational robotics in promoting collaborative and interdisciplinary learning [1]. However, as Scaradozzi et al. (2019) cautions, the use of robots does not inherently improve learning outcomes unless thoughtfully integrated into instructional strategies [2]. Educational robotics, while promising, often lacks clear alignment with curricular goals and replicability without robust assessment frameworks. Nonetheless, when well-implemented, these systems can help develop digital competencies and support broader goals such as inclusion and sustainability [3].

This study is motivated by both a scholarly need to address overlooked ethical and cybersecurity risks in robot assistant deployment and a personal interest in their societal impact. The past decade has seen an unprecedented acceleration in AI and robotics, creating both exciting possibilities and complex challenges. As institutions increasingly rely on intelligent, autonomous systems, it is crucial to establish responsible design and governance models. This thesis seeks to contribute to that effort by developing an integrated ethical and security framework to guide the safe, equitable, and effective use of robot assistants in HE.

1.2 Research Problem

While robot assistants offer notable instructional and administrative advantages in HE, their adoption introduces complex ethical and cybersecurity challenges. Concerns such as algorithmic bias, erosion of student privacy, and diminished human interaction intersect with technical risks like unauthorized access, data leaks, and system-level manipulation. These issues are frequently overlooked during early development phases, where function-

ality is prioritized over security-by-design or ethical safeguards. The situation is further complicated by the forthcoming rollout of the EU Artificial Intelligence Act, which introduces compliance obligations that remain unclear for many institutions and developers. As enforcement phases continue through 2030, HE institutions often lack the policies, risk frameworks, and technical readiness needed to interpret and implement the Act's requirements effectively. This regulatory uncertainty, combined with the absence of comprehensive internal safeguards, risks undermining student rights, institutional integrity, and the long-term resilience of digital academic environments.

1.3 Research Goal

This thesis examines the ethical and cybersecurity issues of robot assistants in higher education, identifying key risks, assessing implications, and proposing practical guidelines for secure, responsible adoption. By aligning innovation with accountability, the study aims to facilitate the sustainable integration of robotic systems within academic environments. A key focus of this research is understanding how emerging regulatory frameworks particularly the EU Artificial Intelligence Act shape the deployment of robot assistants in educational contexts. The study examines the Act's implications and highlights the key compliance considerations that institutions and developers must navigate to ensure lawful and ethical implementation.

To achieve this, the thesis conducts a comprehensive review of current literature, identifying core security themes such as physical safety, data privacy, communication integrity, and ethical governance. It also evaluates the limitations of existing standards, originally developed for industrial applications, and develops tailored recommendations suited to the unique operational realities of education. The insights gained aim to support researchers, policymakers, developers, and institutions in advancing secure and trustworthy robotic solutions that align with legal, ethical, and societal expectations.

1.4 Research Scope

This thesis focuses on the ethical and cybersecurity dimensions of robot assistants deployed in HE. It reviews physically embodied robots used in teaching and administration, including telepresence robots, humanoid and social robots, service and delivery robots, lab robots, and quadrupedal platforms. These systems are analyzed in relation to their integration into academic workflows such as classroom interaction, attendance automation, student support, and campus operations.

The study prioritizes literature and case studies addressing ethical risks (e.g., bias, surveillance, autonomy) and cybersecurity challenges such as data privacy, system vulnerabilities, and compliance with emerging regulations. Technical studies were included where they help illustrate robot capabilities or inform assessments of ethical and security implications. However, works focused solely on mechanical design or functional performance, without relevance to broader risk, were not emphasized.

Robots originating in other sectors such as healthcare or industry were considered when their functionality or risk profile was applicable to academic settings. In contrast, systems designed explicitly for K–12 education, entertainment, or toy use were excluded unless their capabilities were clearly transferable to the HE context.

1.5 Research Questions

To guide the investigation, the following research questions (RQ) are addressed:

- **RQ1:** What ethical and cybersecurity challenges arise from deploying robot assistants in higher education, particularly regarding autonomy, bias, surveillance, transparency, and system-level vulnerabilities?
- **RQ2:** What implications does the European Union Artificial Intelligence Act have for robot assistants in higher education, and what key compliance considerations must institutions and developers be aware of?
- **RQ3:** How can individuals, institutions, or developers approach the design and deployment of robot assistants in a way that ensures ethical alignment and cybersecurity resilience in higher education?

1.6 Novelty

This thesis addresses a critical and underexplored intersection: the ethical and cybersecurity risks associated with robot assistants in HE. While research highlights pedagogical benefits such as engagement and accessibility, limited integrated analysis in combination with ethical and cybersecurity risks exists.

Ethical and cybersecurity concerns are frequently treated in isolation. Ethical discussions tend to focus on issues like AI bias, surveillance, and autonomy, while cybersecurity research emphasizes institutional IT threats. However, robot assistants combine physical presence with AI-driven decision-making, introducing distinct vulnerabilities that span both domains. Their growing presence in post-pandemic academic settings raises pressing

questions about data privacy, autonomous assessments, and the resilience of connected systems.

By analyzing existing literature and standards, this thesis develops a thematic framework and tailored guidelines that consolidate security considerations and identify gaps in current practices. It offers a consolidated reference point that synthesizes fragmented research into a cohesive, structured perspective. In doing so, it identifies emerging trends, maps key areas of concern, and surfaces influential works that might otherwise remain disconnected. The resulting framework serves as a foundation for researchers, institutions, developers, and policymakers to guide the secure, ethical, and informed deployment of robot assistants in education.

1.7 Contributions

This thesis contributes a systematic, integrated, and practice-oriented account of the ethical and cybersecurity challenges posed by robot assistants in higher education. First, it presents a structured mapping study synthesising insights from 84 sources published between 2018 and 2025. This mapping identifies recurring risk themes across autonomy, bias, surveillance, data protection, and system-level vulnerabilities, and connects them to concrete HE use cases.

Second, the thesis proposes an integrated, control-based framework that unifies ethical and cybersecurity domains and explicitly maps them to relevant provisions of the EU Artificial Intelligence Act for high-risk educational AI systems. The framework organises requirements into practical control domains and sub-controls that support both design-time planning and deployment-time assessment.

Third, the thesis evaluates the framework through a qualitative analysis of three real-world robot assistant deployments in higher education. Using the framework, the evaluation identifies gaps in areas such as data privacy, system integrity, vulnerability management, and human oversight, and shows how the approach can surface latent risks not addressed by generic IT security policies.

Finally, the thesis translates these contributions into an actionable compliance and governance roadmap for institutions. This roadmap provides stepwise recommendations for universities and developers, including prioritised controls, governance measures, and robotics-specific safeguards aligned with EU AI Act requirements.

2. Literature Review

While numerous studies have explored the effectiveness of robotic teaching aids in enhancing student engagement and learning outcomes, few have directly addressed the dual dimensions of ethics and cybersecurity within the context of HE. This review aims to bridge that gap by systematically mapping literature that discusses robot assistants through both an ethical and security-conscious lens.

This systematic mapping study (SMS) adopts a structured approach to categorize and synthesize literature covering key security and ethical concerns, including physical safety, data protection, system vulnerabilities, algorithmic fairness, and compliance with emerging regulations such as the EU AI Act. Rather than focusing solely on technical performance or pedagogical applications, the SMS evaluates how existing work addresses responsible design, institutional risk, and governance.

The study critically assesses the current body of research, identifies gaps in ethical and cybersecurity coverage, and proposes a thematic framework and actionable guidelines to support the safe and accountable deployment of robot assistants in HE settings.

2.1 Search Strategy

This study followed the Systematic Literature Review (SLR) methodology outlined by Kitchenham [4], ensuring rigor, transparency, and repeatability in synthesizing knowledge on ethical and cybersecurity aspects of robot assistants in HE. The review process included protocol formulation, iterative keyword development, database selection, and the application of clear inclusion and exclusion criteria. Additionally, the structure and procedural refinement of this review were informed by Budagov et al. [5], whose recent SLR on the applicability of robot assistants in HE served as a relevant domain-specific reference.

To supplement traditional database searches, snowballing techniques were applied as described by Wohlin [6]. Backward and forward citation tracking were used to uncover additional literature that may have been missed by keyword-only queries. Manual inclusion ("single picking") was also performed, targeting specific robot platforms—such as Pepper, NAO, and Temi once they were revealed in initial searches.

The review acknowledges that, due to the applied nature of educational robotics and the fast

pace of technological development, relevant information is not limited to academic sources. Therefore, credible non-academic references such as manufacturer documentation, official product pages, technical whitepapers, and verified media reports—were also consulted for details on capabilities, specifications, deployments, and real-world incidents. These sources are especially critical when discussing commercial robots and current implementations in HE. Although the main scope focuses on HE, the search was purposefully extended to adjacent institutional settings when ethical or cybersecurity implications were relevant and transferable.

Searches were conducted in four major databases and limited to English-language publications from January 2018 to March 2025. Table 1 outlines the search configuration.

Table 1. Search Strategy

Attribute	**Details**
Electronic Databases	IEEE Xplore, Web of Science, Scopus, SpringerLink
Type of Literature	Peer-reviewed journal and conference papers
Search String	("robot assistants" OR "educational robots" OR "social robots" OR "robotic teaching assistants") AND ("data privacy" OR "cybersecurity" OR "ethics" OR "AI risk" OR "robot misuse" OR "facial recognition") AND ("HE" OR "university" OR "college" OR "postsecondary")
Language	English
Publication Period	January 2018 – March 2025
Supplemental Methods	Snowballing, Manual Inclusion (Robot-Specific Sources)

2.2 Inclusion and Exclusion Criteria

To ensure focus and relevance, the review included sources that met the following criteria: (a) peer-reviewed journal articles or conference papers published between January 2018 and March 2025; (b) publications discussing robot assistants within HE contexts; (c) works addressing ethical, cybersecurity, or regulatory concerns related to educational robotics; (d) technical documentation or official manufacturer sources used to describe real-world capabilities, deployments, or risk factors; and (e) additional relevant studies identified through citation snowballing and manual selection based on earlier findings.

Studies were excluded if they (a) focused solely on non-postsecondary contexts such as K–12 education, entertainment, or corporate training without clear applicability to HE; (b) were not accessible in full text at the time of review; or (c) provided only technical or functional analysis without engaging with ethical, cybersecurity, or governance dimensions.

2.3 Literature Search and Selection

The initial search retrieved articles from four major academic databases: IEEE Xplore (n = 111), Web of Science (n = 245), SpringerLink (n = 136), and Scopus (n = 1087). After removing titles that fell outside the scope and eliminating duplicates, 271 records remained for screening. Of these, 135 were excluded for not addressing robot-assisted learning or lacking relevance to HE. Full-text review of the remaining articles led to the exclusion of an additional 74 that did not sufficiently cover ethical or cybersecurity concerns.

To broaden the review, citation snowballing was used to identify additional relevant sources. Manually curated literature such as product documentation, technical specifications, and official developer resources was also included, based on robot models and deployments discovered during the initial review. In total, 84 sources were included in the final analysis from the literature selection. Figure 1 illustrates the complete search and selection process.

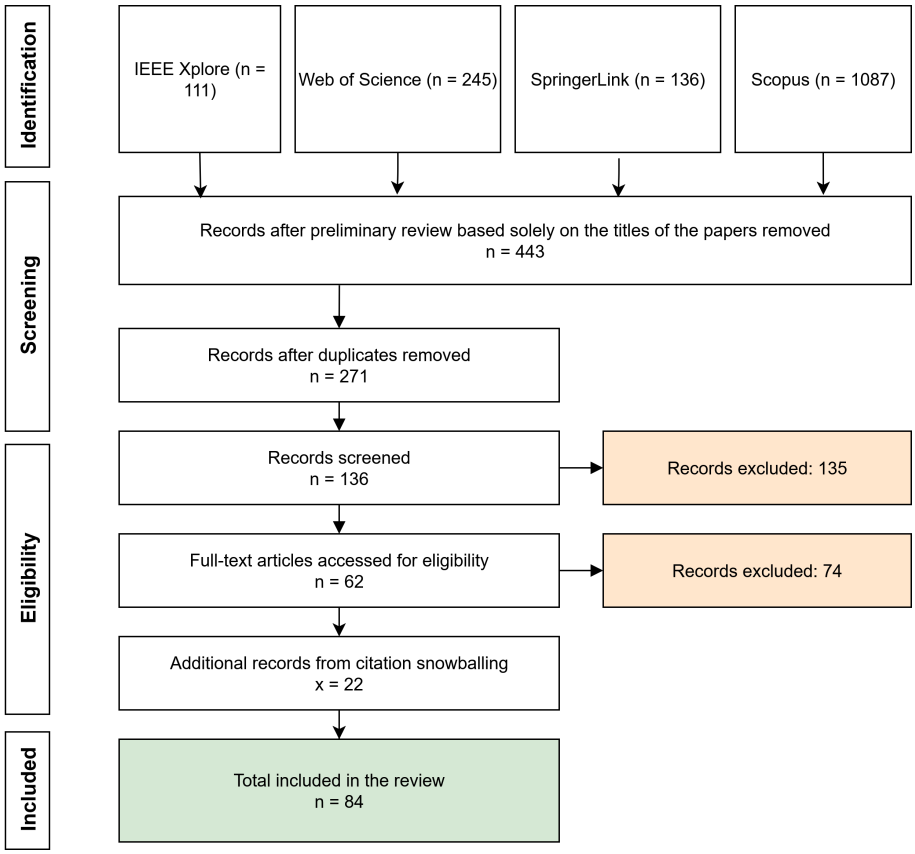

Figure 1. Diagram of the literature selection process

3. Methodology

3.1 Overview

This study adopts a three-phase methodology to explore and apply ethical and cybersecurity considerations related to the use of robot assistants in HE. Rather than relying solely on abstract analysis, the methodology culminates in three applied case studies at TalTech to test the framework against real deployments. Each phase contributes to the development of a comprehensive guideline-based framework that institutions can use both as a planning resource and an evaluative tool.

3.2 Mapping Challenges

The first phase involves collecting and synthesizing known ethical and cybersecurity challenges associated with robot assistants in academic environments. This is achieved through a systematic review of academic literature, regulatory sources, technical documentation, and institutional practices. Particular emphasis is placed on risk areas such as data privacy, algorithmic bias, surveillance, autonomy, accountability, access control, and system-level vulnerabilities.

In this phase, the requirements outlined in the European Union Artificial Intelligence Act (EU AI Act) are also incorporated into the analysis. Relevant AI Act provisions—particularly those relating to high-risk systems, transparency obligations, human oversight, and documentation requirements—are mapped onto the identified challenge domains. This ensures the framework is grounded in current academic discourse and aligned with emerging regulatory standards that institutions will increasingly need to meet through 2030 and beyond.

Importantly, this phase of research considers the full spectrum of robot assistants currently deployed or proposed in HE. Each category introduces distinct risk vectors and socio-technical interactions, making it essential to account for their varying roles in classroom, administrative, and remote engagement settings. This comprehensive inclusion ensures that the challenges identified are not limited to any single robot form factor or function but reflect the diverse applications emerging across academic contexts.

The objective is to develop a structured foundation of recurring themes, dilemmas, and

technical threats that are consistently highlighted in the literature. These findings are organized into thematic categories to allow for easier reference and later application.

3.3 Categorization into a Guideline-Driven Framework

Following risk identification, findings are structured into a guideline-driven framework intended primarily as a collection of practical guidelines, controls, and best practices. Each control is accompanied by specific evaluation indicators, enabling institutions to assess their degree of alignment and readiness. Although not intended as a rigid compliance checklist, the framework can function as a self-assessment tool to evaluate both individual control implementation and overall deployment maturity.

The framework categorizes risks into two main domains:

- **Ethical Considerations:** including fairness, transparency, consent, inclusivity, autonomy, and human oversight.
- **Cybersecurity Considerations:** including authentication, encryption, system integrity, network security, and access control.

3.4 Case Study Evaluation

The third phase applies the developed framework to a real-world use case at Tallinn University of Technology (TalTech). This case study examines the live deployments of a robot assistant as part of a pilot tests [7, 8, 9]. With direct access to the environment, robots, and relevant personnel, it enables a grounded evaluation of how institutional practices align with the framework's guidelines.

Using the guideline-driven framework, the case study examines:

- Which ethical and cybersecurity risks have been explicitly addressed,
- Which risks remain unresolved or are accepted within institutional policy,
- What mitigation strategies are in place or under development,
- How ethical, regulatory, and operational trade-offs are managed by stakeholders.

This approach offers a practical validation of the framework, while also generating insights into implementation challenges, policy gaps, and institutional preparedness. It highlights the value of proactive evaluation and the benefits of aligning robot deployments with both emerging standards and best practices.

4. Study of Challenges

This section presents a structured investigation into the different types of robot assistants used in HE and analyzes the key cybersecurity and ethical challenges they pose.

4.1 Types of Robot Assistants in HE

A structured investigation into the different types of robot assistants used in HE and analysis of key cybersecurity and ethical challenges they pose is presented here. It explores the diverse types of robot assistants currently used or studied within HE environments and, drawing on interdisciplinary research, examines the technical capabilities and roles of platforms ranging from humanoid tutors and telepresence devices to collaborative lab robots and animal-inspired companions. By analyzing these categories, the section aims to understand how different forms of robotic assistance are reshaping teaching, learning, accessibility, and campus life.

The five categories considered are not mutually exclusive: some platforms span multiple roles (e.g., Temi as both telepresence and service robot; humanoid robots acting as delivery assistants), and any given university deployment may combine several robot types into an ecosystem. Nonetheless, this typology clarifies key axes of variation. **Embodiment and sociality** range from utilitarian carts to highly anthropomorphic humanoids, shaping users' expectations, emotional responses, and ethical concerns. **Mobility and operating environment** vary from stationary lab cobots to outdoor delivery and quadrupedal robots, influencing safety requirements and exposure to environmental risks. **Autonomy and control** extend from fully teleoperated systems to highly autonomous navigation and interaction, with direct implications for accountability, transparency, and security considerations. **Function and integration** span narrow logistics tasks to broader roles in teaching, research, and campus life, determining both the sensitivity of the data processed and the criticality of the services provided. Understanding these dimensions is essential for the subsequent analysis of cybersecurity and ethical challenges, as each category presents a distinct attack surface, risk profile, and set of governance needs within HE institutions.

4.1.1 Telepresence and Remote Presence Robots

Telepresence robots enable remote participation by combining video conferencing with a mobile platform. They are remotely controlled devices with wheels and a screen/camera,

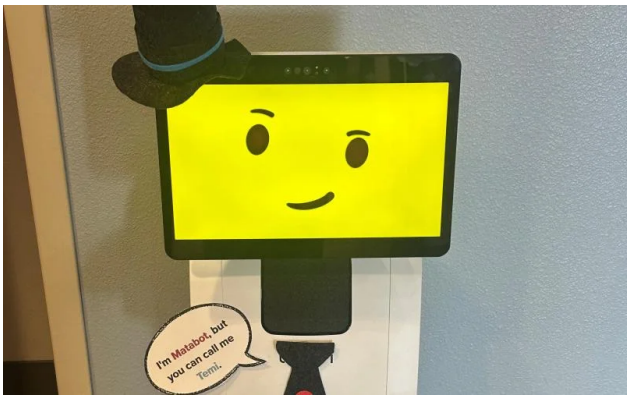

Figure 2. "MATABOT" a Temi V3 robot. Image source: Arizona Western College [10].

allowing an off-site user to navigate through a campus or classroom and interact in real time [11]. These robots can support students or staff who cannot be physically present and are increasingly used to enhance inclusion in HE. They enhance distance learning and inclusion by providing a physical proxy for remote learners [11]. Key examples include:

Double 3 (Double Robotics) – A two-wheeled balancing telepresence robot with an iPad-like screen that remote users can drive around hallways and classrooms. Universities have used Double to let homebound or distant students attend on-campus classes and interact with peers and instructors via video [12, 11].

Ohmni / OhmniLabs Telepresence – A lightweight telepresence robot offering high-definition video, remote control via web or mobile app, and auto-docking. It is commonly used for telecommuting to classes, enabling guest lecturers to "beam in" to lecture halls, and allowing students to tour campuses virtually [13, 14].

PadBot P2 (Inbot Technology) – A lightweight telepresence robot with a tilting screen, Wi-Fi and 4G LTE connectivity, and remote driving capabilities. According to the FCC manual, it features a 10-hour battery life and supports encrypted WebRTC connections, and has been deployed for mobile advising sessions or to assist students unable to physically attend university spaces [15, 16].

Beam (Suitable Technologies/Awabot) – A telepresence unit consisting of a stable base and large screen, with dual cameras and secure WebRTC-based communication. Beam has been used in academic conferences and classrooms to allow presenters or students to roam and interact from afar in a natural way [17, 18].

Temi (Robotemi) – A personal robot that functions both as an autonomous assistant and a telepresence unit, capable of autonomous navigation, following people, and facilitating video calls using voice commands and facial recognition. For example, Arizona Western College introduced a Temi V3 robot ("MATABOT") to guide students around campus and provide in-class support using AI-based voice interaction and autonomous navigation [19, 20, 10].

InTouch Vita / RP-Vita (InTouch Health) – A medical-grade remote presence robot originally designed for telemedicine, featuring autonomous navigation, high-definition video, encrypted communications, and remote-control capabilities. In teaching hospitals, it allows remote experts or professors to make "virtual rounds" with students and participate in training and meetings [21, 22].

Kubi (Xandex) – A tabletop telepresence device that holds a tablet and provides pan-tilt motion to give remote users limited control over their viewpoint. Kubi has been used in lecture halls and seminar rooms to give remote students a controllable view of discussions [23].

Capabilities: Telepresence robots typically feature live two-way audio/video, remote steering, and sometimes autonomous driving or self-docking. Some include obstacle avoidance, voice interaction, and saved location navigation (e.g., Temi can "go to" preset destinations), giving remote users a physical proxy that can increase social presence and engagement compared to static video calls [14, 20, 11].

Use Cases in HE: These robots are used for remote class attendance (e.g., homebound students attending labs), administrative meetings, and virtual campus tours [11]. Gallon et al. [24] show that telepresence robots can preserve social links and class continuity for students recovering from illness, especially when integrated with connected learning environments (shared whiteboards, VPN-based lab access). Wernbacher et al. [25] highlight their role in supporting long-term absence and hybrid teaching, while also noting infrastructure barriers and data privacy concerns. Overall, telepresence robots contribute to inclusion, digital literacy, and potentially reduced travel emissions, but require careful pedagogical integration and governance [24, 25].

4.1.2 Humanoid and Social Robots

Humanoid social robots are assistant robots designed to interact with people using human-like communication (speech, gestures, eye contact) and often a human-inspired body plan. In universities, these robots are used as interactive teaching aids, guides, or research platforms for human-robot interaction and social robotics. Key examples include:

Pepper (SoftBank Robotics) – A human-shaped robot about 1.2 meters tall with a touchscreen on its chest, articulating arms, and expressive LED eyes. Equipped with cameras, touch sensors, microphones, and speech recognition, Pepper has been used as a greeter and educational assistant, leading quizzes, giving interactive presentations, and practicing languages with students, often drawing on its emotion and face detection

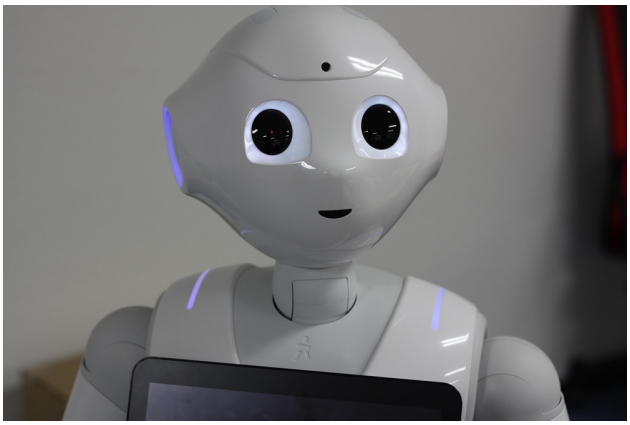

Figure 3. Pepper the robot (SoftBank Robotics). Image source: Wikimedia Commons, licensed under CC BY-SA 4.0 [26].

capabilities [27, 28, 29]. It is widely adopted as a research platform in academic settings focused on human-robot interaction and "Enactive Robot Assisted Didactics" [28, 29].

NAO (SoftBank Robotics) – A small humanoid robot (58 cm) widely used in K–12 and HE to teach programming, AI, and language skills. NAO can walk, dance, speak, and recognize faces, and has been shown to support vocabulary learning and engagement in adult classrooms through interactive exercises and feedback [30]. Its programmability via tools like Choregraphe or Python makes it a common platform for hands-on robotics and human-robot interaction experiments [31, 30].

ARI (PAL Robotics) – A modern humanoid robot with a torso, two arms, expressive LCD eyes, and a touchscreen on its chest. ARI is designed for advanced social interaction with integration of facial recognition, natural language processing, and emotion detection, and has been deployed in projects such as PRO-CARED to support language learning and personalized feedback in Catalan [32, 33]. Its ROS-compatible architecture and support for AI services make it a flexible research and teaching platform [32, 34].

SIFROBOT (SIFSOF) – A line of humanoid service robots (e.g., SIFROBOT-5.2) featuring a touchscreen torso, expressive head, and mobile base with autonomous navigation, voice interaction, facial recognition, and gesture-based control [35]. On campus, such robots can act as greeters, information kiosks, or delivery assistants in dynamic spaces like lobbies or libraries, leveraging adaptive interaction models described by Bello et al. [36].

Sophia (Hanson Robotics) – An advanced humanoid robot known for lifelike facial expressions and conversational abilities, capable of up to 62 facial expressions using a "Frubber" face [37, 38]. Sophia is frequently invited to university events and conferences as a guest lecturer or discussion facilitator, serving as a platform to explore ethics, the uncanny valley, and public perceptions of AI and robot personhood [37].

Romeo (SoftBank Robotics) – A taller humanoid robot (approx. 147 cm) developed as a research platform for assistive robotics, with 37 joints, multiple cameras, and tactile

sensors [39]. In academic settings, Romeo is used in labs to experiment with vision-based manipulation and human-like motion control, for example detecting and grasping door handles in dynamic environments [39].

Atlas (Boston Dynamics) – A high-mobility humanoid robot designed for agility and dynamic motion research, employing hydraulic actuators, lidar, and stereo vision [40]. Although not a teaching assistant, Atlas is widely referenced in robotics courses and public demonstrations as a benchmark for advanced locomotion, balance, and robot-environment interaction [40].

Toyota THR-3 – A teleoperated humanoid controlled via a motion-tracking suit and head-mounted display, with 5G-based low-latency operation [41, 42]. THR-3 serves as an experimental platform for embodied telepresence with potential applications in remote education, healthcare, and campus demonstrations [41, 42].

Capabilities: Social humanoid robots typically feature speech recognition and synthesis, gesture and facial recognition, touch sensors, cameras, and microphones, enabling multimodal interaction and basic emotion recognition [43, 31, 29]. Platforms like Pepper and NAO provide programmable APIs and visual programming environments for custom educational applications [43, 31], while more advanced robots such as Romeo and Atlas offer complex sensorimotor systems for real-time balancing, manipulation, and navigation [39, 40]. Robots like ARI and SIFROBOT integrate touchscreen interaction with facial tracking and multilingual conversational AI, and THR-3 introduces full-body teleoperation with high-speed wireless communication [32, 35, 41].

Use Cases in HE: Humanoid robots are used as teaching assistants, tutors, lab demonstrators, and administrative support tools. In classrooms, NAO and Pepper can lead quizzes, practice vocabulary, and provide interactive feedback, often boosting learner engagement and confidence in language and AI-related courses [30, 28]. ARI, SIFROBOT, and Temi serve as mobile guides and information kiosks in libraries or lobbies [35, 10], while research platforms such as Romeo and Atlas support advanced studies in locomotion, manipulation, and human-robot interaction [39, 40]. Sophia and THR-3 are often used as demonstration platforms or research subjects in courses on ethics, robotics, and telepresence [37, 41].

4.1.3 Service and Delivery Robots

Service robots are designed to perform practical tasks such as delivering items, guiding people, or maintaining facilities. In campus environments, they take on roles like couriering materials across buildings, assisting in libraries or dorms, or providing logistics support to

Figure 4. Closeup of Starship delivery robot. Image source: Wikimedia Commons, licensed under CC BY-SA 4.0 [44].

staff. These robots typically prioritize function over social interaction and often resemble carts or mobile boxes rather than humanoids. Examples in this category include:

Savioke Relay (Relay Robotics) – A compact autonomous delivery robot with a secure internal compartment for transporting small items such as food, mail, or lab samples. It uses elevators, onboard sensors, and navigation software to traverse hallways and deliver payloads point-to-point, and has been piloted on campuses for book and equipment delivery between departments or within libraries [45, 46].

Aethon TUG (T3) – An autonomous mobile robot widely deployed in healthcare and research facilities that resembles a motorized cabinet and can transport carts or bins across departments [47]. In teaching hospitals and large research institutes, TUG robots deliver meals, medications, and sensitive lab samples, supporting 24/7 logistics in shared spaces [47, 48].

Keenon Dinerbot T8 – A multifunctional autonomous service robot designed for hospitality and commercial environments, capable of delivering food and beverages, carrying luggage, and providing information. It uses an Android-based OS, integrated camera, voice commands, and sensors for obstacle detection, with up to 20 kg load capacity and 15 hours of battery life [49].

Care-O-bot 4 (Fraunhofer IPA / Mojin) – A general-purpose mobile manipulator with an omnidirectional base, expressive head, and optional arms for manipulating objects and interacting with users via touch display and gestures [50]. On campus, Care-O-bot can be configured to fetch books, deliver items, answer questions, or guide visitors in student centers and libraries [51].

Starship Technologies Delivery Robots – Small, autonomous wheeled robots widely

used on university campuses for contactless delivery of meals, groceries, and packages. Equipped with cameras, sensors, and ML-based navigation, they have supported essential deliveries during crises such as the COVID-19 pandemic and are now part of broader sustainable, AI-driven campus logistics [52, 53, 54].

Autonomous Campus Shuttles and Carts – Low-speed autonomous shuttles used to support sustainable mobility across large campuses. For example, the University of South Florida piloted a COAST Autonomous P1 shuttle with over 500 participants, reporting improved user trust and acceptance [55]; CETYS University integrates autonomous carts with IoT sensors and solar energy within a smart campus initiative to support accessibility and efficiency for students and staff [56].

Capabilities: Service and delivery robots typically feature autonomous navigation using SLAM or preloaded maps, along with obstacle detection via lidar, sonar, ultrasonic sensors, or depth cameras. Many offer payload management systems—secured compartments unlockable via app or PIN (e.g., Starship, Relay), open shelves (e.g., Keenon Dinerbot), or articulated arms (e.g., Care-O-bot), and can interface with elevators or doors to support end-to-end delivery [45, 50, 53]. Communication with users is often via synthesized voice, touchscreen displays, LEDs, or mobile alerts, and larger platforms like TUG or autonomous shuttles add remote monitoring and accessibility features [47, 55, 56].

Use Cases in HE: Service and delivery robots streamline campus logistics, improve accessibility, and reduce staff workload. Common deployments include food and grocery delivery to students, autonomous mail distribution, and library support (e.g., book returns and retrieval), as well as transport of clinical supplies and lab samples in academic medical centers [52, 53, 48]. Platforms like Care-O-bot can act as mobile info desks or event assistants [51], while autonomous shuttles and carts support sustainable transport for people and goods on large or distributed campuses [55, 56].

4.1.4 Collaborative and Laboratory Robots

This category includes robots that typically work in labs, workshops, or classrooms to assist with technical tasks rather than roaming campus hallways. They often have manipulator arms or tools for precise actions and operate alongside human students and staff, making them central to engineering education, research labs, and medical training. Examples include:

ABB YuMi – A dual-arm collaborative robot engineered for high-precision tasks alongside humans, featuring compliant joints and fine motor control. In HE, YuMi is used to teach

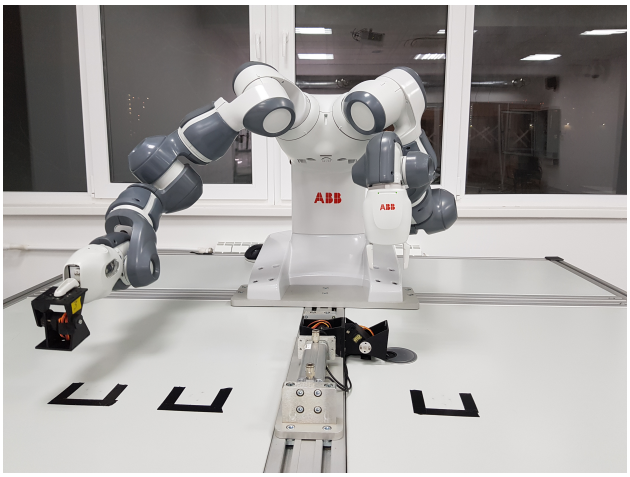

Figure 5. ABB YuMi collaborative dual-arm robot. Image source: Wikimedia Commons, licensed under CC BY-SA 4.0 [57].

motion planning, human-robot interaction, and collaborative assembly; at TalTech it has been showcased in interactive demonstrations and used in thesis work on nonlinear model predictive control for pick-and-place tasks [58, 59]. ABB has also proposed a mobile lab assistant concept integrating YuMi onto a navigable base for laboratory automation [60].

KUKA LBR iiwa and Universal Robots (UR) Cobots – Collaborative arms used widely in HE to teach industrial automation, control systems, and safe human-robot interaction. The KUKA LBR iiwa features torque sensors for compliant operation, while UR5/UR10 arms are valued for ease of programming, ROS compatibility, and accessible safety controls, supporting lab exercises in sorting, manipulation, and ROS-based control [61]. The KUKA youBot mobile manipulator further supports hands-on education in mobile manipulation, systems integration, and sensor fusion [62].

Nextage (Kawada Robotics) – A humanoid torso robot with two arms and a camera-equipped head, used in research labs for human-robot collaboration, dual-arm coordination, and light industrial tasks. Nextage has been used to demonstrate soft-object folding with deep learning (DCAE + TDNN) and to evaluate hardware-level reusable learning-from-observation systems that translate demonstrations into robot-agnostic symbolic sequences [63, 64].

Da Vinci Surgical Robot (Intuitive Surgical) – A multi-arm robotic system for minimally invasive surgery, teleoperated by a surgeon at a console with tremor filtering and fine motion control [65]. Medical schools and teaching hospitals use Da Vinci in simulation labs to train residents, and evaluations by SAGES confirm its safety and effectiveness for many procedures while noting higher costs compared to traditional methods [66, 65].

Baxter and Sawyer (Rethink Robotics) – Collaborative robots that were widely adopted in education before being discontinued, offering safe interaction, screen-based feedback, and programming by demonstration. Studies show Baxter provides reliable performance in

semi-structured tasks such as pick-and-place for educational experiments [67], and Sawyer has been used in dynamic HRI experiments where robots learn task sequences and timing from a single human demonstration [68].

Capabilities: Collaborative and laboratory robots emphasize precision, safety, and programmability, often using force-limited joints and collision detection to operate in shared workspaces. Many can be programmed by demonstration or through software interfaces, and high-resolution cameras and vision systems support object detection, pose estimation, and complex manipulation, as in Nextage and YuMi [63, 58]. Surgical systems like Da Vinci offer haptic feedback, tremor filtering, and VR-based simulation modules, while platforms such as KUKA youBot support real-time control through interfaces like EtherCAT for deeper integration into engineering education [62, 65, 60].

Use Cases in HE: Collaborative and lab robots are primarily used for technical education and research training. Engineering and computer science students use cobots such as UR5 or LBR iiwa in lab courses on vision-guided manipulation, dynamic task planning, and ROS-based control, while platforms like YuMi and Nextage support research in human-robot collaboration and manipulation [59, 63]. In medical education, systems like Da Vinci are used to simulate real-world procedures long before clinical practice [65, 66], and robots such as Baxter and Sawyer have been foundational in teaching collaborative robotics and safe physical interaction in academic labs [67, 68].

4.1.5 Quadrupedal and Animal-Inspired Robots

Figure 6. Boston Dynamics SpotMini quadruped robot. Image credit: © David Pérez (DPC), Wikimedia Commons, licensed under CC BY-SA 4.0 [69].

An emerging category in HE involves robots inspired by animals, such as four-legged "robot dogs" and other bio-mimetic designs. These robots often have unique mobility

advantages (e.g., traversing stairs or rough terrain), making them useful for campus security, research data collection, and novel forms of student engagement. Key examples include:

Boston Dynamics Spot – A widely recognized quadrupedal robot capable of walking, climbing stairs, recovering from disturbances, and carrying modular payloads. In academic settings, Spot is used as a mobile research platform for autonomous navigation, infrastructure inspection, and human-robot collaboration, including radiation surveys at Los Alamos National Laboratory and interdisciplinary courses on live human-robot interaction at Princeton [70, 71, 72]. Its 3D vision, SLAM navigation, and dynamic gait control make it suitable for both indoor and outdoor campus environments [73].

Unitree Quadrupeds (e.g., A1, Go1) – Smaller, lower-cost robot dogs that offer agile locomotion and are widely used in research and teaching because they are more affordable, open platforms. Recent work shows that Unitree robots can robustly execute locomotion under varied loads and terrain, including stable trotting with payloads up to 125% of nominal mass, supporting their role as versatile educational platforms for locomotion control and reinforcement learning studies [74].

Sony AIBO – A robotic dog originally introduced as an entertainment robot, with sensors, touch sensitivity, and voice interaction. In HE, AIBO has a strong legacy as a standard platform in RoboCup and as a tool to broaden participation in computer science, particularly among underrepresented groups, by providing an approachable and playful interface for programming and social robotics [75]. Universities use AIBO in outreach, introductory CS courses, and social AI research.

Other Animal-inspired Robots – This category also includes snake-like robots, aquatic robots modeled after fish, aerial robots with bird-like wings, hexapods, robotic cheetahs, and robotic bats. These platforms support research into biomechanics, adaptive control, and locomotion in environments where traditional wheeled robots are less effective, and are used in university projects spanning environmental monitoring, facility inspection, and bio-inspired engineering education [76].

Capabilities: Animal-inspired robots in HE combine mechanical agility, autonomy, and often human-centric interaction design. Quadrupeds like Spot and Unitree's A1/Go1 use force sensors, IMUs, and vision-based SLAM to maintain balance and situational awareness, support significant payloads, and self-correct after falls, enabling reliable operation indoors and outdoors [71, 74]. Social robots such as AIBO emphasize emotional engagement through touch and voice responsiveness and adaptive behaviors, supporting education and HRI, while other bio-inspired platforms (fish, birds, insects) enable exploration of specialized locomotion paradigms such as slithering, gliding, or aquatic propulsion [75, 76].

Use Cases in HE: Quadrupeds like Spot are used for autonomous patrol, facility inspection, and environmental data collection (e.g., radiation, air quality, and 3D mapping), often in hazardous or hard-to-reach areas [70, 71]. They also serve as platforms for coursework on human-robot coexistence and reinforcement learning, including choreographed performances and social interaction experiments [72, 74]. Companion robots like AIBO support outreach and introductory programming, helping broaden participation in computer science and offering robotic pet therapy in student spaces [75]. More broadly, animal-inspired robots allow universities to investigate bio-mimetic engineering while enhancing research, safety, and student engagement on campus [76].

4.2 Cybersecurity Challenges Associated with Robot Assistants

Introducing robot assistants in academia raises concenrs that must be addressed to ensure responsible use. This section elaborates these challenges to derive concrete requirements for the proposed framework and its guidelines. Figure 7 summarizes the eight cybersecurity challenge areas that structure Sections 4.2.1–4.2.8:

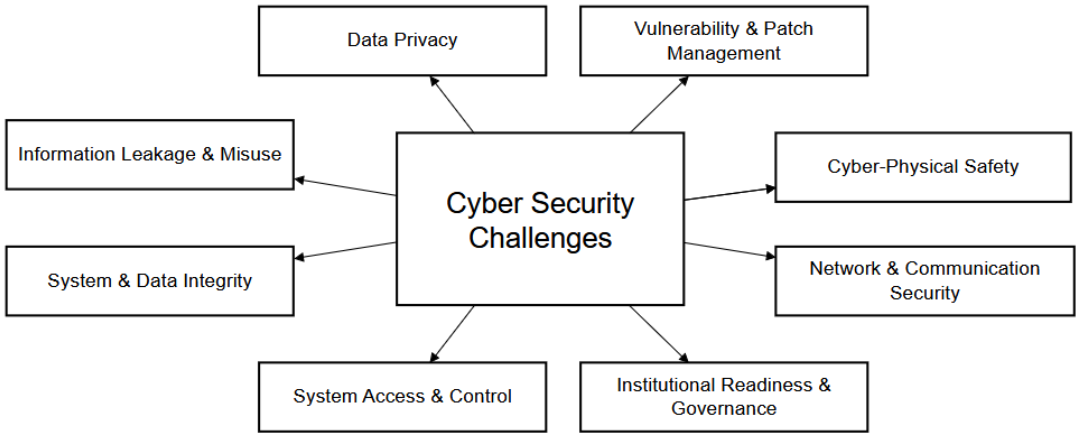

Figure 7. Cybersecurity challenge domains considered in this thesis

4.2.1 Data Privacy and Confidentiality

As intelligent robotic systems become increasingly embedded in the digital and physical fabric of HE, they generate and interact with large volumes of sensitive data. These datasets often include personally identifiable information (PII), academic records, biometric identifiers, behavioral interaction logs, and real-time audio-visual streams. Robot assistants such as *Pepper*, *Temi*, or virtual agents like *Jill Watson* are capable of processing data related to attendance, quiz responses, facial recognition, or even emotion detection each of

which carries distinct privacy risks.

The exposure surface of these systems is significantly broader than traditional software tools. A telepresence robot like *Double 3* or *Ohmni* not only transmits high-definition video across networks but may also store logs of past sessions, stream to cloud platforms, or remain connected unintentionally. Focus group studies reveal that participants fear unregulated recordings, surveillance without notice, and misuse of visual/audio data in telepresence contexts [77]. These risks are amplified when institutions lack clear user notification and consent mechanisms for robotic data processing [78].

Encryption at rest and in transit is a foundational requirement, but technical implementations vary. Systems should employ modern cryptographic protocols—such as AES-256 for data at rest and TLS 1.3 for transmission—but not all vendors meet these standards by default. A systematic security review of commercially deployed social robots has revealed issues such as the use of unencrypted communication channels, open administrative ports, and unsecured Wi-Fi protocols—exposing sensitive data to passive attackers or man-in-the-middle threats [78].

The EU Artificial Intelligence Act (AI Act) reinforces these concerns by codifying strict requirements for data governance and quality. High-risk AI systems—such as educational robots that influence student learning, performance, or well-being—must ensure their training, validation, and testing datasets are relevant, representative, complete, and as error-free as possible for their intended purpose (EU AI Act, Recital 67, p. 19) [79]. Additionally, providers must document the origin of data and its original collection purpose when personal data is involved (EU AI Act, Article 10(2), p. 57) [79], ensuring data usage aligns with the context in which it was collected.

Access control is equally critical. Role-based mechanisms should ensure that only authorized faculty or IT administrators can view sensitive data, with audit logs maintained. Robots like *NAO* or *ARI* that utilize facial recognition or voice templates must handle biometric data with elevated care. The AI Act mandates that if special categories of personal data (e.g., ethnicity, health indicators) are ever used—for example, to detect and correct systemic bias—such use must be demonstrably necessary and subject to strict access controls, confidentiality obligations, and robust documentation (EU AI Act, Article 10(5)(a)–(c), p. 58) [79].

Importantly, the Act outright prohibits certain types of AI-based biometric surveillance. For instance, using AI to infer emotions in education contexts is banned unless specifically permitted for safety or medical purposes (EU AI Act, Recital 43) [79]. This directly affects

humanoid classroom robots with affective computing capabilities, limiting their use of emotion recognition tools. Such protections help mitigate mass surveillance risks and preserve student dignity.

Furthermore, the AI Act mandates that data-intensive AI systems adhere to the GDPR and broader EU privacy frameworks, clarifying that the Act does not override existing personal data protections (EU AI Act, Recital 7, p. 3) [79]. Institutions using AI-based learning tools must therefore perform Privacy Impact Assessments (PIAs), enforce minimal data retention periods, and prefer anonymized or synthetic data unless real personal data is strictly necessary (EU AI Act, Article 10(5), p. 58) [79].

Finally, institutional compliance must extend beyond vendors. Universities must establish robust governance frameworks, including data classification rules, automated retention policies, breach notification procedures, and vendor oversight protocols [80]. Without such infrastructure, systems handling academic performance, health details, or student interactions risk violating GDPR or FERPA. In sum, the AI Act elevates privacy from a soft ethical guideline to a legal obligation: educational AI systems must be designed and deployed in ways that rigorously protect personal data and prevent misuse or unwarranted surveillance of students.

4.2.2 System and Data Integrity

The integrity of robotic systems in educational environments is critical to ensuring reliable and trustworthy operations. Robot assistants rely on intricate combinations of software, firmware, and machine learning models to perform tasks ranging from autonomous navigation to student engagement and assessment. Any compromise in these layers can lead to data corruption, unauthorized behavior, or complete system failure—jeopardizing not only academic workflows but also institutional trust.

Many commercial and educational robots—such as *Pepper*, *NAO*, and *Temi*—run on general-purpose operating systems like Linux or Android. While these platforms support broad compatibility and developer accessibility, they also present familiar attack surfaces. Robots deployed with default settings, legacy services, or unpatched kernels are vulnerable to known exploits. For instance, a security analysis revealed that due to missing integrity checks in its update process, a modified Temi Android app could be used by attackers to gain remote privilege escalation on the robot [81]. Similarly, the Pepper robot has been found to expose unprotected network services, insecure remote access features, and open telnet/SSH ports—allowing attackers to manipulate its software and underlying OS remotely [82].

Firmware update mechanisms are another common vector for compromise. While many educational robots support over-the-air (OTA) or USB-based updates for classroom convenience, these mechanisms are not always protected. A widely used hospital logistics robot (Aethon TUG) was found vulnerable to remote hijacking due to insecure update and configuration channels, potentially enabling attackers to override navigation or steal operational data [83]. These findings echo broader issues in the IoT ecosystem, where systems often suffer from long patch cycles, exposed debugging interfaces, or outdated libraries—flaws that are equally present in networked educational robots [84].

Middleware vulnerabilities are also of concern. Educational robots increasingly rely on Robot Operating System (ROS) or similar middleware, which—while flexible—has exhibited critical security weaknesses. ROS 2's default DDS (Data Distribution Service) layer has been shown to lack authentication, enabling attackers to inject false commands, manipulate sensor data, or access unauthenticated interfaces [85]. A 2022 investigation found 15 exploitable ROS/DDS vulnerabilities affecting academic and industrial robots, prompting calls for immediate adoption of secure variants like SROS2.

Ensuring the technical integrity and cybersecurity of AI-driven robot assistants is also a legal requirement under the EU Artificial Intelligence Act (AI Act). High-risk AI systems—such as those influencing student safety, rights, or academic outcomes—must be designed and developed to ensure appropriate levels of accuracy, robustness, and cybersecurity throughout their lifecycle (EU AI Act, Article 15(1), p. 54) [79]. This includes the ability to handle errors, unexpected inputs, or environmental anomalies without unsafe or unpredictable behavior (Article 15(4), p. 55) [79]. Robots operating in classroom environments must incorporate fail-safes and fallback mechanisms to prevent escalation of faults, and may require interlocks or override features to maintain safety during component failures.

Furthermore, the AI Act explicitly requires resilience against intentional attacks. Article 15(5) mandates that high-risk AI systems be protected against unauthorized alteration of their performance, including AI-specific threats like data poisoning, model tampering, and adversarial examples. This aligns with real-world cases where robotic systems have been hijacked, manipulated, or degraded due to security lapses. It underscores the necessity of strong safeguards to prevent a student, insider, or external actor from exploiting a robot to behave unsafely or access sensitive information.

To mitigate these risks, institutions must enforce secure update pipelines and runtime verification. Cryptographic signing of firmware and software should be mandatory, with all updates delivered via authenticated channels. Robots should perform integrity checks

at boot and runtime using hash validation, secure boot processes, and application-level sandboxing. Vulnerability assessments on platforms like Pepper show that absent such controls, unauthorized changes to robotic behavior—including camera access or motion control—are entirely feasible [82].

While various standards such as ISO/SAE 21434 and IEC 62443 have been developed for automotive and industrial systems respectively, their underlying principles—such as structured patch management, intrusion detection, and subsystem isolation—can inform best practices in the educational robotics domain [78]. For example, a compromised chatbot interface or mobile API should not be able to access critical components such as grading engines or identity modules. This paper builds upon such foundational ideas while developing a tailored framework specifically for HE environments.

Auditability is equally important. All critical operations—such as student data access, system reconfigurations, and AI-driven decision overrides—should be logged in tamper-evident formats. These logs provide the forensic trail needed during investigations or accountability reviews, helping determine whether integrity has been violated and by whom.

Finally, system partitioning and sandboxing must be considered essential design features. Sensitive components like authentication modules, grading engines, and attendance tracking must be isolated from general-purpose interfaces. This prevents a compromise in one subsystem (e.g., a user-facing chatbot or mobile API) from cascading into deeper layers of control.

The AI Act also addresses broader obligations for providers of general-purpose AI models—such as those powering robotic conversational assistants. Article 55(d) [79] requires that such models be secured along with their physical and digital infrastructure, especially if designated as carrying "systemic risk." This offers downstream protection for educational institutions, as upstream providers are compelled to deliver hardened components and report serious incidents.

As educational institutions adopt increasingly complex robotic systems, system and data integrity become foundational to operational trust, student safety, and legal compliance. Without robust technical safeguards, architectural isolation, and adherence to regulatory standards, even the most advanced learning assistant may become a liability. The AI Act reinforces these concerns by embedding cybersecurity and resilience as essential conditions for any high-risk AI application in education.

4.2.3 Information Leakage and Misuse

Even without direct exploitation or malicious interference, robot assistants can become inadvertent vectors of sensitive information leakage if not configured or used correctly. Operating in semi-autonomous and persistent modes, these systems often remain active across multiple user interactions, increasing the risk of unintentional data disclosure. A high-risk scenario, for example, involves telepresence robots that continue streaming after sessions end—accidentally capturing private conversations, screen content, or student behaviors [86].

Social and service robots further challenge user privacy through data personalization and persistent recognition. As explored by Reig et al. [87], users are more accepting of personalization when its benefits are clear, but many perceive identification and data reuse across roles as invasive, especially when robots re-embody or change functions without explicit user consent. In institutional settings like universities, these tensions are heightened due to a mix of formal and informal interaction contexts where expectations of both informational and psychological privacy are deeply rooted.

Design factors also influence the degree of perceived data risk. According to Heilmann et al. [88], students and educators prefer robots that provide physical cues—such as lights or screen prompts—when active or recording, and show a preference for devices that allow local data processing over cloud-based storage. These preferences point to a broader demand for transparency and perceived control, especially when robotic systems are embedded in educational environments.

Additionally, poorly configured chatbots or AI-based campus assistants can leak data in response to ambiguous queries—such as inadvertently exposing student grades, schedules, or administrative contacts [89]. Without robust filtering or context-aware query handling, systems may respond with information intended only for authenticated users. Such incidents also underline a critical issue in human-robot communication: systems can unintentionally violate contextual norms due to a lack of situated understanding.

Unauthorized misuse adds another layer of concern. Attackers exploiting telepresence links or voice-enabled robots may impersonate legitimate users to access sensitive institutional data or manipulate communication channels. These "passive compromise" scenarios, while subtle, fall outside many traditional security assessments and remain underaddressed [86]. Moreover, as Zhu et al. [90] emphasize, privacy in human-robot interaction spans several overlapping constructs—including informational, physical, and social privacy—which are often insufficiently distinguished in system design. Failing to address this diversity can

result in perceived intrusions even in the absence of technical vulnerabilities.

Mitigating such risks requires a multifaceted approach. Robots should enforce conservative privacy defaults: disabling cameras and microphones by default, implementing auto-logouts, and anonymizing or obfuscating data wherever feasible. User-facing controls must be available to terminate sessions, mute microphones, or revoke permissions in real time. Furthermore, natural language output must be filtered to ensure that personal or institutional data is not disclosed in generic responses.

Training and policy alignment are equally essential. As proposed in the CONFIDANT system by Fritsch et al. [89], integrating user-driven privacy control mechanisms and contextual awareness into the robot's design fosters responsible behavior. Institutions must extend cybersecurity awareness campaigns to include robot-specific risks such as re-embodiment, phishing via chatbots, or unexpected surveillance. Educational stakeholders must also be sensitized to the nuances of privacy perception, acknowledging that design preferences are not just technical decisions but ethical commitments to user autonomy.

Ultimately, institutional governance must evolve. Robots and AI assistants should be fully integrated into data protection policies, including documentation on what types of data may be collected, how long it is stored, who has access, and which legal obligations (e.g., GDPR, FERPA) apply. Without these policies, even well-intentioned deployments may lead to significant privacy breaches or reputational damage.

4.2.4 System Access and Control

Robot assistants introduce unique access vectors into institutional digital infrastructures and physical environments. Unlike traditional networked devices, robots often possess actuators, mobility systems, and live audio-visual capabilities—making unauthorized access not only a privacy risk but also a potential safety threat. Attackers may target robot systems to gain unauthorized control over their behaviors, interfaces, or data channels. Such breaches can arise from weak authentication mechanisms, unprotected network endpoints, or misuse of default configurations commonly found in off-the-shelf robotic platforms.

Real-world security analyses have demonstrated that many commercially available educational and service robots ship with dangerously weak access controls. For instance, McAfee's investigation into the Temi robot uncovered multiple critical vulnerabilities, including hardcoded credentials (CVE-2020-16170), missing authentication for critical functions (CVE-2020-16167), and authentication bypass using alternate paths (CVE-2020-

16169). These flaws allowed attackers to remotely operate Temi, intercept video calls, and access sensitive data without any authentication [91].

Similarly, Alias Robotics' penetration testing of SoftBank's Pepper robot revealed significant security issues, such as default root credentials and unsecured web interfaces. These vulnerabilities enabled unauthorized users to manipulate Pepper's behavior and access its systems, posing risks of eavesdropping and physical disruption [92].

Once compromised, attackers can misuse the robot's physical or audiovisual capabilities. With root access, malicious users may repurpose the robot to harass others, trigger alarms, or navigate into restricted areas. As noted by researchers, poorly secured robots can effectively be weaponized into "cyber-physical threat agents." Similar concerns are echoed in studies of multimodal assistive robots like AMRSys, where improper handling of sensor data and command privileges can compromise both safety and privacy in sensitive environments [93].

To address these threats, robots must be equipped with robust access control policies. This includes enforcing session-based access tokens, user-specific authentication, and ideally, two-factor authentication (2FA) for administrative privileges. Meeting links and command sessions should expire automatically, and robots should be logically segmented via VLANs or dedicated Wi-Fi SSIDs to minimize lateral movement in the event of a breach. Critical interfaces—such as USB debug ports, web dashboards, and developer APIs—must be hardened or disabled entirely unless explicitly needed.

Role-based access control (RBAC) should be implemented to differentiate privileges between user roles (e.g., instructor, technician, student). According to recent trust model research, systems that account for context-aware authorization—adjusting access permissions based on trust scores, environment, or user-device behavior—provide stronger safeguards than static credential checks alone [94]. Integration with institutional IAM frameworks should support features such as credential rotation, role inheritance, and termination of orphaned accounts.

Access monitoring is equally vital. As emphasized in both industrial practice and emotion-aware security research [95], real-time auditing and behavioral anomaly detection help flag unusual interactions or brute-force login attempts. Access logs should be cross-referenced with timestamps, device IDs, and geolocation metadata to identify suspicious activity patterns.

Finally, as robotic infrastructure increasingly relies on cloud-based platforms [96], institu-

tions must ensure that authentication protocols, encryption standards, and regional data residency policies are consistently applied across both on-premises and remote control endpoints. Robots must be treated not as novelty devices but as operationally critical endpoints in institutional threat models—subject to the same security expectations as servers or IoT gateways. Without such parity, classroom environments remain vulnerable to both digital intrusion and physical compromise.

4.2.5 Vulnerability and Patch Management

Robot assistants deployed in HE frequently rely on a heterogeneous mix of software frameworks, proprietary firmware, and consumer-grade operating systems. These components—while essential for enabling AI integration, sensor communication, and user interactivity—can also introduce significant cybersecurity risks if not properly maintained. Many educational robots ship with embedded Linux or Android variants and custom firmware stacks that include outdated packages, insecure libraries, and publicly known vulnerabilities [97].

A deeper concern lies in the underlying development culture within the robotics field. An empirical study by Fernandes et al. [98] found that 76% of robotics practitioners surveyed had never performed formal security testing on their systems, and nearly half believed cyberattacks against robots were unlikely to occur in practice. This lack of perceived risk results in real-world vulnerabilities remaining unpatched across many robots deployed in academic settings, creating long-lived exposure windows. Developers often prioritize functionality over security and delay patching due to limited institutional mandates or vendor oversight.

This situation is exacerbated by the absence of coordinated vulnerability disclosure practices. Unlike conventional IT ecosystems where patch cycles align with vulnerability announcements, robotics vendors tend to lag behind in addressing known flaws. The Robot Vulnerability Database (RVD) [99] was established in response to this gap. It catalogues robot-specific CVEs and aims to pressure manufacturers into issuing timely security fixes. However, the effectiveness of this initiative is still limited by the inconsistent participation of vendors and the lack of regulatory enforcement, leaving educational institutions burdened with monitoring and mitigating these issues independently.

Compounding the problem is the widespread use of the Robot Operating System (ROS), which, although widely adopted in research and education for its flexibility, was not designed with security as a foundational principle. As Dieber et al. [100] highlight, ROS lacks basic access control, message authentication, and encryption mechanisms. While

frameworks like SROS and ROS2 address many of these shortcomings, their uptake in educational deployments remains minimal. Consequently, ROS-based systems remain vulnerable to attack vectors such as topic hijacking, man-in-the-middle interception, and unauthorized execution of control commands [101].

To mitigate these risks, universities should develop structured vulnerability and patch management frameworks tailored to robotics. This includes subscribing to sources like RVD for vulnerability notifications, maintaining a formal update cadence based on severity disclosures, and verifying cryptographic signatures on firmware or software before installation. Robots should support secure over-the-air (OTA) updates, or alternatively, enable secure manual update procedures with version validation and change logging.

Routine vulnerability scanning—using tools like OpenVAS, vendor-supplied scanners, or ROS-specific introspection utilities—should be mandatory. In parallel, institutions should manage firmware baselines and software states within a configuration management database (CMDB) or version-control system, helping detect unauthorized modifications or version drift. Security-critical components such as data processors, AI inference modules, and networking stacks should be sandboxed and continuously monitored for anomalies.

Finally, robots running unsupported or end-of-life (EOL) operating systems should be phased out of active use in sensitive environments. Procurement decisions must favor vendors that offer transparent update cycles and long-term patch support. By embedding patch management into the full lifecycle—from acquisition to decommissioning—academic institutions can reduce the attack surface of their robotic deployments and safeguard both digital and physical campus infrastructure.

4.2.6 Cyber-Physical Safety

As robot assistants increasingly operate in shared human environments—equipped with mobility platforms, actuators, and manipulators—their presence introduces new dimensions of risk beyond traditional digital vulnerabilities. These systems are no longer confined to screens or passive observation; they traverse hallways, assist in laboratory tasks, and interact physically with students and staff. Consequently, cyber-physical safety concerns emerge at the intersection of mechanical behavior, software reliability, and environmental unpredictability [102].

Unlike static digital tools, mobile robots like Savioke Relay and Aethon TUG must navigate dynamic, often congested academic spaces. Recent research highlights that collision risk increases in areas where students move unpredictably or group density

changes rapidly [102]. Safety issues are not only technical but spatial and behavioral. For instance, a delivery robot failing to respond to a sudden obstacle in a hallway—be it a dropped item or a wheelchair user—may block passageways or trigger minor accidents.

Even in controlled lab environments, collaborative robots such as ABB YuMi or Universal Robots cobots present tangible safety challenges. Improper configuration or sensor miscalibration can lead to joint overshoot or contact at unsafe force thresholds [103]. The educational setting adds complexity: students are still learning safe interaction boundaries, and instructors may not be trained to handle robotic safety faults in real-time.

Malfunctioning or misused telepresence platforms—such as Temi or Care-O-bot—can also create harm. Although lighter and less forceful, their autonomous navigation routines can misinterpret visual input, fail to identify edge cases (e.g., seated individuals), or continue motion during remote disconnections. As highlighted by Dieber et al. [100], poor modular separation between movement and sensor input in ROS-based systems can cause navigation failures to cascade into physical hazards.

To mitigate these risks, institutions must enforce safety-by-design principles. Robots deployed in any shared environment should feature hardware kill-switches, remote shutdown interfaces, and reliable user authentication for control operations. Movement commands should be sandboxed through safety-layer verification, and control subsystems (e.g., drive motors, arm actuators) isolated from higher-level, potentially untrusted, applications.

Simulation-based motion verification is essential prior to deployment. As recommended in the lab safety literature [103], robots should be tested in confined testbeds to validate joint limits, torque ranges, and proximity sensors under near-realistic workloads. Telemetry data—including actuator strain, acceleration patterns, and emergency stops—must be logged and reviewed regularly for incident prediction or post-hoc analysis.

Cyber-physical safety also extends into perceptual and psychological space. As Schieb et al. [104] show, even robots that do not make physical contact can cause discomfort through proximity, sudden motion, or unclear intent. Perceived safety is a significant factor in acceptance, especially among vulnerable groups or in unfamiliar classroom scenarios. Response protocols must therefore account for not only mechanical malfunctions but also user discomfort and anxiety.

Ultimately, cyber-physical safety is foundational to the sustainable integration of robot assistants in HE. From navigation to manipulation, and from system design to psychological impact, all aspects of robotic behavior must be scrutinized through the lens of human

safety, ethical responsibility, and institutional accountability. Without robust safeguards, these technologies risk becoming sources of physical disruption, reputational harm, or legal liability.

4.2.7 Network and Communication Security

Robot assistants in HE environments function as always-connected devices, interfacing with Wi-Fi networks, cloud services, campus APIs, and peripheral systems such as learning platforms or digital signage. While these capabilities enable real-time responsiveness and remote interaction, they also open critical communication pathways that, if improperly secured, pose serious risks to both data integrity and broader institutional networks.

A key vulnerability stems from the architectural decisions behind many robotic frameworks. The Robot Operating System (ROS), widely adopted in both research and commercial robotics, was originally designed without basic security features such as encryption or authentication. A 2019 global scan by researchers revealed over 100 robots running ROS exposed directly to the internet, including educational units with active sensors and actuator access [105]. These nodes were left unprotected due to the ROS master's lack of access control, allowing unauthenticated users to take control of the robot, intercept data streams, or inject malicious commands. In one documented case, attackers were able to remotely move a robot arm in a university lab and stream its sensor data without any form of verification [105].

Even with the advent of Secure ROS (SROS) and improved middleware in ROS 2, adoption remains limited in academic settings, often due to technical complexity or compatibility constraints [106]. Without these security enhancements, robot control topics, service calls, and parameter servers remain exposed to sniffing, spoofing, or denial-of-service attacks. Moreover, many robots rely on WebSocket APIs, MQTT brokers, or proprietary peer-to-peer communication layers, which often lack rigorous authentication or encryption—leaving them vulnerable to session hijacking and man-in-the-middle (MITM) attacks [107].

The issue extends beyond ROS. Many commercial educational robots use insecure firmware that communicates over open ports or uses hard-coded credentials to connect with backend services. These robots are often deployed on shared campus Wi-Fi, where a compromised client can scan for and reach robot endpoints. This creates lateral movement opportunities, allowing attackers to pivot from a single misconfigured robot to critical IT infrastructure [108].

To counter these risks, institutions must treat robots as high-risk cyber-physical nodes within their threat model. Network segmentation is essential: robots should operate on dedicated VLANs or logically isolated SSIDs, with strict firewall rules governing traffic to and from only necessary endpoints. TLS encryption must be enforced for all remote communication, especially for control APIs, firmware updates, and telemetry channels. Endpoint validation mechanisms, such as certificate pinning and mutual TLS, should be used to prevent impersonation and MITM scenarios.

Beyond static defenses, behavioral monitoring and anomaly detection are critical. Robots sending unexpected payloads, communicating with unknown domains, or initiating large data transfers during off-hours should raise automated alerts. Intrusion detection systems (IDS), traffic rate limiting, and authentication throttling can help mitigate abuse. Institutions should also schedule periodic penetration tests targeting robot platforms, simulating exploits such as rogue device insertion or protocol fuzzing, to proactively identify attack surfaces.

Modern robotics cybersecurity frameworks are converging around zero-trust principles: no device should be assumed safe solely based on network location or vendor assurance. Instead, continuous verification, isolation, and policy enforcement are essential. As campus robots increasingly integrate with student records, learning systems, and physical navigation, robust communication security is not an optional enhancement—it is a baseline requirement to ensure safety, privacy, and institutional resilience.

4.2.8 Institutional Readiness and Governance

As HE institutions adopt increasingly complex AI-driven robotic systems, traditional IT governance models must evolve to address emerging ethical, legal, and cybersecurity risks. Institutional readiness is no longer about incremental policy tweaks—it demands a cohesive strategy that spans procurement, deployment, oversight, and liability attribution.

Current governance gaps are significant. Many universities lack clear frameworks regulating the use of robot assistants in classrooms, laboratories, or administrative spaces. This often results in robotic deployments occurring without centralized approval, risk assessment, or integration into institutional cybersecurity and ethics protocols. As Villaronga emphasizes in the healthcare context, robotic systems require a shift from broad, principle-based policies to risk-based impact assessments tailored to specific deployments [109]. Similar risk assessment frameworks should be adopted in academia to evaluate physical safety, autonomy suppression, data collection practices, and social impacts prior to robot adoption.

The EU AI Act reinforces this need for systemic oversight by mandating that institutions (as deployers of high-risk AI systems) maintain clear governance mechanisms, including assignment of roles, documentation procedures, and compliance reporting (EU AI Act, Art. 29, p. 47; Art. 70, p. 94) [79]. Member States must designate competent authorities and single points of contact, with institutions expected to align their oversight procedures with national and EU-level frameworks (EU AI Act, Art. 70(1–3), p. 94) [79].

Clear institutional ownership is fundamental. Decision rights over procurement, configuration, maintenance, and monitoring of robotic systems should not be left to individual faculty or ad hoc committees. As emphasized by Gräf et al. in their study of AI governance in HE, universities must designate responsible bodies to oversee compliance, accountability, and human oversight in all autonomous system deployments [110]. This may include central ethics committees, IT security teams, and legal advisors trained in AI and robotics governance.

Governance readiness must also extend into the legal domain. Guerra et al. note that robots challenge traditional tort frameworks, especially in cases of autonomous behavior leading to harm. The "responsibility gap" arises when it becomes difficult to assign blame between the robot, its operator, or the institution that deployed it [111]. HE institutions must develop internal liability models that anticipate these gaps. This includes clarifying who is accountable in case of surveillance misuse, physical harm, or AI-driven decisions that adversely affect students or faculty.

The EU AI Act further anticipates such concerns by requiring institutions to implement post-market monitoring mechanisms for high-risk AI systems, enabling proactive detection of system failures or non-compliant behavior throughout their lifecycle (EU AI Act, Art. 72(1–3), p. 101) [79]. In the event of serious incidents—such as physical harm or significant rights violations—deployers must report to market surveillance authorities within tight deadlines (EU AI Act, Art. 73(1–4), p. 102) [79]. Universities, especially public institutions, will need to develop internal processes to comply with these timelines and reporting expectations.

Moreover, cross-disciplinary training and incident protocols are critical. Instructors, lab personnel, and administrative staff must be taught how to identify anomalous robotic behavior, respond to hardware or software malfunctions, and escalate issues through well-defined response channels. Drills and tabletop exercises—simulating robot misbehavior, AI bias events, or system compromise—should be institutionalized, reflecting the scenario-based preparation advocated in emerging AI governance models [110]. The AI Act supports this with provisions allowing real-world testing of high-risk systems under supervised condi-

tions, but only when such tests comply with strict safety, transparency, and documentation requirements (EU AI Act, Art. 76(1–3), p. 108) [79].

Finally, collaboration beyond the university is key. As Villaronga argues in the context of healthcare robotics, regulatory innovation must accompany technical progress. Universities must therefore participate in inter-institutional working groups, contribute to standard-setting efforts, and adopt best practices from domains where robotic governance is more mature [109]. Article 74 of the AI Act specifically encourages coordinated joint investigations and knowledge sharing among competent authorities, with the aim of strengthening market oversight and compliance across multiple Member States (EU AI Act, Art. 74(11), p. 105) [79].

Legal scholars have also proposed assigning robots a form of limited legal personhood or contractual accountability to bridge liability gaps in robotic torts—a proposition that may eventually shape university policy when robots act with high levels of autonomy [111]. Even absent such formal recognition, universities must adopt legal models that capture the nuances of autonomous behavior and its consequences within institutional settings.

In sum, governance of educational robot systems must mature in tandem with their deployment. Rather than viewing robots as isolated gadgets, institutions should treat them as embedded cyber-physical actors—subject to the same scrutiny, legal preparation, and ethical oversight as any critical infrastructure. A robust governance framework ensures not only operational safety but also long-term trust in the university's ability to harness automation responsibly.

4.3 Ethical Challenges Associated with Robot Assistants

Alongside cybersecurity, there exists ethical concerns when deploying robot assistants in HE. Key ethical challenges include:

4.3.1 Privacy, Consent and Surveillance

The integration of robot assistants into HE introduces profound shifts in the nature and scope of surveillance within academic environments. These systems are equipped with a wide range of sensors—including cameras, microphones, depth sensors, and GPS modules—to support telepresence, context-aware navigation, and interactive services. While these capabilities enhance engagement and accessibility, they also blur the boundaries between pedagogical assistance and continuous monitoring, raising serious concerns

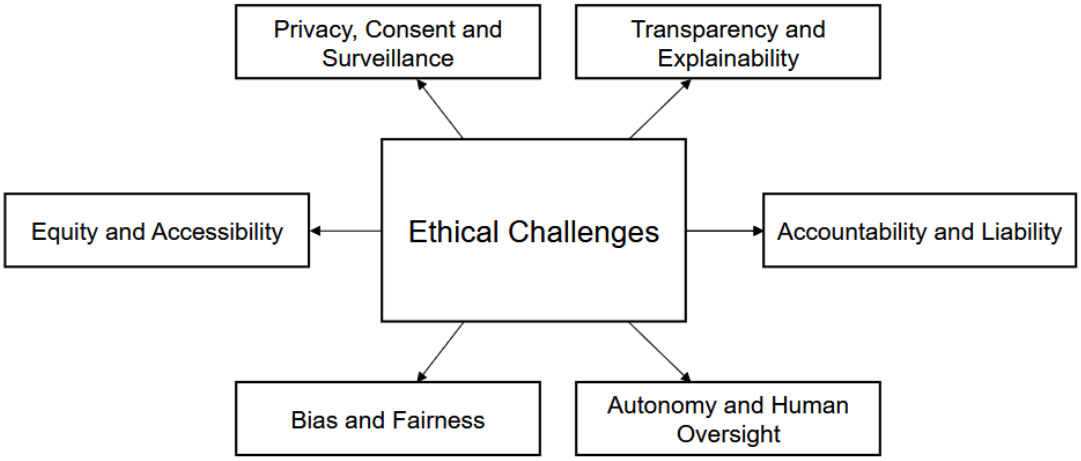

Figure 8. Ethical challenge domains considered in this thesis

regarding autonomy, transparency, and informed consent [93].

As highlighted in recent legal and sociotechnical analyses [112, 113], robot assistants do not merely process information—they mediate space, behavior, and power. Telepresence devices such as *Temi*, *Double 3*, and *Ohmni* are increasingly used for inclusive learning, yet students have reported feeling "watched" or uncomfortable when unsure who controls the robot or whether it is recording [114]. Unlike traditional surveillance cameras, mobile robots introduce dynamic, persistent observation, often lacking clear visual cues about their operational state. These perceptions—regardless of whether recording is actually taking place—can erode the sense of psychological and social privacy necessary for safe and equitable learning environments.

Similarly, delivery and service robots such as *Starship* or *Savioke Relay*, though not designed primarily for observation, may inadvertently collect environmental or conversational data through their autonomous navigation systems. In shared spaces like dormitories or health clinics, even passive data collection may capture private behavior, incidental speech, or location patterns. According to Marchang and Di Nuovo [93], multimodal assistive robots also present heightened risk due to their ability to simultaneously process facial, vocal, and gesture-based inputs—raising concerns about biometric profiling and the long-term storage of sensitive data.

More critically, humanoid robots like *NAO*, *Pepper*, and *ARI* challenge traditional legal notions of surveillance and responsibility. As Cardiell argues, their human-like embodiment and social interface capabilities invoke not only informational privacy concerns, but also dimensions of emotional, bodily, and relational privacy [112]. These robots' capacity for

facial recognition, emotional inference, and adaptive behavior introduces an "affective surveillance" dynamic, where users may feel compelled to modify their behavior even without active data collection.

To address these concerns, privacy-by-design must become a foundational principle in educational robotics. Robots should collect only the minimum data required for their function, and must make their operational status visible through LED indicators, audio prompts, or on-screen notices. Microphones and cameras must be opt-in rather than default-on, with session-based controls and clear expiration mechanisms. These requirements are not newly introduced by the EU AI Act, which instead reinforces existing obligations under the GDPR—particularly for systems capable of biometric processing or behavioral analysis. Compliance with GDPR remains the primary legal foundation for privacy and data protection in robot deployments.

Consent must be conceptualized as dynamic and context-specific. As shown in both educational and clinical settings, users often underestimate the scope of data being collected by social robots [115]. Institutions must implement repeatable, explainable consent procedures—not just checkbox agreements at deployment. Policies should differentiate between passive background logging and active recording, clarify data retention periods, and ensure that access to logs is tightly controlled.

Finally, universities must go beyond technical mitigation and implement institutional governance for robotic surveillance. Data access must be logged and monitored, and responsibilities for privacy breaches must be clearly assigned. Institutions deploying robots capable of persistent observation should conduct impact assessments with stakeholder input—including students, accessibility advocates, and legal advisors. These assessments should consider not only compliance with legal mandates, but also broader questions of academic freedom, psychological safety, and equitable participation [93, 115].

In sum, surveillance by robot assistants is not a narrow issue of camera placement or encryption—it is a reconfiguration of the classroom's ethical and legal fabric. Safeguarding consent, minimizing overreach, and ensuring trust are essential not only to protect rights, but to enable the ethical and sustainable use of robotic technologies in HE.

4.3.2 Transparency and Explainability

A foundational principle in the ethical deployment of robot assistants in HE is transparency—both in terms of the system's identity and the mechanisms driving its decision-making processes. As AI-driven assistants become increasingly sophisticated, the boundary

between human and machine interaction can blur, potentially leading to confusion, over-attribution of agency, or even unintentional deception. This is particularly salient in educational contexts, where students rely on feedback, tutoring, and guidance that may be shaped by opaque algorithmic decisions.

One illustrative example is the case of "Jill Watson," a virtual teaching assistant deployed at Georgia Tech. Initially embedded into online forums without disclosing its non-human identity, the agent successfully interacted with hundreds of students before its nature was revealed. While the deployment demonstrated the feasibility and efficiency of AI-driven assistance, it also raised critical questions about disclosure, informed consent, and ethical design [116]. The project sparked broader debate about whether AI agents should be required to identify themselves, particularly in contexts involving power imbalances or expectations of interpersonal trust.

Such issues align with Wachter, Mittelstadt, and Floridi's framework for "meaningful transparency" in AI, which goes beyond technical explainability to encompass user comprehension, contestability, and institutional accountability [117]. In educational settings, explainability must support not only technical audits but also pedagogical trust—ensuring that students understand what a robot can and cannot do, how it makes decisions, and whether those decisions are contextually appropriate or biased. As Gordon emphasizes, the construction of "moral robots" involves not just building safe algorithms but embedding moral reasoning in ways that are interpretable to human users [118].

Humanoid robots such as *Pepper* and *NAO*, designed with anthropomorphic features and conversational capabilities, further complicate the issue. Their appearance and behavior often trigger anthropomorphism, which can lead users—particularly younger students or those unfamiliar with AI—to attribute agency or emotional understanding where none exists. Without visual or verbal cues indicating their machine nature, users may misplace trust or assume empathetic capacities [117]. This raises pedagogical risks, as overreliance or confusion could affect learning outcomes, classroom dynamics, or interpersonal boundaries.

Telepresence robots such as *Double 3* and *Ohmni* present similar concerns from a different angle. While their purpose is to enable audiovisual presence, their mobile embodiment often conceals the operator's identity or the session's recording status. Students and faculty may be unaware of who is virtually "in the room," whether the interaction is being logged, or how the data is later used—highlighting the need for transparent usage protocols and live disclosure interfaces.

Design affordances must therefore include real-time transparency tools: indicators of AI control, explicit system introductions, and user-accessible logs detailing how and why decisions are made. For instance, when an AI tutor generates a recommendation or response, it should link back to the rule, dataset, or model rationale underpinning its conclusion. Institutions should also develop clear disclosure policies mandating that robot assistants identify themselves at the beginning of interactions, especially in classes, office hours, or sensitive discussion forums [116].

Crucially, as Chatterjee argues, transparency also shapes user intention and trust in robot adoption [119]. Students and educators are more likely to support or engage meaningfully with AI systems when they perceive institutional safeguards, ethical grounding, and user-facing explanations as part of the deployment process. The lack of such clarity can breed suspicion or resistance, undermining the technology's intended benefits.

Recent regulatory developments affirm these ethical imperatives. The EU AI Act sets out specific obligations for high-risk AI systems—such as those used in education—to ensure both system-level transparency and user comprehension. Article 13 requires providers to furnish detailed instructions for deployers, including: the intended purpose of the system; known limitations and foreseeable misuse scenarios (Art. 13(3)(b)) [79]; performance accuracy; and explanations of how outputs are generated and interpreted (Art. 13(3)(b)(iv, vii)) [79]. These requirements serve to enhance both auditability and usability, ensuring deployers can make informed decisions about system behavior and outcomes.

Furthermore, Article 13(3)(d) [79] mandates technical measures for human oversight, including interpretability tools that allow users to monitor and intervene where needed. Article 13(3)(f) [79] extends this to log collection and review, stipulating that AI systems must include mechanisms for capturing interpretable operational records. In educational settings, these features would apply to robot tutors, proctoring systems, or learning assistants, ensuring stakeholders can assess whether output was contextually appropriate or ethically problematic.

Complementing this, Article 50 focuses on public-facing transparency. It stipulates that AI systems which interact directly with natural persons—such as conversational robots—must clearly disclose that interaction is with a non-human entity (Art. 50(1)) [79]. This includes physical and virtual robots, unless the artificial nature is self-evident. Additionally, Article 50(5) [79] requires that such disclosures be made "in a clear and distinguishable manner" at the point of first contact—reinforcing the need for visible cues or system prompts, particularly in educational environments where power asymmetries and trust relationships matter.

Together, these provisions underscore that transparency is not merely a best practice but a legal obligation for high-risk AI systems in education. When implemented thoughtfully, they promote accountability, facilitate contestability, and help ensure that AI systems operate in ways that uphold student dignity and institutional trust.

Ultimately, transparency and explainability are not merely technical challenges but social contracts. Robot assistants, especially in education, mediate human relationships, knowledge access, and authority structures. Their deployment must reflect a commitment to openness, fairness, and human-centered design—not just to meet regulatory thresholds, but to uphold the integrity of the learning experience.

4.3.3 Accountability and Liability

The integration of autonomous or semi-autonomous robot assistants in HE presents a growing challenge in delineating accountability for actions taken or decisions made by these systems. As robot assistants increasingly participate in educational processes—such as grading, attendance monitoring, or advising—they transition from being mere tools to active participants in academic governance. This shift necessitates a fundamental reassessment of responsibility structures within institutions, particularly when errors or harm occur.

For example, humanoid robots like *Pepper* and *NAO* are often used as interactive tutors or facilitators in classroom environments. Should these robots deliver incorrect instructional content or respond inappropriately to a student's question, the downstream effects may include confusion, misinformation, or loss of trust. Likewise, AI-powered virtual assistants, exemplified by systems like the Jill Watson AI teaching assistant, have been deployed to autonomously answer student queries or manage forum discussions. If such systems provide erroneous feedback or inadvertently reinforce biased interpretations, students may unknowingly act on flawed guidance.

The issue is further complicated when administrative tasks are automated. For instance, a service robot managing access control to examination venues or delivering confidential documents—such as medical accommodations—must execute these functions with a high degree of precision and confidentiality. A misdelivery, unauthorized disclosure, or system misconfiguration can result in significant institutional liability. Telepresence robots, similarly, if left active during sensitive meetings, may lead to breaches of confidentiality that universities are legally and ethically bound to prevent.

As Krupp et al. [77] observe, public perception of responsibility and privacy risks posed

by telepresence robots varies widely, and universities cannot rely solely on technical correctness to maintain trust. Clear human oversight and transparent escalation channels are vital, especially in contexts of data exposure or behavioral interventions.

Legal scholarship has emphasized the importance of defining and enforcing meaningful human control in human-robot teams. Verhagen et al. [120] argue that in safety-critical environments, such as firefighting, varying levels of autonomy must be explicitly accompanied by policies that anchor final decision-making authority in human agents. A similar principle should apply in education, where the stakes may not be physical safety, but involve cognitive, emotional, and developmental outcomes. High-stakes decisions—grades, disciplinary outcomes, or welfare referrals—must never be left solely to machines.

Emerging scholarship on algorithmic and robotic liability underscores how current legal frameworks—rooted in tort, contract, and product liability law—struggle to assign blame when harm arises from systems that learn or adapt over time. Barfield [121] warns that as robots evolve beyond predictable scripts, the law faces novel challenges in tracing causality and responsibility. Algorithms that generate unanticipated behavior may act in ways that no human—designer, deployer, or user—can foresee or directly control. In such cases, assigning fault becomes legally and ethically ambiguous.

Tóth et al. [122] introduce the notion of "accountability clusters," recognizing that responsibility in AI robot contexts is often dispersed across a constellation of actors—designers, institutional users, vendors, and regulators. Their proposed framework highlights how different contexts—such as educational institutions—demand tailored ethical and legal regimes that consider both the locus of moral agency (human vs. system) and the moral intensity of the task at hand. As accountability disperses, universities must formalize multi-level governance mechanisms to document, review, and adjudicate robot-related incidents.

This ethical demand finds regulatory reinforcement under the EU AI Act, which codifies accountability through detailed procedural and oversight requirements. Providers of high-risk AI systems must implement a comprehensive, iterative risk management framework that assesses and updates safety measures throughout the AI lifecycle (EU AI Act, Art. 9(1–2), p. 56) [79]. This includes anticipating misuse, addressing bias, and proactively mitigating new harms as they emerge in real-world contexts—for instance, unexpected student behaviors or socio-cultural dynamics in classrooms.

Providers must also maintain technical documentation (Art. 11, p. 58) [79] and logging capabilities (Art. 12, p. 59) [79], establishing traceability for decisions made by robot

systems. These artifacts enable both internal reviews and external audits, and they are especially critical when the AI system is involved in sensitive functions like grading or behavioral monitoring. When errors occur, logs provide an evidentiary foundation for redress.

Moreover, Article 14 of the Act mandates robust human oversight mechanisms. Robot assistants must include interfaces and procedures that empower designated humans to monitor, interpret, and override system outputs (Art. 14(4), p. 60) [79]. Oversight is not merely technical; it must aim to minimize risks to fundamental rights, including fairness in assessment and equitable treatment (Art. 14(2), p. 60) [79]. In practice, this might involve teacher dashboards, kill switches, or explainability tools that illuminate why a robot flagged a student for concern.

Educational institutions as deployers are not exempt. They must assign qualified individuals to supervise robot systems (Art. 26(2), p. 68) [79], follow the provider's usage instructions, and ensure compliance with operational constraints. This legal linkage ensures that there is always a named, competent individual responsible for the robot's actions within the academic setting.

Accountability is further enforced through conformity assessment mechanisms. Before deployment, high-risk systems must pass formal certification conducted by impartial "notified bodies" (Art. 43, p. 80; Art. 31, p. 71) [79], who verify adherence to requirements related to data, transparency, cybersecurity, and more. These bodies must be independent, technically competent, and free from conflicts of interest (Art. 28(3), p. 70) [79].

Confidentiality provisions (Art. 78, p. 105) [79] assure that providers can disclose system details without risking trade secrets during audits, encouraging transparency without compromising intellectual property. Meanwhile, regulators and notified bodies are empowered to conduct joint investigations, monitor compliance, and intervene when violations occur.

Finally, the Act introduces strong penalties for non-compliance. For instance, breaching prohibitions or omitting risk mitigations can result in fines of up to €35 million or 7% of global turnover (Art. 99(3), p. 115) [79], with public institutions like universities also subject to sanctions (Art. 100) [79]. Transparency violations and failure to implement oversight mechanisms are also punishable under Article 99(4) [79] reinforcing the importance of procedural diligence.

In sum, the EU AI Act transforms scholarly calls for educational accountability into a structured legal obligation. Through risk management, human oversight, technical

documentation, conformity assessments, and enforceable penalties, it ensures that robot assistants in education operate within a system of responsible human governance. Institutions must move beyond ad hoc measures to proactive legal, technical, and procedural planning that anticipates harms, embeds oversight, and protects both institutional integrity and student well-being.

4.3.4 Autonomy and Human Oversight

As robot assistants become increasingly autonomous, the question of appropriate human oversight takes on critical importance. In HE, where pedagogical authority, academic evaluation, and student welfare are at stake, allowing AI-driven systems to operate without meaningful supervision can result in unanticipated and sometimes damaging outcomes. Robots such as *Pepper*, *NAO*, and *ARI* have been integrated into classrooms to deliver instructional content, lead quizzes, or answer student questions autonomously. While these features reduce faculty workload, they also risk delegating key academic decisions—such as evaluating student engagement or suggesting remedial materials—to opaque algorithmic processes.

A central ethical concern is the erosion of accountability when decision-making is partially or wholly transferred to an AI agent. For instance, if an AI teaching assistant provides incorrect feedback on an assignment or fails to recognize a student's need for academic support, it is not always clear who should be held responsible: the instructor, the software vendor, or the system administrator [123]. The diffusion of responsibility can delay remediation and diminish student trust in the educational process. This is particularly salient for systems used in assessment contexts, such as auto-grading platforms or performance-monitoring robots, which may be embedded into virtual assistants or classroom robots without rigorous validation of their accuracy and fairness.

The problem is compounded when physical robots with embedded AI (e.g., *Temi*, *Care-O-bot*, or *Starship*) make real-time decisions about mobility or task prioritization in dynamic campus environments. For instance, a service robot autonomously navigating a hallway may need to decide whether to interrupt a group conversation to complete a delivery. Without carefully designed intervention thresholds, these systems can cause social disruption, discomfort, or even physical risk—especially in crowded settings like university lobbies or labs. In such scenarios, clear policies must define when and how human operators can override or intervene in robotic behavior.

Recent research suggests that autonomy in robots must be matched with safeguards to ensure users retain agency and moral control. Verhagen et al. [120] propose a model

of “meaningful human control” for variable-autonomy systems, emphasizing that human oversight should be scalable to the robot’s context and capabilities. Their findings from high-risk fields like firefighting underscore the need for contextual transparency, where the AI system continuously signals when and how humans can intervene. This principle is directly transferable to the classroom, where students’ learning trajectories or emotional responses must never be determined by unreviewed automated judgment.

This need for structured oversight is partially codified in the EU AI Act. Article 14 specifically mandates that high-risk AI systems—including those used in education under Annex III—be developed with built-in mechanisms for effective human supervision during their use (EU AI Act, Art. 14(1), p. 60) [79]. It requires systems to be designed with interfaces that allow human agents to monitor, understand, and intervene in AI operations when necessary. The Act explicitly states that oversight must prevent or minimize risks to health, safety, and fundamental rights, particularly in scenarios involving foreseeable misuse (Art. 14(2), p. 60) [79].

Moreover, the Act requires that human oversight measures be proportionate to the system’s level of autonomy and the context of deployment. These measures must either be built into the system by the provider (e.g., with interrupt capabilities or alert systems), or defined as responsibilities to be executed by the deployer (e.g., requiring human staff to verify actions before implementation) (Art. 14(3), p. 60) [79]. For example, a university using a robot to flag academic misconduct or attendance discrepancies must ensure that such outputs are reviewed and confirmed by a trained educator before taking any formal action—thus aligning with Article 14(4)(d) [79], which gives humans the right to override or reverse AI decisions.

The Act also seeks to mitigate automation bias, reminding deployers of their obligation to ensure that staff using AI outputs remain critical and cautious rather than deferring to automated recommendations uncritically (Art. 14(4)(b), p. 60). This directly addresses classroom environments where instructors may be tempted to treat AI-generated performance summaries or behavioral assessments as infallible, when in fact such systems may reflect hidden biases or limited contextual understanding.

In line with these mandates, best practices in educational robotics should include tools that allow real-time visualization of AI decision paths, alert systems for anomalies, and manual override capabilities such as kill switches or administrative backends. Feedback and logging systems must allow educators to document why AI decisions were accepted or rejected, which supports post-hoc auditing and system improvement. Saunderson and Nejat [124] warn that the illusion of AI authority—especially in social robots—can prompt

users to comply without question. In the classroom, such persuasion must be tempered by mechanisms that remind users that robots are tools, not authorities.

Janina Loh [123] emphasizes that ethical robot deployment requires not just technical oversight but philosophical clarity around moral agency and responsibility attribution. Educational institutions must therefore cultivate a governance culture that treats robot autonomy as a managed variable—not a fixed trait. Robotic agency should always be subject to pedagogical and ethical review, with ultimate decision authority retained by qualified humans.

Ultimately, the EU AI Act reinforces the view that autonomy must not come at the expense of accountability or safety. While Article 14 does not delve into the full ethical complexity of education-specific AI deployments, it provides a legal scaffold that supports human-centered oversight and empowers deployers to tailor control structures to academic contexts. With enforcement slated for August 2025, universities have a limited window to bring their robotics strategies into compliance—redefining not just how AI assists education, but how it is governed in the classroom.

4.3.5 Bias and Fairness

Ensuring fairness and mitigating algorithmic bias are among the most urgent challenges in deploying AI-powered robot assistants in HE. As robots are increasingly entrusted with instructional, evaluative, and administrative roles, concerns arise about whether these systems treat all users equitably across dimensions such as race, gender, language, and ability. Emerging research in robotics ethics and learning algorithms underscores that fairness is not simply a question of avoiding overt discrimination, but also of actively auditing and reshaping the underlying data, interaction patterns, and learning objectives embedded within these systems [125].

A primary source of unfairness in robot-assisted education lies in biased training data. Londono et al. [125] note that robot learning systems—especially those relying on reinforcement learning or imitation learning—are highly sensitive to data quality and scope. If the system is trained on data collected from a narrow demographic, it may generalize poorly to diverse populations. This limitation is particularly salient in educational contexts, where student diversity includes not only visible traits like ethnicity and gender, but also less overt dimensions such as accent, learning style, or neurodiversity.

In recognition of this, the EU Artificial Intelligence Act (AI Act) imposes strict obligations for bias detection, mitigation, and prevention in high-risk systems, including those used

in education. Article 10 mandates that training, validation, and testing datasets must be "relevant, sufficiently representative, and to the best extent possible, free of errors and complete" (EU AI Act, Art. 10(3), p. 57) [79]. This means developers must ensure their datasets reflect the diversity of the target population—such as students from varying cultural backgrounds, linguistic profiles, and neurotypes. Inadequate representation can lead to discriminatory outcomes, and failure to address these gaps constitutes non-compliance with the regulation (Art. 10(2)(f–h), p. 57).

Moreover, to safeguard against discrimination, Article 10(5) [79] permits the exceptional processing of special categories of personal data—such as racial or ethnic origin—strictly for the purpose of bias detection and correction, provided that appropriate safeguards like pseudonymization and access controls are implemented. This allowance underscores the legal and ethical seriousness of ensuring AI fairness, even when privacy-sensitive data must be used for that purpose (Art. 10(5)(a–f), p. 57) [79].

Bias also emerges through physical and interactional design. In their study of gender bias in educational robots, Cesaro et al. [126] found that robot embodiment and behavior often reflect and reinforce gender stereotypes. Robots are frequently anthropomorphized as male when associated with authority (e.g., grading, instruction) and as female when designed for supportive or emotional tasks. This can subconsciously affect how students interpret their authority or emotional intelligence, and may influence gendered expectations about roles in educational and professional domains. Educational robots, especially those embedded with social and affective interfaces, must be carefully evaluated for the symbolic roles they play in classroom hierarchies and student identity formation.

These risks fall under the scope of the risk management system required by Article 9, which must be established and maintained across the AI system's entire lifecycle (EU AI Act, Art. 9(1–2), p. 56) [79]. This includes identifying and mitigating risks related to health, safety, and fundamental rights—with a specific directive to evaluate impacts on children and other vulnerable groups (Art. 9(9), p. 56) [79]. The risk management measures must address foreseeable misuse (e.g., stereotyping or excluding students) and ensure that the residual risk is acceptable (Art. 9(5), p. 56) [79]. Importantly, these measures must not be static: AI systems must undergo continuous testing, including real-world deployment trials, to confirm that they behave as intended and do not introduce unintended disparities (Art. 9(6–8), p. 56) [79].

To address systemic bias in robotic behavior, recent work by Zhu et al. [127] introduces fairness-sensitive policy gradient methods in reinforcement learning. Their approach modifies the reward structure to penalize interactional imbalances between different user

groups, thus enabling robots to adapt their assistance strategies in real time to achieve more equitable engagement. Such methods hold significant promise for classroom robotics, where unequal attention or response quality could reinforce existing disparities in learning outcomes. For example, a robot that preferentially responds to students with louder voices or standard dialects may unintentionally marginalize quieter or non-native speakers unless its policy is fairness-adjusted.

Preventing bias and ensuring fairness is not only a technical best practice, but now also a legal compliance issue under the AI Act. AI systems, if unchecked, could reinforce prejudices or treat individuals unequally—a serious concern in education where AI might influence grading, tutoring, or student engagement. Providers must take active and proactive steps to detect, prevent, and mitigate discriminatory behavior or outcomes. This includes incorporating fairness metrics into evaluation, diversifying training datasets, and adapting model behavior through dynamic risk management practices (EU AI Act, Art. 9(2)(d), p. 56; Art. 10(2)(g), p. 57) [79].

Additionally, the AI Act outright bans certain forms of algorithmic unfairness, such as "social scoring" (e.g., ranking students based on perceived personality or social behavior rather than academic performance) and any disproportionately detrimental profiling (EU AI Act, Art. 5(1)(c–d), p. 38) [79]. These prohibitions draw a clear ethical and legal boundary, ensuring that robot assistants cannot assign persistent "low potential" labels or use prior behavioral data to penalize students in unrelated contexts.

The issue of fairness is not limited to individual robot-student interactions. It also extends to broader institutional structures. Londono et al. [125] argue for multi-level auditing processes that include not only technical validation, but also social and legal oversight to ensure that robot deployments align with human rights and educational equity principles. They propose interdisciplinary collaboration between roboticists, ethicists, and educators to anticipate bias amplification before systems are deployed in real-world environments.

The post-market monitoring obligations under the AI Act (Art. 72, p. 102) [79] reinforce this view, mandating that developers continuously assess and address emerging risks related to fairness, especially when systems are introduced to new institutional or cultural contexts.

Importantly, bias mitigation must extend beyond one-time fairness metrics. Educational robots require continuous monitoring and recalibration as they interact with new student populations and contexts. Dynamic classroom environments present shifting variables in terms of language, culture, and technology access. Systems that may appear unbiased in

lab conditions can exhibit problematic behavior when deployed at scale, particularly when deployed across institutions with different infrastructure or pedagogical models. Feedback loops and redress mechanisms—such as logging incidents, accepting user-flagged concerns, and offering alternate interaction modalities—should be integrated from the start.

In sum, fairness in educational robotics is not a static feature that can be tested and signed off. It is an ongoing process of reflection, intervention, and co-design. Universities must treat bias not merely as a technical flaw to be patched, but as an institutional risk that demands transparency, stakeholder participation, and an explicit commitment to equitable access to the benefits of educational automation. The EU AI Act strengthens this mandate by embedding fairness into both technical design and regulatory accountability. Developers and deployers of robot assistants who fail to uphold these standards risk not only ethical censure but also legal liability for non-compliance.

4.3.6 Equity and Accessibility

The deployment of robot assistants in HE has the potential to foster more inclusive, accessible learning environments. However, this potential can only be realized through intentional efforts to center equity from the earliest stages of system design and implementation. Without explicit consideration of accessibility, representation, and social context, robot deployments risk reinforcing the very disparities they claim to address.

Inclusive robotics in education must begin with the acknowledgment that not all users interact with technology from a level playing field. Saille et al. [128] argue that traditional top-down approaches to robotic development often sideline the needs of underrepresented groups. Their work on upstream co-creation emphasizes that equity must be built into the design process itself—not retrofitted after deployment. Through participatory workshops and stakeholder engagements involving disabled students, caregivers, and educators from marginalized communities, Saille et al. show how co-creative design methods can surface accessibility concerns and equity goals that are otherwise ignored by technocentric design teams.

Ostrowski et al. [129] further reinforce the need for equity-aware frameworks in human-robot interaction (HRI). In their review of ethics, equity, and justice in robotics research, they propose a Design Justice lens that interrogates not just how robots function, but who they are designed for, who participates in their creation, and who controls their deployment. Their framework invites educators and technologists to reflect on how power, privilege, and marginalization shape robot-student interactions—particularly in relation to race, gender, ability, and socioeconomic status. For instance, if robot tutors or assistants respond better to

standard dialects or Eurocentric gestures, students from linguistically or culturally diverse backgrounds may experience diminished engagement or recognition. Design Justice in this context means involving those most affected by educational inequality in setting the terms of robot use and evaluation.

Equity also intersects with identity-sensitive robotics. Poulsen et al. [130] highlight how robots that interact with vulnerable populations, such as LGBTIQ+ elders, must be designed with awareness of identity and community values. Though focused on eldercare, their insights apply directly to educational contexts, where students bring complex, intersectional identities into the classroom. A robot that fails to account for gender diversity, neurodivergent communication patterns, or non-traditional learning pathways may unintentionally exclude or stigmatize the very students it aims to support. As Poulsen et al. argue, ethical robotics must include robust engagement with community norms and cultural variation, particularly when these systems are designed to build trust and provide sensitive support.

From a regulatory standpoint, the EU AI Act codifies many of these principles as legal obligations and policy imperatives. Notably, as a signatory to the United Nations Convention on the Rights of Persons with Disabilities, the European Union requires that AI systems, including robot assistants in education, be designed for accessibility "on an equal basis with others" and in full recognition of the diversity and dignity of users with disabilities (EU AI Act, Recital 80, p. 23) [79]. This extends beyond interface tweaks to mandate universal design principles embedded in the technical and interactional architecture of AI systems.

In addition to these baseline requirements, Article 95 encourages the voluntary application of fairness, accessibility, and inclusion safeguards even to AI systems not classified as high-risk. Providers are specifically urged to assess and prevent negative impacts on vulnerable groups, including people with disabilities and marginalized genders (EU AI Act, Art. 95(1)(e), p. 113) [79]. In educational contexts, this provision directly supports efforts to ensure that robot tutors or navigational assistants are usable by all students, regardless of physical or cognitive ability.

The Act also highlights the importance of inclusive and interdisciplinary design practices. Recital 142 encourages Member States to prioritize funding and development support for AI systems that address social inequalities, such as improving accessibility for students with disabilities or mitigating the digital divide in HE (EU AI Act, Recital 142, p. 36) [79]. These projects are expected to bring together roboticists, educators, disability experts, and social scientists—aligning closely with the participatory, co-creative methodologies

advocated in recent HRI scholarship.

Even outside the domain of high-risk systems, the promotion of voluntary codes of conduct (Recital 165) calls on all AI developers—including those in education—to integrate "inclusive and diverse design and development" as a standard practice (EU AI Act, Recital 165, p. 41) [79]. This includes balancing gender in development teams, engaging civil society and academic partners, and measuring equity goals with clear performance indicators. Such initiatives do not merely reflect ethical ideals but create concrete pathways for accountability and transparency in educational robotics.

Equity, therefore, is not merely a logistical challenge of distributing robots across campuses or accommodating disability access protocols. It is an epistemic and ethical commitment that reshapes how robotic systems are conceived, tested, and integrated. HE institutions should adopt participatory design practices, conduct equity impact assessments, and establish inclusive governance mechanisms for robot deployment. These steps not only fulfill principles of justice and non-discrimination but also align with the EU's legal and policy frameworks for trustworthy, rights-respecting AI.

In sum, the pursuit of accessibility and equity in educational robotics must move beyond compliance and toward genuine co-creation. This means asking not just "can all students use the robot?" but also "whose values and voices shaped its development?" and "whose needs remain unmet?" Only through such reflection and redesign can robot assistants become tools of educational empowerment rather than agents of exclusion—fulfilling both ethical duties and legal obligations under the EU AI Act.

5. Guideline-Driven Framework

This chapter introduces a comprehensive framework of guidelines aimed at addressing the ethical and cybersecurity challenges posed by robot assistants in HE. Rather than serving solely as an assessment tool, the framework is intended to inform the design, deployment, and governance of robotic systems by offering clear, actionable recommendations.

Structured across a set of defined control domains, each area outlines both the "risks to be mitigated" and the "practices to be adopted". The framework supports developers, instructors, IT staff, and policymakers by breaking down each domain into specific sub-controls—labeled systematically (e.g., T.E.-1 for a Transparency and Explainability control)—that include concrete goals, expected implementation practices, and examples of compliance.

Institutions can use this framework as a blueprint to:

- Develop internal policies and technical requirements for robot-assisted systems,
- Evaluate current deployments for ethical and security alignment,
- Guide future development with built-in oversight, inclusivity, and resilience in mind.

While the framework can be used for scoring and analysis it is first and foremost a collection of "best-practice guidelines" designed to help educational institutions deploy robotic technologies responsibly and sustainably.

5.1 Positioning Relative to Existing Security and AI Governance Frameworks

The guideline-driven framework proposed in this thesis does not exist in isolation. It draws on and complements prior work in robotics security as well as established information security and AI governance standards. This section briefly situates the framework in relation to these bodies of work and clarifies its distinctive contribution.

Research on robot cybersecurity has highlighted significant vulnerabilities in platforms such as ROS and ROS 2, including unauthenticated topics and services, exposed debugging interfaces, and weak default configurations. Secure variants such as SROS and SROS 2, together with penetration studies on commercial robots, provide concrete recommendations

for hardening middleware, enabling encryption and authentication, and isolating critical components. These efforts are primarily *stack-focused*: they concentrate on specific robotic platforms or middleware layers and aim to improve their technical resilience.

In contrast, the framework in this thesis is *deployment-focused*. Rather than targeting a single middleware or robot family, it is designed to be agnostic to underlying platforms and to cover the full socio-technical environment in which robot assistants operate in HE. The control domains explicitly include institutional governance, human oversight, equity and accessibility, and transparency, in addition to technical topics such as authentication, vulnerability management, and network security. As a result, the framework can be applied both where ROS/SROS-based stacks are used and where robots rely on proprietary cloud services or custom AI pipelines.

The framework also relates to general IT and OT security baselines such as NIST Special Publication 800–53, ISO/IEC 27001, and IEC 62443. Several cybersecurity domains parallel familiar control families: for example, *Data Privacy & Confidentiality* and *System Access & Control* echo NIST 800–53 Access Control (AC) and System and Communications Protection (SC) families, as well as ISO/IEC 27001 Annex A controls on information security policies and asset management. Likewise, the *Vulnerability & Patch Management* and *System & Data Integrity* domains reflect principles embedded in IEC 62443 and ISO/SAE 21434 regarding structured patch processes, secure update mechanisms, and integrity verification. The intention is not to replace these standards but to provide a higher-education robotics layer that translates their abstract requirements into robot- and classroom-specific practices.

At the same time, the framework extends beyond traditional security baselines by integrating ethical and regulatory dimensions that are central to educational robotics. Ethical domains such as *Bias and Fairness*, *Autonomy and Human Oversight*, and *Equity and Accessibility* are explicitly aligned with requirements in the EU Artificial Intelligence Act for high-risk AI systems in education, including obligations on data governance, transparency, and human-in-the-loop decision-making.

5.2 How to Use This Framework

This framework is a practical tool for evaluating and improving the ethical and cybersecurity readiness of robot assistant deployments in HE. It is designed to be flexible, allowing institutions to tailor control selection and implementation depth based on their specific context, budget, and risk profile.

The following scenario as an example: a university intends to introduce the *Temi* robot as an autonomous teaching assistant in an undergraduate Psychology classroom. Temi will be responsible for interacting with students, asking formative questions, collecting assignments, and delivering basic tutoring aligned with pre-programmed content. Prior to deployment, the institution applies this framework to perform an initial assessment, beginning with the **Control Evaluation Matrix** (Table 3) and the accompanying **Control Assessment Scale** (Table 2). Each control is rated from 0 to 3, depending on how comprehensively it is addressed in the proposed implementation.

After assigning scores, the team consults the **Control Scoring per Domain Table** (Table 4) to examine strengths and gaps across key areas such as Privacy, Human Oversight, and Transparency. For example, while Temi might demonstrate high coverage in Network Security, it may score lower in Explainability or Informed Consent. These diagnostic results help the institution identify priorities for remediation. Through regular reevaluation of these controls—especially following technical upgrades or policy revisions—the institution can progressively strengthen its deployment, lowering risk exposure while improving ethical and legal compliance. Although each control is independently actionable, their collective application fosters a culture of secure, equitable, and responsible robotics use in academia.

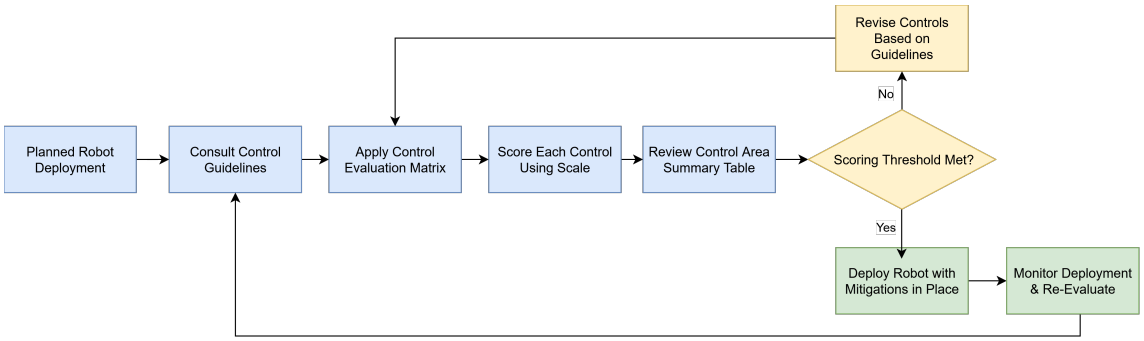

Figure 9. Workflow for Applying and Iterating the Control Framework

5.3 Control Assessment Overview

Each control in the framework is evaluated using a standardized scoring system (Table 2) that reflects the degree to which an institution or deployment fulfills a given requirement. The score ranges from 0 (Not Addressed) to 3 (Fully Implemented), allowing for quick benchmarking of current readiness and identification of gaps.

Evaluators begin by referencing the Control Evaluation Matrix, where each control is described alongside an implementation observation. Based on available evidence—such as policy documentation, technical configuration, or observed behavior—a score is assigned for each control. This score is not only a reflection of technical compliance but also

considers accessibility, usability, and transparency from a stakeholder perspective.

Once individual scores are assigned, the Control Area Summary Table (Table 4) is used to group the results by domain (e.g., Privacy, Transparency, Accountability). This enables institutions to assess both the average maturity of their deployment and the distribution of strong or weak areas. The "% ≥ 2" column shows how many controls in each domain are at least partially fulfilled. This helps highlight which areas need prioritization or further improvement.

Table 2. Control Assessment Scale

Score	Label	Definition / Criteria
0	Not Addressed	No indication that the control is considered or implemented. No policy, design, or documentation in place.
1	Minimally Addressed	The control is informally mentioned or implemented in an ad hoc way. It may be incomplete, inconsistently applied, or lack user awareness.
2	Partially Fulfilled	There is a clear implementation, but it may be narrow in scope or lack consistency, regular auditing, or full stakeholder buy-in.
3	Fully Implemented	The control is formally defined, systematically applied, supported by technical tools or policies, and regularly reviewed for effectiveness.

Table 3. Example Evaluation Matrix

Control ID	Control Description	Assessment	Observations
P.C.-1	Users must explicitly opt in to data collection.	1	Consent is assumed by use; no opt-in option provided.
P.C.-2	Users are told what data is collected and why.	2	General purpose stated, but details are missing.
P.C.-3	Sensors must show when active; users can opt out.	1	No clear indicators or user controls available.
P.C.-4	Users can access or delete their personal data.	0	No access or deletion features available.

Table 4. Example Control Scoring per Domain

Control Domain	1	2	3	4	Avg. Score	% ≥ 2
Privacy, Consent (P.C.)	1	1	2	1	1.25	25%
Transparency & Explainability (T.E.)	3	2	1	1	1.75	50%
Accountability & Liability (A.L.)	1	2	1	2	1.50	50%
Autonomy & Human Oversight (H.O.)	1	1	1	–	1.00	0%
Overall					**1.16**	**26.0%**

5.4 Cybersecurity Guidelines

Data Privacy (D.P.)

Protecting sensitive academic data handled by robots from unauthorized access or leaks.

Table 5. Data Privacy (D.P.) Control Guidelines

Control ID	Control Name	Guideline Description
D.P.-1	End-to-End Encryption	All data collected or transmitted by robot assistants must be encrypted using strong, modern protocols (e.g., TLS 1.3, AES-256). This includes: ■ Facial recognition feature vectors and audio logs. ■ Attendance or performance data sent to servers or emails. ■ Local database storage on shared or mobile robots. ■ Live streams or real-time interactions over telepresence links.
D.P.-2	Minimized Biometric Retention	Robots should avoid storing raw biometric data (e.g., facial images, voice samples) unless strictly necessary. Where necessary: ■ Convert to anonymized feature templates immediately. ■ Purge raw inputs within seconds of processing. ■ Provide user-facing options to delete stored biometrics.
D.P.-3	Secure Transmission Enforcement	Robots must actively prevent insecure communication practices. Systems must: ■ Enforce strict SSL/TLS certificate validation on all outbound and inbound connections. ■ Disable fallback to deprecated protocols (e.g., SSLv3, TLS 1.0/1.1). ■ Log and alert administrators when transmissions occur over insecure Wi-Fi, Bluetooth, or peer-to-peer modes. ■ Apply the same enforcement to real-time services like telepresence streams or cloud data syncs.
D.P.-4	Purpose-Limited Data Use and Retention	Robots must only collect and use data for clearly defined, educationally justified purposes. Institutions should: ■ Clearly define the data collection purpose and ensure alignment with the original context (e.g., attendance tracking only). ■ Limit data retention to the minimum necessary duration; apply automated deletion policies post-course or term. ■ Avoid secondary use of student data (e.g., emotion data for behavioral profiling) unless re-consent is explicitly obtained. ■ Conduct regular reviews to ensure compliance with data minimization and purpose-limitation principles.

System & Data Integrity (S.I.)

Ensuring software, firmware, and AI models used by robot assistants cannot be tampered with to introduce errors or malicious actions.

Table 6. System & Data Integrity (S.I.) Control Guidelines

Control ID	Control Name	Guideline Description
S.I.-1	Immutable Audit Logs	Critical actions—such as attendance edits or data deletions—must be recorded in tamper-evident logs, including: ■ Timestamps, actor IDs, and data operation types. ■ Digitally signed or hash-chained log formats. ■ Exportable formats for forensic or audit review.
S.I.-2	Secure Update Pipeline	Software, firmware, and AI models must only be updated through authenticated channels. Institutions should: ■ Verify signatures on updates before application. ■ Log all update events, including version and installer identity. ■ Reject updates over unsecured networks.
S.I.-3	Runtime Integrity Checks	Robots must include self-checks and validation routines to detect unauthorized modification at runtime. This may include: ■ File integrity monitoring for critical binaries and application scripts. ■ Runtime cryptographic verification of AI models (e.g., hash matching against signed baseline models). ■ Alert mechanisms for checksum mismatches, tampering attempts, or unauthorized behavior changes.
S.I.-4	Isolation of Trusted Components	Security-critical functions (e.g., identity matching, biometric verification, system logging) must operate within isolated and hardened execution environments to reduce the risk of compromise. This includes: ■ Segregating trusted components (e.g., authentication modules, attendance logging, encryption services) from general-purpose layers such as user interaction, navigation, or chatbot services. ■ Using OS-level sandboxing, containerization, or hardware-backed enclaves (e.g., ARM TrustZone or TPMs) to isolate critical processes. ■ Ensuring inter-process communication (IPC) between trusted and non-trusted components is minimal, authenticated, and logged. ■ Designing the robot's architecture such that a vulnerability in peripheral components (e.g., ROS nodes, touchscreen interfaces, cloud connectors) cannot escalate to control or corrupt sensitive logic.

Information Leakage & Misuse (I.L.)

Even unintentionally, robots may leak sensitive data through interfaces, responses, or audio/video feeds. This control area seeks to reduce inadvertent data exposure.

Table 7. Interaction & Logging (I.L.) Control Guidelines

Control ID	Control Name	Guideline Description
I.L.-1	Safe Defaults for Audio/Video	Robots must not record or transmit audio/video unless explicitly required. Defaults should include: ■ Microphone and camera disabled by default. ■ Visible indicators (e.g., LED) when recording is active. ■ Explicit session consent for live interactions. ■ Local processing preferred over cloud-based storage or streaming where feasible.
I.L.-2	Session Expiry & Timeout Controls	Robot sessions should automatically close when inactive to avoid lingering access. Recommendations include: ■ Timed auto-logout after inactivity. ■ Auto-disconnect from telepresence after use. ■ Restriction of background processes after end of session.
I.L.-3	Output Filtering & Response Censorship	Robot-generated responses must be monitored to avoid accidental data disclosure. This includes: ■ Preventing unintended sharing of other students' data. ■ Filtering responses based on user role (e.g., student vs. admin). ■ Blocking external prompts from revealing private data. ■ Ensuring responses are contextually appropriate to prevent misinterpretation. ■ Using intent detection to restrict sensitive outputs to authorized roles only.
I.L.-4	Retention & Disclosure Policies	Clear institutional policies must define: ■ What data is stored, for how long, and by whom. ■ What logs or interactions may be disclosed and under what authority. ■ How and when users can request deletion or review of their data.

System Access & Control (S.A.)

Restricts administrative privileges, enforces access control, and protects interfaces from abuse or compromise.

Table 8. System Access & Control (S.A.) Control Guidelines

Control ID	Control Name	Guideline Description
S.A.-1	Role-Based Access Control (RBAC)	System access must be segmented by roles such as instructor, student, and admin. Each role must have: ■ Minimum necessary permissions. ■ Regular access reviews (e.g., every semester). ■ Logged role escalation events.
S.A.-2	Credential and Interface Security Hardening	Robots must disable default credentials and restrict exposed access interfaces prior to deployment. Additional requirements include: ■ Passwords must meet institutional complexity requirements and be rotated periodically. ■ No shared or hardcoded credentials allowed. ■ Admin interfaces should support 2FA and restrict access to known IP ranges. ■ Debugging ports (e.g., USB, SSH, developer APIs) must be disabled unless explicitly required and secured.
S.A.-3	Access Attempt Logging & Anomaly Monitoring	All login attempts, successful and failed, must be logged with sufficient metadata. Logs should include: ■ User ID (if available), timestamp, IP or device ID. ■ Geolocation or network zone indicators to detect anomalous access. ■ Threshold-based alerting for brute-force attempts or behavioral anomalies (e.g., rapid login cycling). ■ Integration with institutional SIEM systems for centralized security monitoring.
S.A.-4	Privilege Escalation Controls and Re-Authentication	Sensitive operations (e.g., firmware updates, system resets) must be gated with strong verification. Requirements include: ■ Re-authentication for all high-impact administrative tasks. ■ Session timeout enforcement and automatic logout from idle admin consoles. ■ Detailed logs recording action, actor identity, justification, and timestamp. ■ Optional alerting to secondary administrators for peer verification.

Vulnerability & Patch Management (V.P.)

Prevents exploitation of unpatched or poorly maintained robot systems.

Table 9. Vulnerability & Patch Management (V.P.) Control Guidelines

Control ID	Control Name	Guideline Description
V.P.-1	Scheduled Patch Application	Robots must support and follow a patch schedule (e.g., monthly or security-driven). Controls should include: ■ Automatic update checks. ■ Admin notifications for required upgrades. ■ Test environments for patch staging before deployment.
V.P.-2	Periodic Vulnerability Scanning	Institutions must scan robot firmware, operating systems, and applications on a recurring basis to detect known weaknesses. This should include: ■ Regular checks against CVE databases and robotics-specific vulnerability listings (e.g., RVD). ■ Use of scanning tools appropriate for robot platforms. ■ Manual validation of vendor security bulletins and changelogs when automated scanning is not supported.
V.P.-3	Secure Baseline Tracking	Maintain version and configuration baselines so changes can be identified. This includes: ■ Hardware firmware versions. ■ Installed app lists and checksums. ■ Change logs maintained via version control or CMDB.
V.P.-4	EOL and Unsupported Component Avoidance	Robots must not operate using operating systems, firmware, or software components that are: ■ Past end-of-life (EOL) and no longer receiving security updates. ■ Undocumented or closed-source with no vendor support guarantees. ■ Lacking a defined upgrade path or subject to vendor lock-in.

Cyber-Physical Safety (C.P.)

Protects against risks of physical harm or unintended actions from robotic systems with actuators or sensors.

Table 10. Cyber & Physical Safety (C.P.) Control Guidelines

Control ID	Control Name	Guideline Description
C.P.-1	Emergency Stop Access	Robots with mobility or motion must have: ■ A visible, accessible emergency stop button. ■ Remote "soft-stop" functionality for supervisors. ■ Safety override documentation and signage.
C.P.-2	Motion Sandbox Testing	Movement scripts (e.g., navigation, gestures) must: ■ Be tested in simulation environments before real-world use. ■ Be validated in controlled testbeds for speed, torque, joint limits, and proximity sensitivity. ■ Include sensor calibration checks (e.g., LIDAR, ultrasonic, encoders) prior to deployment. ■ Fail safely when interrupted or misfired.
C.P.-3	Restricted Control Interfaces	Only verified admin roles may control actuators. This requires: ■ Credential checks before issuing movement commands. ■ UI isolation between untrusted user applications and motion control functions. ■ Logs of all actuator commands, including timestamp and user ID. ■ Sandboxing of actuator interfaces from higher-level scripts or untrusted API calls.
C.P.-4	Physical Activity Logging	Unexpected or hazardous actions (e.g., collisions, abrupt turns) must be: ■ Logged with timestamp, cause, and affected area. ■ Flagged to IT or safety teams. ■ Correlated with control inputs and sensor readings.

Network & Communication Security (N.C.)

Secures communications between robots, administrators, and any external systems or services.

Table 11. Network & Communication Security (N.C.) Control Guidelines

Control ID	Control Name	Guideline Description
N.C.-1	Encrypted Network Protocols	Robots must use secure and up-to-date communication protocols (e.g., HTTPS, SSH, MQTT over TLS 1.3) and avoid deprecated standards such as SSLv3 or plain HTTP. All sensitive traffic—including: ■ Control messages from administrators, ■ Cloud-based logging or updates, ■ User data transmission to institutional systems— must be encrypted using modern cryptographic suites. Mutual TLS or equivalent peer-authentication should be used for sensitive endpoints.
N.C.-2	Network Segmentation and Firewalls	Robots must operate on segmented networks (e.g., VLANs or logically isolated SSIDs), fully separated from: ■ Student Wi-Fi or public wireless zones, ■ Untrusted IoT environments, ■ Direct internet exposure (unless proxied through hardened gateways). Firewall rules must enforce least-privilege access—allowing only approved protocols, IP addresses, and ports. Robots must also have service discovery protocols (e.g., mDNS, SSDP) disabled unless explicitly required within secured network contexts.
N.C.-3	Endpoint Authentication and Validation	All communications must validate source identity via: ■ Certificate pinning, ■ Token-based mutual authentication, ■ MAC/IP filtering and per-device trust lists.
N.C.-4	Replay and Spoofing Protection	Systems must prevent message reuse or impersonation through: ■ Nonce or timestamp mechanisms, ■ Input rate-limiting and anomaly detection, ■ Session key rotation and timeout policies.

Institutional Readiness & Governance (I.R.)

Ensures educational institutions have the organizational structure and preparedness to manage robot security long-term.

Table 12. Institutional Readiness & Governance (I.R.) Control Guidelines

Control ID	Control Name	Guideline Description
I.R.-1	Mandatory IT Security Involvement	All robot deployments must be reviewed and supported by institutional IT and cybersecurity teams. Reviews must cover: ■ Network integration plans. ■ Identity and access protocols. ■ Ongoing patch and audit responsibilities.
I.R.-2	Robot-Aware Policy Integration	Institutions must update existing information security, privacy, and classroom technology policies to: ■ Cover AI and robotics use cases. ■ Define boundaries for deployment and recording. ■ Set responsibilities for users, admins, and developers. ■ Align with regulatory frameworks (e.g., EU AI Act Art. 70–73) for compliance, monitoring, and reporting. ■ Designate institutional roles for oversight (e.g., ethics committee, security officer, AI governance team).
I.R.-3	Incident Response Inclusion	Incident response procedures must include robotic edge cases such as: ■ Unauthorized access to control panels. ■ Physical tampering or hijacking. ■ User-reported harm or data leakage. Annual simulations or tabletop exercises must include robot-related scenarios involving technical, legal, and ethical roles across departments. Exercises should reflect post-market monitoring and incident reporting duties required by the EU AI Act.

5.5 Ethical Guidelines

Privacy, Consent & Surveillance (P.C.)

Protects user autonomy and minimizes unnecessary data collection or surveillance by robots, particularly in environments like classrooms.

Table 13. Privacy, Consent & Surveillance (P.C.) Control Guidelines

Control ID	Control Name	Guideline Description
P.C.-1	Opt-in Consent for Data Collection	Users must explicitly opt in to data collection before interaction. Consent must be: ■ Granular (e.g., separate toggles for audio, video, and logs). ■ Time-bound (e.g., "valid for one semester"). ■ Easily revocable without penalty.
P.C.-2	Transparent Data Usage Notices	Prior to consent, users must receive clear explanations about: ■ What data is collected (e.g., facial data, voice recordings). ■ Why the data is needed and how it will be used. ■ How long the data will be stored and when it will be deleted. Notices must be written in accessible, student-friendly language.
P.C.-3	Surveillance Indicators and Controls	All active sensors (e.g., cameras, microphones, GPS) must be: ■ Clearly labeled and visible on the device. ■ Accompanied by on-screen or audio alerts when in use. ■ Capable of being toggled off by the user unless safety overrides are necessary. ■ Designed to minimize perceived surveillance by clearly signaling when data collection is inactive.
P.C.-4	Data Portability and Erasure Options	Users must be able to: ■ View what data the robot has collected about them. ■ Export a machine-readable copy of their data (e.g., JSON, CSV). ■ Delete all personal data from the system via a self-service portal or request. Institutions must honor these rights within reasonable response times (e.g., 14 days).

Transparency & Explainability (T.E.)

Ensures that students and educators understand when they are interacting with a robot assistant and how its decisions are made. This fosters user trust, minimizes confusion, and supports human oversight.

Table 14. Transparency & Explainability (T.E.) Control Guidelines

Control ID	Control Name	Guideline Description
T.E.-1	Disclosure of Non-Human Identity	Robot assistants must clearly indicate their non-human nature during any interaction involving academic or administrative tasks. This can be implemented through: ■ Visual cues (e.g., robot profile icons, display badges) ■ Spoken introductions (e.g., "Hello, I'm an AI teaching assistant.") ■ Course-level disclosures via syllabi or LMS announcements Especially in cases where human–robot distinctions may not be visually apparent (e.g., text-based agents), explicit labeling is critical to avoid user confusion.
T.E.-2	Explanation Request Interface	Users must be able to request a clear rationale for AI-generated decisions or actions. At minimum: ■ An "Explain" button or feedback prompt should be available in response interfaces. ■ The system should provide a natural language explanation describing key factors in user-understandable terms (e.g., "You were marked absent because your face was not detected during check-in."). ■ Where full explainability is not feasible, the system should state limitations (e.g., "This recommendation is based on statistical trends; further details are unavailable.") and highlight that the output should not be treated as final without human review.
T.E.-3	Documentation of Decision Logic	Developers must create and maintain internal documentation that: ■ Describes the decision logic, model types, and training data assumptions. ■ Identifies known interpretability limitations (e.g., "This system does not explain recommendations based on neural network weights."). ■ Specifies under what conditions the model outputs may be unreliable or unreviewed. This documentation must be accessible to institutional reviewers and ethical oversight bodies.
T.E.-4	Role Transparency in Educational Tasks	Students must be informed when robot assistants are used for core tasks such as grading, advising, or attendance. Disclosure may include: ■ Course introduction presentations or welcome emails stating the system's role. ■ LMS announcements clarifying which responses or feedback are AI-generated. ■ Consent interfaces when students first interact with the system. Lack of transparency in such roles undermines student trust and impedes informed participation in the learning process.

Accountability & Liability (A.L.)

Prevents the deflection of blame from humans to machines by clearly assigning roles and responsibilities for robotic decisions or system failures.

Table 15. Accountability & Liability (A.L.) Control Guidelines

Control ID	Control Name	Guideline Description
A.L.-1	Assigned Human Supervisor	Every deployed robot assistant must have a designated human supervisor responsible for oversight. Responsibilities include: ■ Monitoring system outputs for anomalies. ■ Responding to student appeals and questions. ■ Serving as a liaison with IT or academic administration during incidents. Supervisors should be trained in the system's capabilities and limitations.
A.L.-2	Defined Student Appeal Pathways	Institutions must ensure that students can appeal or contest robot-generated decisions. The process should include: ■ A clear appeal submission method (e.g., online form, academic portal). ■ Defined response timelines and points of contact. ■ Transparent review criteria and documentation of outcomes. Appeals must be handled by qualified staff who can override or amend the robot's decisions.
A.L.-3	Source Attribution Logging	All major robot actions must be traceable to their initiation source—whether human-in-the-loop, autonomous, or hybrid. Logs should record: ■ Input type (e.g., manual trigger, automated scheduling) ■ Origin (e.g., instructor override, student command) ■ Timestamps and session data ■ Retention periods and access permissions These logs are essential for audits, incident resolution, and demonstrating regulatory compliance. Access must be limited to authorized personnel and retained for a defined audit window.
A.L.-4	Institutional Responsibility Policies	Universities must publish clear policies outlining: ■ Who is legally and contractually responsible for robot use and its outcomes. ■ What constitutes acceptable use or misuse of robot capabilities. ■ How harm or errors will be reported, investigated, and compensated if applicable. ■ How responsibilities are distributed throughout the robot lifecycle, including conformity assessments, deployment approval, and emergency procedures. These policies should be part of broader data governance and digital ethics charters.

Autonomy & Human Oversight (H.O.)

Maintains appropriate boundaries between automated support and human-led academic governance.

Table 16. Autonomy & Human Oversight (H.O.) Control Guidelines

Control ID	Control Name	Guideline Description
H.O.-1	Human Oversight for Critical Decisions	All high-impact educational decisions (e.g., academic probation alerts, final grade assignments) must involve a human checkpoint. This can be ensured by: ■ Configuring the robot to flag high-risk decisions for instructor review. ■ Enforcing dual-authentication before implementing sensitive decisions.
H.O.-2	Manual Override Capabilities	Human supervisors must be able to halt, revise, or reverse a robot's automated actions, particularly in case of: ■ Misbehavior or bugged behavior loops. ■ Student complaints or harm indicators. ■ Unexpected outcomes (e.g., incorrect attendance marking).
H.O.-3	Integrated Feedback and Error Reporting	Students and instructors must have access to an always-available channel for reporting robot issues. The system should: ■ Capture logs and screenshots tied to user reports. ■ Send alerts to assigned supervisors upon critical submissions. ■ Provide resolution status updates to the reporter.

Bias, Fairness & Inclusion (B.F.)

Ensures that robot assistants serve all students equitably and avoid replicating or amplifying social or algorithmic biases in educational processes.

Table 17. Bias, Fairness & Inclusion (B.F.) Control Guidelines

Control ID	Control Name	Guideline Description
B.F.-1	Use-Case Specific Bias Evaluation	Institutions must assess the robot assistant's performance across relevant use cases (e.g., grading, tutoring) to identify differential treatment. This includes: ■ Simulation-based testing for performance consistency across demographic groups (e.g., by gender, ethnicity, language fluency). ■ Scenario testing for behavioral interactions with diverse students. ■ Documentation of known edge cases or performance gaps.
B.F.-2	Inclusive Performance Audits	Institutions must perform regular performance audits across population groups. These audits should include: ■ Tracking system accuracy, error rates, or engagement outcomes by student subgroup. ■ Sampling user feedback from underrepresented groups. ■ Reporting and remediating identified disparities through tuning or retraining.
B.F.-3	Equity Monitoring Metrics	System dashboards and reports must include equity-focused KPIs such as: ■ Usage distribution by demographic. ■ False positive/negative rates in automated grading. ■ Helpfulness ratings by subgroup. These metrics should be monitored internally and reviewed quarterly.
B.F.-4	Inclusive Design Requirements	Robot interfaces and language models must be designed with universal access in mind. This includes: ■ Screen reader compatibility and voice control for students with visual or motor impairments. ■ Multi-language support or simplified English modes. ■ Gender-neutral voice and behavior defaults, along with periodic evaluation of embodiment and symbolic roles (e.g., ensuring that robot roles do not reinforce stereotypes based on appearance or assigned function). Developers must conduct inclusive design reviews prior to deployment.

Equity & Accessibility (E.A.)

Ensures that robot assistants are inclusively designed and accessible to all students, regardless of ability, background, or available resources. Institutions must actively identify and eliminate access disparities.

Table 18. Equity & Accessibility (E.A.) Control Guidelines

Control ID	Control Name	Guideline Description
E.A.-1	Inclusive Interface Design	Robot systems must support diverse accessibility needs to ensure equitable participation. This includes: ■ Screen reader and text-to-speech compatibility. ■ Voice control or switch input for students with mobility impairments. ■ Multi-language support for students with limited proficiency in the primary language of instruction. ■ Adjustable display brightness, contrast, and font size for visual impairments.
E.A.-2	Equal Service Distribution	Access to robot assistants must not be limited to a specific campus location or in-person students. Examples include: ■ Ensuring virtual equivalents (e.g., chatbot or video-linked robot assistants) for remote learners. ■ Deployment in satellite or off-campus facilities, not just central buildings. ■ Logging and analyzing usage to detect under-served groups.
E.A.-3	Socioeconomic Inclusion Measures	Institutions must ensure that robot-assisted learning or services do not create advantage gaps for students with fewer resources. This includes: ■ Avoiding reliance on student-owned smart devices or paid apps to access robot features. ■ Providing university-funded internet or device loans when robot assistants depend on digital access. ■ Making robot-enhanced services (e.g., guided navigation, tutoring) available to all students at no additional cost.
E.A.-4	Institutional Monitoring for Disparities	Regular evaluations must be conducted to ensure inclusive use of robotic systems. Institutions should: ■ Collect anonymous metrics on who uses the robots (e.g., by major, disability status, enrollment mode). ■ Run periodic accessibility reviews in collaboration with student diversity and disability offices. ■ Use findings to adapt deployment strategies or redesign features that exclude or discourage certain groups.

6. Evaluation: Results of Applying the Framework to Case Studies

This chapter applies the guideline-driven framework developed in Chapter 5 to three robot assistant deployments at Tallinn University of Technology (TalTech). Together, these case studies illustrate how the framework can be used to diagnose cybersecurity and ethical risks, and to derive concrete improvement steps in realistic HE settings.

The analysis focuses on:

1. the Attendance Check Application (ACA) deployed on a Temi robot in classroom settings;
2. a robot teaching assistant used to answer students' course-related questions during lab sessions;
3. a robot teaching assistant that autonomously evaluates student tasks and conducts oral knowledge assessment.

All three assessments were carried out by the author in close collaboration with Fuad Budagov, who was involved in the design and implementation of each deployment. Technical and procedural details were confirmed with him and other members of the TalTech team to ensure that the framework was applied to an accurate description of the systems.

For each case, the Control Evaluation Matrix and domain-specific guidelines from Chapter 5 were applied. To keep this chapter readable, detailed control-by-control tables are moved to the appendix; here, the emphasis is on cross-case patterns and an actionable roadmap toward framework compliance.

6.1 Overview of TalTech Case Studies

Attendance Check Application (ACA)

The Attendance Check Application (ACA) is a robot assistant-based attendance system deployed on a Temi V3 platform in a HE classroom. In the initial pilot, students voluntarily registered their face with the robot, which then used facial recognition to mark attendance when they approached the device before class. Attendance records and biometric templates were stored locally on the robot, with access restricted to the research team. While usability

feedback was positive, many students expressed neutrality or concern about data privacy and long-term biometric data use, indicating early trust and consent challenges.

Building on this pilot, a newer ACA version integrates with Estonia's national ID card infrastructure. Instead of storing student faces, the robot reads the ID card, retrieves the official photo, and performs a live 1:1 face comparison on-device. Once a match is confirmed, the system records that *student X was present at time Y* and discards the images, avoiding persistent biometric storage. This architectural shift significantly reduces biometric retention risk and strengthens identity assurance.

However, the updated design introduces new questions around the use of national ID data in an academic environment, the boundaries of consent, and the robustness of supporting infrastructure. The ACA case therefore provides a rich test-bed for the framework's cybersecurity and ethical control domains, especially data privacy, human oversight, transparency, accessibility, and institutional governance.

Robot Assistant for Answering Student Questions

The second case study is a Temi-based robot teaching assistant that answers students' questions during practical laboratory sessions using the Interactive Mobile Teaching Assistant (IMTA) system. The robot operates in a pre-mapped computer lab, where students can walk up to Temi and pose spoken questions via a voice user interface (VUI). When a student asks a question, the robot transcribes the utterance, retrieves relevant course materials via semantic search, constructs a prompt, and queries an LLM for an answer. The answer is then sent back to the robot, which both speaks it aloud and displays it on the tablet.

During the pilot, the system was deployed in multiple 90-minute lab sessions. Around half of the enrolled students chose to interact with the robot, posing a modest number of questions per session. Most of the course-related questions were answered correctly as verified by the course lecturer, with only a small fraction misinterpreted due to speech recognition errors or ambiguous phrasing. Post-pilot UX survey results indicated that students generally found the robot easy to use and helpful, and they reported positive attitudes toward having a robot assistant in the lab. At the same time, usability scores were moderate, and students reported issues with response timing, speech recognition sensitivity, and the system's handling of complex or contextually nuanced questions.

From this thesis's perspective, the pilot represents a pedagogically significant robot assistant that mediates knowledge access, influences engagement, and introduces new data

flows and cloud dependencies.

Robot Assistant for Automated Task Evaluation

The third case study examines a robot teaching assistant that autonomously evaluates student configurations and conducts oral assessments in a computer networking lab using IMTA. In this setup, the Temi robot is stationed at the classroom entrance and waits in an idle state. When a student requests evaluation (e.g., by saying "I want to defend work" and providing their desk number), the robot navigates autonomously to the specified workstation.

The evaluation process follows a three-stage workflow:

1. **Configuration check:** The robot asks the student to connect a USB adapter between Temi and the networking device. IMTA retrieves the live configuration via console commands, compares it against a stored reference configuration, and tests connectivity via ping requests. Major configuration parameters must be correct for the work to be accepted.
2. **Oral examination:** If the configuration passes, the robot proceeds to an oral knowledge assessment, drawing open-ended questions from a question bank implemented in IMTA. The student is asked up to five questions and must answer at least three correctly for the oral exam to be validated.
3. **Feedback delivery:** Based on configuration and oral exam results, the robot provides structured feedback, such as identifying errors, indicating failure with referral to instructors, or confirming successful validation.

During the pilot, the robot handled a substantial number of evaluation requests. Most connection attempts succeeded and advanced to automated configuration checks, with a subset proceeding to oral examinations, resulting in nearly two hundred robot-posed questions. Performance data show the robot reliably completed configuration checks and oral exams in most cases, with few failures caused by speech recognition or network interruptions. Survey responses indicated that students found the robot helpful and fair but emphasized the need for improved speech interaction and robustness in noisy lab environments. Importantly, as this was a pilot, negative or ambiguous outcomes were manually rechecked by the instructor to protect students from unfair grading.

This case is particularly relevant for the EU AI Act perspective: even with the pilot limitations, the robot performs task evaluation and oral assessment in a way that is structurally similar to high-risk educational AI systems. It therefore provides a realistic context to

test how the proposed framework can capture and mitigate risks around fairness, human oversight, accountability, and data protection.

6.2 Summary of Framework Application Across Case Studies

Across all three deployments, the full Control Evaluation Matrix and domain guidelines from Chapter 5 were applied. The assessments draw on system documentation, published case studies, and the TalTech project team's experience, with key technical and organisational details validated through discussions with collaborators. In all cases, scores were assigned using the standardised 0–3 scale, but this section focuses on qualitative patterns rather than numeric values (detailed matrices are provided in the appendices).

Cybersecurity Domains

Data Privacy & Confidentiality (D.P.). The ACA deployment shows strong progress: the newer ID-card integration minimizes biometric retention by performing face matching in real time and discarding images immediately. Attendance records are purpose-bound and need not include biometric data at all. In the IMTA-based pilots, no biometric or directly identifying data are collected; logs are stored in a university database and linked to desk numbers or session IDs. Participation is voluntary and governed by ethics approval and GDPR-compliant consent procedures. Taken together, the three cases show relatively mature data minimisation practices and a clear separation between instructional data and sensitive identifiers.

System & Data Integrity (S.I.) and System Access & Control (S.A.). By contrast, integrity and access controls emerge as weaker, cross-cutting areas. None of the deployments yet implements secure, signed update pipelines, explicit runtime integrity checks, or systematic isolation of security-critical components (e.g., biometric matching logic or IMTA services). Administrative access to the robots and backend systems relies largely on standard OS-level accounts and project-based practices. The IMTA-based deployments similarly do not describe hardened update or key management processes; their focus is on functional robustness and pedagogical outcomes rather than formal DevSecOps controls. In framework terms, these domains tend to cluster around "minimally addressed" to "partially fulfilled" rather than fully implemented.

Vulnerability & Patch Management (V.P.). All three deployments depend on complex software stacks: the Temi operating system, custom IMTA components, networking libraries, and large language model infrastructure. Although teams maintain system functionality and monitor failures, evidence of structured vulnerability management such

as CVE tracking, software bills of materials, or documented patch timelines Evidence is mixed: ACA shows strong patching/baseline practices, but lacks vulnerability scanning, while the IMTA-based pilots show more consistently ad hoc vulnerability management. Given the relatively small scale of current pilots, this is understandable but becomes a significant gap if similar systems are scaled across programmes or campuses.

Network & Communication Security (N.C.). IMTA-based deployments are network-intensive, relying on local Wi-Fi, console access to networking devices, and outbound connections to server-side IMTA components and cloud-based LLM services. Pilot results explicitly report occasional failures or delays due to network interruptions and serial connection issues. From the framework's perspective, this raises questions about encrypted communication enforcement, certificate validation, and robustness against misconfigured or insecure networks. For the ACA system, which now integrates with national ID infrastructure, ensuring that all communication with state systems and institutional databases uses modern, strictly enforced encryption becomes a central requirement.

Cyber-Physical Safety (C.P.). In all three cases, the robot operates in controlled indoor environments with limited speed and established navigation maps. Students approach the robot voluntarily, and interaction patterns are constrained (e.g., scanning an ID card, connecting a console cable, answering questions). There are no reports of safety incidents, and physical risk is inherently low. As a result, Cyber-Physical Safety is high for ACA and moderate for the IMTA-based deployments, with remaining gaps mainly in formal safety assurance, documentation, and review..

Institutional Readiness & Governance (I.R.). The case studies also reveal an emerging but incomplete governance layer. Each deployment was reviewed by the university's ethics committee and involved collaboration between researchers, lecturers, and IT staff. Responsibility structures remain project-based, with no university-wide robot assistant policy, governance body, or standardized process for adopting the framework when new systems are proposed. As the number and scope of robot assistants grows, these gaps risk becoming a bottleneck for AI Act compliance.

Ethical & Sociotechnical Domains

Privacy, Consent & Surveillance (P.C.). All pilots follow baseline ethical research practices: students are informed of the purpose, participation is voluntary, and consent is obtained in accordance with GDPR and institutional requirements. In the ACA pilot, students expressed notable concern or ambivalence about biometric data, which motivated the move to the ID-based architecture. At the same time, None of the systems offer self-service tools

for students to inspect, correct, or delete personal data; redress is possible only through lecturers or data protection officers. Surveillance cues are minimal: although cameras or microphones are visible, students lack explicit indicators or controls beyond choosing whether to use the robot.

Transparency & Explainability (T.E.). Role transparency is high across deployments, as students clearly understand they are interacting with a robot performing support or evaluation tasks for the institution. Decision logic is documented in papers and internal technical descriptions (e.g., how configuration checks or RAG pipelines work), but this documentation is not exposed to end-users in a structured way. The ACA system provides no user-facing explanations for failed face matches or attendance errors, and the evaluation robot offers only generic feedback for insufficient results. The Q&A assistant can show answer text and source passages but does not provide structured justifications or confidence indicators. Overall, the deployments are transparent about *what* the robot is doing but weak on accessible explanations of *why* particular outcomes occurred.

Accountability & Human Oversight (A.L., H.O.). A positive pattern across all three cases is the consistent presence of human oversight. Instructors retain responsibility for courses, can verify or override robot outputs, and are directly involved when negative outcomes occur, such as failed evaluations. The ACA and IMTA-based systems are all embedded in supervised teaching contexts rather than operating independently. However, these arrangements are informal: roles such as "system supervisor", "appeals handler", or "log reviewer" are not formally assigned, and appeals are handled conversationally rather than via documented procedures. Human oversight is present in all cases, but the formality differs: ACA and the Q&A assistant rely mainly on informal instructor-mediated redress, while the automated evaluation deployment includes a clear, built-in review pathway for negative outcomes. As these systems move from research pilots toward routine use, formalising accountability structures will be essential to align with AI Act expectations.

Bias, Fairness & Inclusion (B.F.). The ACA deployment includes fairness validation under the research programme, including segmented monitoring during trials, but this is not yet institutionalised as a university-wide operational audit.

In the IMTA-based pilots, fairness issues emerge primarily through speech recognition and interaction design. Students with strong accents or varied speech patterns may experience misrecognitions, which in evaluation contexts could lead to unfair outcomes if not carefully mitigated. The Q&A and evaluation studies both report that speech-related issues were a primary source of technical errors and student frustration, even though overall perceptions of fairness and usefulness remained positive. There is no evidence of systematic fairness

auditing by demographics or accessibility needs, although the the automated evaluation pilot includes a strong use-case fairness assessment (alignment with instructor grading and student fairness perceptions), but does not perform subgroup disparity audits.

Equity & Accessibility (E.A.). All three deployments are tied to physical presence in a specific lab or classroom. Remote students, those on satellite campuses, or students who cannot comfortably interact with a voice-based robot have limited or no access to equivalent support. Despite simple interfaces, the systems rely on speech and standard visuals, lacking explicit support for screen readers, alternative inputs, or multilingual accessibility. Planned future work includes a silent interaction mode using on-screen controls, partially addressing accessibility concerns, though a comprehensive equity strategy remains undeveloped.

6.3 Actionable Compliance Roadmap

Building on the case study evaluations, this section translates the framework into an actionable plan for TalTech (and similar institutions) to bring robot assistant deployments into closer alignment with the proposed controls and the EU AI Act. The recommendations are structured along clusters of control domains and assume a multi-year, iterative process.

Institutionalise Governance and Risk Management

A university-wide robotics governance structure should be established to ensure that robot assistant deployments are coordinated, accountable, and aligned with both institutional priorities and regulatory requirements. In practice, this can take the form of a cross-functional committee that includes faculty representatives, IT security specialists, legal and data protection experts, and student representation. This body should be responsible for approving new pilots, overseeing ongoing deployments, and maintaining traceable records of compliance activities.

In parallel, each robot deployment should be formally classified under the EU AI Act risk taxonomy, with particular attention to whether the system may fall within, or closely resemble, high-risk use cases in education (e.g., systems that influence evaluation, access, or learning outcomes). Even where a system is not yet used to assign binding grades, deployments such as the ACA and evaluation robot should be treated as at least "AI-act-adjacent" high-risk scenarios, given their functional proximity to educational assessment.

To operationalise this governance model, the Control Evaluation Matrix should be integrated into project initiation procedures. Specifically, new robot projects should be required to complete an initial control assessment and submit a remediation plan as a

precondition for deployment, using the case study matrices as reusable templates. Finally, each deployment should have named human supervisors. A designated system owner should be formally documented as responsible for oversight, periodic log review, and incident handling, with these responsibilities specified in course-level and institutional policies rather than left to informal assumptions.

Strengthen Technical Security Controls

Technical hardening should focus on improving baseline integrity, access control, secure communications, and vulnerability management across both robot platforms and supporting institutional infrastructure. First, signed update and integrity pipelines should be introduced (S.I., V.P.). This includes cryptographically signed software and firmware updates, checksum verification, and basic runtime integrity monitoring for the robot platform as well as custom applications (e.g., ACA and IMTA). Changes to critical components should be logged and handled through established change management procedures to preserve traceability.

Second, access control to robots and supporting backends should be hardened (S.A.). Administrative interfaces on both robot systems and server-side components should enforce strong authentication, preferably using multi-factor mechanisms. Access should be constrained according to least-privilege principles, and role definitions and responsibilities should be documented to support auditing and accountability.

Third, secure communication patterns should be enforced across robot-to-service links (D.P., N.C.). All communication between robots, IMTA services, institutional databases, and external AI providers should use modern TLS configurations with strict validation and no downgrade or fallback to deprecated protocols. These controls should be complemented by monitoring for insecure network configurations and logging mechanisms capable of flagging anomalies.

Finally, vulnerability and patch management should be formalised (V.P.). Each deployment should maintain a software bill of materials (SBOM), track relevant advisories, and define timelines for patch application based on severity. Where feasible, robot systems should be integrated into the institution's existing vulnerability scanning and incident response workflows rather than handled as exceptions.

Enhance Transparency, Consent, and User Rights

Compliance and trustworthy deployment also depend on measures that support transparency, meaningful consent, and student rights in practice. Privacy notices and consent

flows should therefore be standardised (P.C.). Institutions should develop templates that explain, in student-facing terms, what each robot does, what data it collects, the retention period, and how the system may affect learning processes or evaluation. To ensure accessibility and visibility, these notices should be made available through multiple channels, including robot interfaces, course syllabi, and learning management systems.

In addition, explanation and appeal mechanisms should be implemented for systems with evaluative or attendance-related functions (T.E., A.L.). At a minimum, students should have access to structured, human-readable explanations for adverse or unexpected outcomes (e.g., why attendance was not registered, or why an evaluation failed). These explanation features should be coupled with explicit appeal pathways, such as documented contact points or online forms embedded into course materials, to ensure that procedural fairness is not merely theoretical.

Finally, data access and correction mechanisms should be defined as an operational requirement (P.C.). Even if full self-service portals are not feasible in early stages, students should be given clear instructions on how to request access to interaction logs and how to correct errors in attendance or evaluation records.

Embed Fairness, Accessibility, and Alternatives

Fairness and accessibility requirements should be treated as design obligations rather than optional enhancements. One practical step is to introduce multimodal interaction options (B.F., E.A.). IMTA-based systems should be extended beyond voice interaction to include alternatives such as on-screen controls or typed input via connected devices. This supports students with speech and hearing impairments, those with accent-related challenges, and users operating in noisy environments.

Accessibility considerations must also address modality and presence. Remote and hybrid students should have equivalent service channels (E.A.) for attendance and support functions, such as authenticated online check-ins or chat-based IMTA access, enabling comparable outcomes without requiring physical interaction with the robot.

For systems that influence evaluation, regular fairness audits should be conducted (B.F.). This entails tracking error rates and interaction outcomes across relevant student subgroups, such as language background and disability status, where collection and analysis are ethically and legally appropriate. The purpose of such monitoring is to identify systematic disparities and implement mitigations, for example through speech model tuning or interaction script adjustments.

Crucially, robot-mediated services should not become exclusive gateways. Human-equivalent pathways should remain available in parallel to robot services, including manual attendance procedures, direct instructor Q&A, or human-led evaluation sessions. Maintaining these alternatives ensures that students who opt out for privacy, accessibility, or trust reasons are not disadvantaged.

Operational Monitoring and Continuous Improvement

The framework should be implemented as a continuous improvement cycle rather than a one-time compliance exercise. Periodic re-evaluations should therefore be scheduled, treating the Control Evaluation Matrix as a living tool. The assessment should be repeated after significant changes (e.g., a new LLM provider, new identity integration, or new course contexts) and at regular intervals (e.g., annually) to track progress over time.

Moreover, student and lecturer feedback should be integrated into ongoing risk management. Instruments such as UX and SUS surveys should be used not only for evaluation in a research context but also as operational monitoring signals. Reports of reliability issues, perceived unfairness, or discomfort should feed into governance decisions and technical roadmaps.

Finally, institutions should document incidents and near misses through lightweight but consistent procedures. This includes logging events such as misrecorded attendance, incorrect evaluation outcomes, or severe misunderstandings in Q&A interactions. Systematic analysis of such patterns should inform updates to controls, training, and operational safeguards.

6.4 Implications for the Framework

The three TalTech case studies demonstrate that the proposed framework is practically usable across diverse robot assistants, including administrative attendance systems, Q&A agents, and evaluative teaching assistants. They also reveal a recurring pattern: privacy and basic safety mature early, whereas system integrity, governance, transparency, fairness auditing, and accessibility require sustained institutional effort.

By translating control scores into a concrete roadmap, this chapter demonstrates how the framework moves beyond assessment to guide incremental, realistic improvements in technical design and organisational practice. The next chapter builds on these findings to discuss broader implications for robot assistants in HE and the evolving regulatory landscape.

7. Discussion

This research has explored the complex landscape of ethical and cybersecurity challenges surrounding the deployment of robot assistants in HE. Through a structured literature review, EU regulatory analysis, and control-based assessment, the study addressed three guiding research questions.

RQ1: What ethical and cybersecurity challenges arise from deploying robot assistants in higher education, particularly regarding autonomy, bias, surveillance, transparency, and system-level vulnerabilities?

The integration of robot assistants in HE introduces challenges that are socio-technical in nature. Autonomy in robotic systems, while functionally desirable, raises concerns over human oversight and responsibility attribution. As examined in the thesis, robots that perform tasks such as attendance-taking, grading, or tutoring must remain within a controlled loop of human validation. The illusion of machine authority can prompt uncritical compliance, risking inappropriate delegation of pedagogical or disciplinary authority. These risks are amplified in socially assistive or humanoid robots that mimic human expression or behavior.

Bias particularly algorithmic and representational bias also emerged as a pervasive risk. Systems trained on narrow demographic datasets can perform poorly when interacting with diverse student populations, leading to uneven experiences across racial, linguistic, or neurodiverse lines. Further, gendered embodiment and anthropomorphization of robot assistants can reinforce stereotypes, as evidenced by research on how robots performing instructional roles are more often perceived as male, while those offering emotional support are interpreted as female-coded. Such biases are not incidental but structurally embedded in how datasets are curated and interaction models are trained.

The dimension of surveillance presents one of the most potent ethical tensions. Telepresence robots, mobile AI platforms, and context-aware classroom assistants often feature multimodal sensing visual, auditory, spatial which, even when passively activated, blur the boundaries between assistance and monitoring. The perception of being watched, even in the absence of active recording, can influence student behavior, raise anxieties, and deteriorate trust. These systems not only collect sensitive biometric and behavioral data but may do so without sufficient transparency or consent protocols.

On the cybersecurity front, vulnerabilities are rooted both in system architecture and institutional readiness. Many robots operate with legacy security defaults—unencrypted data channels, unsecured ports, or unsegmented network deployments. The case study revealed substantial deficiencies in access control, patch management, and audit logging. A robot that is compromised can serve as an entry point for broader attacks or expose student data to breaches. Furthermore, the absence of formal incident response protocols and governance policies leaves institutions unprepared for escalations when robotic systems malfunction or are exploited.

RQ2: What implications does the European Union Artificial Intelligence Act have for robot assistants in higher education, and what key compliance considerations must institutions and developers be aware of?

The EU Artificial Intelligence Act marks a significant turning point in how robotic and AI systems particularly those deployed in education are designed, implemented, and governed. It translates broad ethical concerns into enforceable legal requirements and repositions robot assistants not merely as innovations, but as high-risk systems operating within a tightly regulated domain. Under Annex III of the Act, robot assistants used in HE are classified as "high-risk AI systems," triggering a comprehensive set of compliance obligations across their entire lifecycle.

Central to these obligations are Articles 9, 10, and 14. Article 9 mandates a continuous risk management system, requiring institutions to move beyond reactive safeguards and proactively identify, evaluate, and mitigate risks to health, safety, and fundamental rights. This process is ongoing extending from pre-deployment design through real-world testing, post-market monitoring, and regular updates. Institutions, therefore, become long-term custodians of ethical and secure system behavior. Article 10 introduces rigorous data governance expectations, requiring that training and validation datasets be representative, unbiased, and complete in the context of deployment. For robot assistants in classrooms, this necessitates careful attention to demographic diversity and fairness particularly when sensitive personal data is involved for bias detection. The regulation permits limited use of such data but demands stringent privacy, access, and security controls throughout.

Article 14 reinforces the necessity of human oversight, insisting that AI systems remain auditable, interruptible, and correctable. Robot assistants must be equipped with mechanisms that allow educators and students to review and challenge decisions such as identity verification or attendance logging—and must include traceable logs, alerts, and override

functions to ensure accountability. These are not optional features; they are mandatory elements of system integrity and trust.

Beyond these articles, the Act's broader framing captured in Recitals 142 and 165 signals a shift toward inclusive, socially beneficial design. Institutions are encouraged not just to meet baseline requirements but to adopt voluntary codes of conduct, embrace interdisciplinary collaboration, and prioritize accessibility and equity. In HE, where values like inclusion, academic freedom, and student welfare are foundational, this means aligning AI use with institutional ethics. Robot deployment becomes not merely a technical project but an educational and civic responsibility.

This thesis only scratches the surface of what this regulatory framework entails. The full implications point to a substantial transformation in how universities and developers engage with intelligent systems. Institutions will need to establish governance structures that encompass legal compliance, algorithmic audit, cybersecurity oversight, and participatory design. Developers will be called upon to integrate explainability, fairness metrics, and secure-by-design practices from the outset. Policymakers and sector leaders may be required to provide domain-specific guidance, while academic collaborations may become vital for sharing best practices and building collective compliance capacity.

Ultimately, the EU AI Act redefines responsible innovation in education. Robot assistants are no longer just interactive tools they are systems operating within a dense matrix of rights, risks, and responsibilities. Navigating this terrain demands not only compliance with legal frameworks, but a commitment to the values that underpin HE itself. In this sense, the Act serves as both a legal mandate and a moral compass, guiding institutions toward a future where AI supports human learning without compromising human dignity.

RQ3: How can individuals, institutions, or developers approach the design and deployment of robot assistants in a way that ensures ethical alignment and cybersecurity resilience in higher education?

This thesis proposes a guideline-driven framework to address precisely that question. The control-based structure introduced organized around domains like privacy, transparency, bias mitigation, access control, and governance allows institutions to translate abstract principles into actionable practices. Each sub-control provides measurable expectations and recommended safeguards, facilitating both pre-deployment review and post-deployment audit.

Institutions should adopt a multi-layered strategy that begins with participatory design. This means engaging students, faculty, and accessibility advocates from the outset to shape system requirements. Technical teams must conduct equity assessments of training data and interaction models, incorporating fairness metrics as a core validation criterion. Privacy-by-design must be a foundational norm, not an afterthought—robotic systems must collect the minimum data required, provide opt-in mechanisms, and visibly indicate when sensing features are active.

Cybersecurity hardening is equally vital. Robots should be onboarded as enterprise devices with strict access controls, encrypted communications, regular patching, and monitored activity logs. Incident response protocols should be rehearsed through drills, and responsibility for robot governance must be assigned formally within institutional structures.

Finally, oversight mechanisms both technical and human must be embedded across the system lifecycle. Real-time explainability tools, manual override features, and logging interfaces empower users and administrators to understand, question, and correct AI behavior. These are not only best practices they are required under the EU AI Act and vital for trust in academic settings.

In conclusion, ethical and cybersecurity alignment in robot assistant deployment is not a static goal but a dynamic process. Through structured evaluation, legal compliance, and participatory design, educational institutions can transition from reactive governance to proactive stewardship of intelligent systems.

8. Conclusion

This thesis investigated the ethical and cybersecurity dimensions of deploying robot assistants in HE. Through detailed analysis, it identified key risk areas including autonomy, surveillance, bias, and system vulnerabilities and mapped them against the regulatory framework of the EU AI Act. A multi-dimensional control framework was introduced and applied to a case study deployment, revealing both promising strengths and critical gaps. The findings affirm that while robot assistants hold significant potential to enhance education, they also introduce risks that must be actively managed through technical safeguards, institutional policy, and legal compliance.

While this thesis offers a structured evaluation and regulatory mapping, its scope is inherently limited. The case studies, while illustrative, are narrow in application and do not encompass the full diversity of robot use cases across global educational contexts. Furthermore, the framework was applied primarily from a developer and institutional perspective, without direct input from students, faculty, or disability groups a gap that future work should address through broader participatory research. Additionally, the analysis focused primarily on current regulatory frameworks and did not explore in depth the complexities of cross-border AI governance or emerging risks in multi-agent systems and generative AI. Future research should aim to validate the proposed control model across diverse deployments, expand its integration with formal assurance mechanisms, and explore interdisciplinary co-design methodologies. There is also a growing need for tools that translate legal compliance (like that mandated by the EU AI Act) into practical development workflows for academic IT teams.

In conclusion, this thesis provides a ready-to-use framework for evaluating the ethical and cybersecurity dimensions of robot assistants, along with a detailed examination of the challenges posed by their integration into educational environments. It offers both a conceptual and practical foundation for institutions seeking to align emerging technologies with legal, ethical, and operational expectations. As educational institutions increasingly embrace automation, their commitment must be to innovation with integrity. This work contributes a meaningful step toward that future equipping stakeholders with the tools and insights necessary to ensure that the deployment of intelligent systems strengthens, rather than undermines, the core values of HE.

References

[1] J. López-Belmonte et al. “Robotics in education: A scientific mapping of the literature in Web of Science”. In: *Electronics* 10.3 (2021), p. 291. DOI: 10.3390/electronics10030291.

[2] David Scaradozzi, Laura Screpanti, and Lorenzo Cesaretti. “Towards a Definition of Educational Robotics: A Classification of Tools, Experiences and Assessments”. In: *New Perspectives in Science Education*. Ed. by Letizia Cinganotto. Springer, 2019, pp. 35–45. DOI: 10.1007/978-3-030-19913-5_3.

[3] Despoina Schina et al. “The Integration of Sustainable Development Goals in Educational Robotics: A Teacher Education Experience”. In: *Sustainability* 12.23 (2020), p. 10085. DOI: 10.3390/su122310085.

[4] Barbara Kitchenham. “Procedures for Performing Systematic Reviews”. In: (2004).

[5] Fuad Budagov et al. “A Systematic Literature Review on Applicability of Robot Assistants in Higher Education”. In: *Methodologies and Intelligent Systems for Technology Enhanced Learning, 14th International Conference*. Cham: Springer Nature Switzerland, 2024, pp. 21–32. ISBN: 978-3-031-73538-7. DOI: 10.1007/978-3-031-73538-7_3.

[6] Claes Wohlin. “Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering”. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE)*. 2014, pp. 1–10. DOI: 10.1145/2601248.2601268.

[7] Fuad Budagov et al. “Attendance Check of Students via Robot Assistant in Higher Education Classes”. In: *Futureproofing Engineering Education for Global Responsibility*. Ed. by Michael E. Auer and Tiia Rüütmann. Cham: Springer Nature Switzerland, 2025, pp. 305–316. ISBN: 978-3-031-85652-5. DOI: 10.1007/978-3-031-85652-5_31.

[8] Janika Leoste et al. “Usage Scenarios for Robot Assistants in Higher Education Settings”. In: *Futureproofing Engineering Education for Global Responsibility*. Cham: Springer Nature Switzerland, 2025, pp. 326–337. ISBN: 978-3-031-98761-8. DOI: 10.1007/978-3-031-98762-5_28.

[9] Tiina Kasuk et al. “Telepresence Robots and Inclusive Hybrid Learning: Bridging Gaps in Higher Education Classrooms”. In: *Futureproofing Engineering Education for Global Responsibility*. Cham: Springer Nature Switzerland, 2025, pp. 498–507. ISBN: 978-3-031-98761-8. DOI: 10.1007/978-3-031-98762-5_42.

[10] Arizona Western College. *Meet MATABOT: SLLC's New Personal Assistant Robot*. Accessed: 2025-04-15. 2023. URL: https://www.azwestern.edu/news/meet-matabot-sllcs-new-personal-assistant-robot.

[11] Rebecca Brent and Richard Felder. "Using Telepresence Robots to Support Students Facing Adversity". In: *EDUCAUSE Review* (2018). URL: https://er.educause.edu/articles/2018/6/using-telepresence-robots-to-support-students-facing-adversity.

[12] Double Robotics. *Double 3 – Telepresence Robot*. Accessed: 2025-04-15. 2024. URL: https://www.doublerobotics.com/double3.html.

[13] OhmniLabs. *OhmniCare Mobile Telehealth Robot*. Accessed: 2025-04-15. 2024. URL: https://ohmnilabs.com/products/ohmnicare-mobile-telehealth-robot/.

[14] OhmniLabs. *OhmniLabs Developer Documentation*. Accessed: 2025-04-15. 2024. URL: https://docs.ohmnilabs.com/.

[15] FCC Filing / Inbot Technology. *PadBot X1 User Manual*. Accessed: 2025-04-16. 2020. URL: https://fcc.report/FCC-ID/2AWJ6-PADBOTX1/4774180.pdf.

[16] Inbot Technology. *PadBot Official Website*. Accessed: 2025-04-16. 2024. URL: https://m.padbot.com/.

[17] Awabot. *Beam Telepresence Robot – Awabot Robotics*. Accessed: 2025-04-16. 2024. URL: https://awabot.com/en/.

[18] Suitable Technologies. *Beam+ User Documentation*. Accessed: 2025-04-16. 2024. URL: https://suitabletech.com/support/documentation.

[19] Robotemi. *Temi Personal Robot*. Accessed: 2025-04-15. 2024. URL: https://www.robotemi.com/.

[20] Robotemi. *Temi User Manual (2022)*. Accessed: 2025-04-15. 2022. URL: https://robotemi.com.tw/wp-content/uploads/2022/01/temi_User_Manual_20220110%E7%89%88.pdf.

[21] BlueMed. *InTouch Health – Telemedicine Robots Overview*. Accessed: 2025-04-16. 2024. URL: https://bluemed.cz/intouch-health/.

[22] InTouch Health. *RP-VITA Remote Presence System: User Guide*. Accessed: 2025-04-16. 2021. URL: https://cloud.kapostcontent.net/pub/29bf3ab3-5974-4a21-9be7-a5c395d8d972/product-brochure-vita-user-guide.pdf.

[23] Revolve Robotics. *Kubi Telepresence Robot — Technical Specifications*. Accessed: 2025-04-15. n.d. URL: https://avcommsolutions.com/specs/RR40-1001_m.pdf.

[24] Laurent Gallon et al. "Using a Telepresence Robot in an Educational Context". In: *Proceedings of the 15th International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*. Available on HAL: hal-02410364. Las Vegas, United States, July 2019. URL: https://hal.science/hal-02410364.

[25] Thomas Wernbacher et al. "TRinE: Telepresence Robots in Education". In: *Proceedings of the 16th Annual International Technology, Education and Development Conference (INTED)*. IATED, Mar. 2022, pp. 6514–6522. DOI: 10.21125/inted.2022.1653.

[26] SoftBank Robotics. *Pepper Robot*. Image licensed under CC BY-SA 4.0, accessed 2025-04-16. 2015. URL: https://commons.wikimedia.org/wiki/File:Robot_pepper.jpg.

[27] Proven Robotics. *Pepper the Humanoid Robot – In Education*. Accessed: 2025-04-16. 2024. URL: https://provenrobotics.ai/pepper-robot-education/.

[28] Hagen Lehmann and Pier Giuseppe Rossi. "Social Robots in Educational Contexts: Developing an Application in Enactive Didactics". In: *Journal of e-Learning and Knowledge Society* 15.2 (2019), pp. 27–41. DOI: 10.20368/1971-8829/1633.

[29] Amit Kumar Pandey and Rodolphe Gelin. "A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind". In: *IEEE Robotics & Automation Magazine* 25.3 (2018), pp. 40–48. DOI: 10.1109/MRA.2018.2833157.

[30] Hossein Banaeian and Ilkay Gilanlioglu. "Influence of the NAO Robot as a Teaching Assistant on University Students' Vocabulary Learning and Attitudes". In: *Australasian Journal of Educational Technology* 37.3 (2021), pp. 71–87. DOI: 10.14742/ajet.6130.

[31] Proven Robotics. *NAO Humanoid Robot for Education*. Accessed: 2025-04-16. 2024. URL: https://provenrobotics.ai/nao-robot-education/.

[32] Sara Cooper et al. "ARI: the Social Assistive Robot and Companion". In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2020, pp. 745–751. DOI: 10.1109/RO-MAN47096.2020.9223470.

[33] PAL Robotics. *PRO-CARED pilots ARI as an education robot with Catalan language*. Accessed: 2025-04-16. 2023. URL: https://pal-robotics.com/blog/pro-cared-pilots-robot-ari-as-education-robot-with-catalan-language/.

[34] PAL Robotics. *ARI – The Social Robot*. Accessed: 2025-04-16. 2024. URL: https://pal-robotics.com/robot/ari/.

[35] SIFSOF. *Humanoid Smart Catering Services Robot - SIFROBOT-5.2*. https://sifsof.com/product/humanoid-smart-catering-services-robot-sifrobot-5-2/. Accessed: 2025-04-16. 2024.

[36] Mukhtar Ibrahim Bello et al. "A Review on the Humanoid Robot and its Impact". In: *Global Journal of Research in Engineering & Computer Sciences* 4.6 (2024), pp. 16–20. DOI: 10.5281/zenodo.14039558.

[37] H. Anjanappa. "Case Study of Sophia – The Humanoid Robot". In: *Proceedings of the National Conference on E-Business, E-Commerce and Management*. Secunderabad, India: St. Martin's Engineering College, 2018, pp. 1–6. URL: https://www.smec.ac.in/assets/images/committee/research/17-18/539.Case%20Study%20of%20Sophia%20%E2%80%93%20The%20Humanoid%20Robot.pdf.

[38] Hanson Robotics. *Sophia the Robot – Hanson Robotics*. Accessed: 2025-04-16. 2024. URL: https://www.hansonrobotics.com/sophia/.

[39] Benoit Heintz. "Development of New Functionalities for the Humanoid Robot ROMEO". Supervisor: Prof. Basilio Bona. Master's thesis. Turin, Italy: Politecnico di Torino, 2017. URL: https://webthesis.biblio.polito.it/6620/.

[40] Erico Guizzo. "By Leaps and Bounds: An Exclusive Look at How Boston Dynamics is Redefining Robot Agility". In: *IEEE Spectrum* 56.12 (2019), pp. 34–39. DOI: 10.1109/MSPEC.2019.8913831.

[41] Toyota Motor Corporation. *Toyota Develops Third-Generation Humanoid Robot THR-3*. Accessed: 2025-04-16. 2019. URL: https://global.toyota/en/detail/19666346.

[42] Robert Bogue. "Humanoid robots from the past to the present". In: *Industrial Robot: The International Journal of Robotics Research and Application* 47.4 (2020), pp. 465–472. DOI: 10.1108/IR-05-2020-0088.

[43] SoftBank Robotics. *Pepper Developer Documentation*. Accessed: 2025-04-16. 2024. URL: http://doc.aldebaran.com/2-5/home_pepper.html.

[44] Wikimedia Commons Contributors. *Starship Robot Up Close*. Licensed under CC BY-SA 4.0. 2018. URL: https://commons.wikimedia.org/wiki/File:Starship_Robot_Up_Close.jpg.

[45] Joanne Pransky. "The Pransky interview: Dr Steve Cousins, CEO, Savioke, Entrepreneur and Innovator". In: *Industrial Robot: The International Journal of Robotics Research and Application* 43.1 (2016), pp. 1–5. DOI: 10.1108/IR-11-2015-0196.

[46] Relay Robotics. *Relay Autonomous Delivery Robots*. Accessed: 2025-04-16. 2024. URL: https://relayrobotics.com/.

[47] Tathagata Chakraborti et al. "A Formal Framework for Studying Interaction in Human-Robot Societies". In: *Proceedings of the AAAI Spring Symposium on Symbiotic Cognitive Systems (SSS-16)*. 2016, pp. 737–741. URL: https://asu.elsevierpure.com/en/publications/a-formal-framework-for-studying-interaction-in-human-robot-societ.

[48] Aethon Inc. *Resources – Aethon T3 Autonomous Intralogistics Robot*. Accessed: 2025-04-16. 2024. URL: https://aethon.com/resources/.

[49] Keenon Robotics. *Dinerbot T8 - Service Robot*. Accessed: 2025-04-16. 2025. URL: https://www.lotsofbots.com/en/keenon-robotics/dinerbot-t8/.

[50] Ralf Kittmann et al. "Let Me Introduce Myself: I Am Care-O-bot 4, a Gentleman Robot". In: *Mensch und Computer 2015 – Tagungsband*. Ed. by Martin Pielot, Sarah Diefenbach, and Niels Henze. Berlin, Germany: De Gruyter Oldenbourg, 2015, pp. 223–232. DOI: 10.1515/9783110443929-024. URL: https://doi.org/10.1515/9783110443929-024.

[51] Fraunhofer IPA / Mojin Robotics. *Care-O-bot 4*. Accessed: 2025-04-16. 2024. URL: https://www.care-o-bot.de/en/care-o-bot-4.html.

[52] Alan-Miguel Valdez, Matthew Cook, and Stephen Potter. "Humans and Robots Coping with Crisis – Starship, Covid-19 and Urban Robotics in an Unpredictable World". In: *Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2021, pp. 2596–2601. DOI: 10.1109/SMC52423.2021.9658581. URL: https://ieeexplore.ieee.org/document/9658581.

[53] Starship Technologies. *National Study Shows Delivery Robots Help Students Feel Safer, Skip Fewer Meals, and Improve Their Mental Health*. Accessed: 2025-04-16. 2023. URL: https://www.starship.xyz/press/national-study-of-more-than-7000-college-students-shows-delivery-

robots-help-students-feel-safer-skip-fewer-meals-and-improve-their-mental-health/.

[54] Miguel Valdez and Matthew Cook. "Examining the Spatialities of Artificial Intelligence and Robotics in Transitions to More Sustainable Urban Mobilities". In: *Norsk Geografisk Tidsskrift - Norwegian Journal of Geography* 78.5 (2024), pp. 313–323. DOI: 10.1080/00291951.2024.2432308.

[55] Pei-Sung Lin et al. *Campus Automated Shuttle Service Deployment Initiative*. Tech. rep. CUTR-NCTR-RR-2018-06. Accessed: 2025-04-16. Center for Urban Transportation Research, University of South Florida, 2020. URL: https://doi.org/10.5038/CUTR-NCTR-RR-2018-06.

[56] Luis C. Básaca-Preciado et al. "Intelligent Transportation Scheme for Autonomous Vehicle in Smart Campus". In: *Proceedings of the 44th Annual Conference of the IEEE Industrial Electronics Society (IECON 2018)*. IEEE, 2018, pp. 3193–3199. DOI: 10.1109/IECON.2018.8592824. URL: https://ieeexplore.ieee.org/document/8592824.

[57] . *MIREA Laboratory Industry 4.0. Digital robotic manufacturing10*. Licensed under CC BY-SA 4.0. 2021. URL: https://commons.wikimedia.org/wiki/File:MIREA_Laboratory_Industry_4.0._Digital_robotic_manufacturing10.jpg.

[58] Renno Juhkam and Margus Ross. "ABB YuMi© High-Speed Pick and Place Game in Action". In: *Proceedings of the 29th DAAAM International Symposium on Intelligent Manufacturing and Automation*. Vienna, Austria: DAAAM International, 2018, pp. 1211–1215. DOI: 10.2507/29th.daaam.proceedings.176. URL: https://www.daaam.info/Downloads/Pdfs/proceedings/proceedings_2018/176.pdf.

[59] Martin Tamm. "Research on Industrial Manipulator Trajectory Optimization by Applying Nonlinear Model Predictive Control". Master's thesis. Tallinn, Estonia: Tallinn University of Technology, 2021. URL: https://digikogu.taltech.ee/en/Download/2c36be77-b2da-4410-9d84-84844b90689e.

[60] ABB Robotics. *ABB Demonstrates Concept of Mobile Laboratory Robot for Hospital of the Future*. Accessed: 2025-04-16. 2022. URL: https://new.abb.com/news/detail/37301/abb-demonstrates-concept-of-mobile-laboratory-robot-for-hospital-of-the-future.

[61] F. Gabriele Pratticò and Fabrizio Lamberti. "Towards the adoption of virtual reality training systems for the self-tuition of industrial robot operators: A case study at KUKA". In: *Computers in Industry* 129 (2021), p. 103446. DOI: 10.1016/j.compind.2021.103446.

[62] Rainer Bischoff, Ulrich Huggenberger, and Erwin Prassler. “KUKA youBot – a Mobile Manipulator for Research and Education”. In: *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 1–6. DOI: 10.1109/ICRA.2011.5980575. URL: https://doi.org/10.1109/ICRA.2011.5980575.

[63] Pin-Chu Yang et al. “Repeatable Folding Task by Humanoid Robot Worker Using Deep Learning”. In: *IEEE Robotics and Automation Letters* 2.2 (2017), pp. 397–403. DOI: 10.1109/LRA.2016.2633383. URL: https://doi.org/10.1109/LRA.2016.2633383.

[64] Jun Takamatsu et al. *Learning-from-Observation System Considering Hardware-Level Reusability*. Preprint. 2022. arXiv: 2212.09242 [cs.RO]. URL: https://arxiv.org/abs/2212.09242.

[65] Kaiser Permanente Bernard J. Tyson School of Medicine. *The cutting edge: Training students in robotic surgery*. Accessed April 2025. 2023. URL: https://medschool.kp.org/news/the-cutting-edge-training-students-in-robotic-surgery.

[66] Shawn Tsuda et al. “SAGES TAVAC safety and effectiveness analysis: da Vinci® Surgical System (Intuitive Surgical, Sunnyvale, CA)”. In: *Surgical Endoscopy* (2015). DOI: 10.1007/s00464-015-4428-y. URL: https://doi.org/10.1007/s00464-015-4428-y.

[67] Sven Cremer, Lawrence Mastromoro, and Dan O. Popa. “On the Performance of the Baxter Research Robot”. In: *2016 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. IEEE. 2016, pp. 106–111. DOI: 10.1109/ISAM.2016.7750722.

[68] Ana Cunha et al. “Towards Endowing Collaborative Robots with Fast Learning for Minimizing Tutors’ Demonstrations: What and When to Do?” In: *ROBOT 2019: Fourth Iberian Robotics Conference*. Springer, 2020, pp. 368–378. ISBN: 978-3-030-35989-8. DOI: 10.1007/978-3-030-35990-4_30.

[69] David Pérez (DPC). *SpotMini 02 by-dpc*. Licensed under CC BY-SA 4.0. Attribution required. 2018. URL: https://commons.wikimedia.org/wiki/File:SpotMini_02_by-dpc.jpg.

[70] Boston Dynamics. *Case Study: Brown University Explores the Future with Spot*. Accessed 2025-04-16. 2023. URL: https://bostondynamics.com/case-studies/brown-university/.

[71] Jeffrey Hyde et al. *Spot Robot Staffing Augmentation at Los Alamos National Laboratory*. Tech. rep. LA-UR-23-24724. Presented at the 2023 INMM/ESARDA Joint Annual Meeting. Approved for public release; distribution is unlimited. Los Alamos National Laboratory, 2023. URL: https://resources.inmm.org/sites/default/files/2023-07/finalpaper_119_0503101544.pdf.

[72] Princeton Engineering. *Meet Spot the bot: Course explores learning with live robots*. Accessed 2025-04-16. 2024. URL: https://engineering.princeton.edu/news/2024/05/16/meet-spot-bot-course-explores-learning-live-robots.

[73] Priyaranjan Biswal and Prases K. Mohanty. "Development of Quadruped Walking Robots: A Review". In: *Ain Shams Engineering Journal* 12.2 (2021), pp. 2017–2031. DOI: 10.1016/j.asej.2020.11.005. URL: https://www.sciencedirect.com/science/article/pii/S2090447920302501.

[74] Milad Shafiee, Guillaume Bellegarda, and Auke Ijspeert. "ManyQuadrupeds: Learning a Single Locomotion Policy for Diverse Quadruped Robots". In: *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3471–3477. DOI: 10.1109/ICRA57147.2024.10610155. URL: https://miladshafiee.github.io/ManyQuadrupeds/.

[75] Maria Gini, Jan Pearce, and Karen Sutherland. "Using the Sony AIBOs to Increase Diversity in Undergraduate CS Programs". In: *Proceedings of the 9th International Conference on Intelligent Autonomous Systems (IAS-9)*. IOS Press, 2006, pp. 1033–1040. URL: https://www-users.cse.umn.edu/~gini/publications/papers/Gini06ias.pdf.

[76] Zhengyue Zhou et al. "Progresses of Animal Robots: A Historical Review and Perspectiveness". In: *Heliyon* 8.11 (2022), e11499. DOI: 10.1016/j.heliyon.2022.e11499. URL: https://doi.org/10.1016/j.heliyon.2022.e11499.

[77] Margaret M. Krupp et al. "A Focus Group Study of Privacy Concerns about Telepresence Robots". In: *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 1451–1456. DOI: 10.1109/ROMAN.2017.8172495. URL: https://doi.org/10.1109/ROMAN.2017.8172495.

[78] Samson Ogheneovo Oruma et al. "Security Aspects of Social Robots in Public Spaces: A Systematic Mapping Study". In: *Sensors* 23.19 (2023), p. 8056. DOI: 10.3390/s23198056. URL: https://doi.org/10.3390/s23198056.

[79] European Parliament and the Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. OJ L, 2024/1689, 12.7.2024. CELEX: 32024R1689. June 13, 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689 (visited on 12/21/2025).

[80] Dimitrios S. Stamoulis. "Management Considerations for Robotic Process Automation Implementations in Digital Industries". In: *Journal of Information System and Technology Management* 7.25 (2022), pp. 35–53. DOI: 10.35631/JISTM.725003. URL: https://doi.org/10.35631/JISTM.725003.

[81] Charlie Osborne. *Black Hat: Hackers can remotely hijack enterprise, healthcare Temi robots*. https://www.zdnet.com/article/black-hat-healthcare-senior-living-temi-robots-can-be-hijacked-remotely-by-hackers/. Accessed: 2025-04-16. 2020.

[82] Alberto Giaretta, Michele De Donno, and Nicola Dragoni. "Adding Salt to Pepper: A Structured Security Assessment over a Humanoid Robot". In: *Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018)*. Association for Computing Machinery, 2018, pp. 1–8. DOI: 10.1145/3230833.3232807. URL: https://doi.org/10.1145/3230833.3232807.

[83] Zack Whittaker. *Autonomous robots used in hundreds of hospitals at risk of remote hijacks*. https://techcrunch.com/2022/04/12/aethon-robots-hospitals-hijacks/. Accessed: 2025-04-16. 2022.

[84] Nicolas Nino et al. "Unveiling IoT Security in Reality: A Firmware-Centric Journey". In: *Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 5609–5626. ISBN: 978-1-939133-44-1. URL: https://www.usenix.org/conference/usenixsecurity24/presentation/nino.

[85] Alias Robotics. *DDS and ROS 2 Vulnerabilities Affect Hundreds of Robots*. Accessed: 2025-04-16. 2022. URL: https://news.aliasrobotics.com/alias-robotics-dds-ros2-vulnerabilities/.

[86] Christoph Lutz, Maren Schöttler, and Christian Pieter Hoffmann. "The privacy implications of social robots: Scoping review and expert interviews". In: *Mobile Media & Communication* 7.3 (Sept. 2019), pp. 412–434. DOI:

10.1177/2050157919843961. URL: https://doi.org/10.1177/2050157919843961.

[87] Samantha Reig et al. “Social Robots in Service Contexts: Exploring the Rewards and Risks of Personalization and Re-embodiment”. In: *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS ’21)*. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 1390–1402. ISBN: 978-1-4503-8476-6. DOI: 10.1145/3461778.3462036. URL: https://doi.org/10.1145/3461778.3462036.

[88] Bengisu Cagiltay et al. “Investigating Family Perceptions and Design Preferences for an In-Home Robot”. In: *Proceedings of the Interaction Design and Children Conference (IDC ’20)*. London, United Kingdom: Association for Computing Machinery, 2020, pp. 229–242. ISBN: 978-1-4503-7981-6. DOI: 10.1145/3392063.3394411. URL: https://doi.org/10.1145/3392063.3394411.

[89] Brian Tang et al. “CONFIDANT: A Privacy Controller for Social Robots”. In: *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI ’22)*. New York, NY, USA: IEEE, 2022, pp. 205–214. ISBN: 978-1-6654-0731-1. DOI: 10.1109/HRI53351.2022.9889540. URL: https://doi.org/10.1109/HRI53351.2022.9889540.

[90] Matthew Rueben et al. “A Taxonomy of Privacy Constructs for Privacy-Sensitive Robotics”. In: *arXiv preprint* (Jan. 2017). DOI: 10.48550/arXiv.1701.00841. arXiv: 1701.00841 [cs.RO]. URL: https://arxiv.org/abs/1701.00841.

[91] Sam Quinn and Mark Bereza. *Call an Exorcist! My Robot’s Possessed!* McAfee Advanced Threat Research Blog. Apr. 2020. URL: https://www.mcafee.com/blogs/other-blogs/mcafee-labs/call-an-exorcist-my-robots-possessed/.

[92] Alias Robotics. *Softbank Robotics Security Case Study: Penetration Testing Humanoid Social Robots*. Accessed: 2025-04-20. 2020. URL: https://aliasrobotics.com/case-study-pentesting-softbank.php.

[93] Jims Marchang and Alessandro Di Nuovo. “Assistive Multimodal Robotic System (AMRSys): Security and Privacy Issues, Challenges, and Possible Solutions”. In: *Applied Sciences* 12.4 (2022), p. 2174. DOI: 10.3390/app12042174. URL: https://www.mdpi.com/2076-3417/12/4/2174.

[94] Andrea F. Abate et al. “Contextual Trust Model With a Humanoid Robot Defense for Attacks to Smart Eco-Systems”. In: *IEEE Access* 8 (2020), pp. 207404–207414. DOI: 10.1109/ACCESS.2020.3037701. URL: https://doi.org/10.1109/ACCESS.2020.3037701.

[95] Yin Zhang et al. “Emotion-Aware Multimedia Systems Security”. In: *IEEE Transactions on Multimedia* 21.3 (2019), pp. 617–624. DOI: 10.1109/TMM.2018.2882744. URL: https://doi.org/10.1109/TMM.2018.2882744.

[96] Eduard Fosch-Villaronga et al. “Cloud Services for Robotic Nurses? Assessing Legal and Ethical Issues in the Use of Cloud Services for Healthcare Robots”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018, pp. 1–6. DOI: 10.1109/IROS.2018.8593591. URL: https://doi.org/10.1109/IROS.2018.8593591.

[97] Bernhard Dieber et al. “Application-level Security for ROS-based Applications”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2016, pp. 4477–4482. DOI: 10.1109/IROS.2016.7759659. URL: https://doi.org/10.1109/IROS.2016.7759659.

[98] Kaitlyn Cottrell et al. “An Empirical Study of Vulnerabilities in Robotics”. In: *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. 2021, pp. 735–744. DOI: 10.1109/COMPSAC51774.2021.00105. URL: https://doi.org/10.1109/COMPSAC51774.2021.00105.

[99] Víctor Mayoral Vilches et al. “Introducing the Robot Vulnerability Database (RVD)”. In: *arXiv preprint arXiv:1912.11299* (2019). DOI: 10.48550/arXiv.1912.11299. URL: https://arxiv.org/abs/1912.11299.

[100] Bernhard Dieber et al. “Security for the Robot Operating System”. In: *Robotics and Autonomous Systems* 98 (2017), pp. 192–203. DOI: 10.1016/j.robot.2017.09.017. URL: https://doi.org/10.1016/j.robot.2017.09.017.

[101] Benjamin Breiling, Bernhard Dieber, and Peter Schartner. “Secure Communication for the Robot Operating System”. In: *2017 Annual IEEE International Systems Conference (SysCon)*. 2017, pp. 1–6. DOI: 10.1109/SYSCON.2017.7934755. URL: https://doi.org/10.1109/SYSCON.2017.7934755.

[102] Pericle Salvini, Diego Paez-Granados, and Aude Billard. “Safety Concerns Emerging from Robots Navigating in Crowded Pedestrian Areas”. In: *International Journal of Social Robotics* 14.2 (2022), pp. 441–462. DOI: 10.1007/s12369-021-00796-4. URL: https://doi.org/10.1007/s12369-021-00796-4.

[103] Xiang Li and Yaping Lu. “Research on Laboratory Safety Management and Teaching in Applied Universities – An Example of Industrial Robot Laboratory”. In: *Proceedings of the 2022 International Conference on Industrial Control, Artificial Intelligence and Education (IC-ICAIE 2022)*. Advances in Human-Centered and Cognitive Computing. 2023, pp. 1549–1554. DOI: 10.2991/978-94-6463-040-4_233. URL: https://www.atlantis-press.com/proceedings/ic-icaie-22/125981170.

[104] Neziha Akalin, Annica Kristoffersson, and Amy Loutfi. “Evaluating the Sense of Safety and Security in Human–Robot Interaction with Older People”. In: *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction*. Human–Computer Interaction Series. Springer, 2019, pp. 237–264. DOI: 10.1007/978-3-030-17107-0_12. URL: https://doi.org/10.1007/978-3-030-17107-0_12.

[105] Nicholas DeMarinis et al. “Scanning the Internet for ROS: A View of Security in Robotics Research”. In: *2019 IEEE International Conference on Robotics and Automation (ICRA)*. 2019, pp. 8514–8521. DOI: 10.1109/ICRA.2019.8794451. URL: https://arxiv.org/abs/1808.03322.

[106] Karolina Krzykowska-Piotrowska et al. “Is Secure Communication in the R2I (Robot-to-Infrastructure) Model Possible? Identification of Threats”. In: *Energies* 14.15 (2021), p. 4702. DOI: 10.3390/en14154702. URL: https://www.mdpi.com/1996-1073/14/15/4702.

[107] George W. Clark, Michael V. Doran, and Todd R. Andel. “Cybersecurity issues in robotics”. In: *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. 2017, pp. 1–5. DOI: 10.1109/COGSIMA.2017.7929597. URL: https://doi.org/10.1109/COGSIMA.2017.7929597.

[108] Eduard Fosch-Villaronga and Christopher Millard. “Cloud Robotics Law and Regulation: Challenges in the Governance of Complex and Dynamic Cyber–Physical Ecosystems”. In: *Robotics and Autonomous Systems* 119 (2019), pp. 77–91. DOI: 10.1016/j.robot.2019.06.003. URL: https://www.sciencedirect.com/science/article/pii/S092188901930051X.

[109] Eduard Fosch-Villaronga. *Robots, Healthcare, and the Law: Regulating Automation in Personal Care*. Routledge, 2019. ISBN: 9780429021930. DOI: 10.4324/9780429021930. URL: https://www.taylorfrancis.com/books/mono/10.4324/9780429021930/robots-healthcare-law-eduard-fosch-villaronga.

[110] Chuhao Wu, He Zhang, and John M. Carroll. “AI Governance in Higher Education: Case Studies of Guidance at Big Ten Universities”. In: *Future Internet* 16.10 (2024), p. 354. DOI: 10.3390/fi16100354. URL: https://www.mdpi.com/1999-5903/16/10/354.

[111] Alice Guerra, Francesco Parisi, and Daniel Pi. “Liability for Robots I: Legal Challenges”. In: *Journal of Institutional Economics* 18.3 (2022), pp. 331–343. DOI: 10.1017/S1744137421000825. URL: https://www.cambridge.org/core/journals/journal-of-institutional-economics/article/liability-for-robots-i-legal-challenges/089EA1B996A5E8974643F8F1BDCD86BB.

[112] Lucas Cardiell. ““A Robot Is Watching You”: Humanoid Robots and the Different Impacts on Privacy”. In: *Masaryk University Journal of Law and Technology* 15.2 (2021), pp. 247–278. DOI: 10.5817/MUJLT2021-2-5. URL: https://cadmus.eui.eu/handle/1814/76820.

[113] Kaori Ishii. “Comparative Legal Study on Privacy and Personal Data Protection for Robots Equipped with Artificial Intelligence: Looking at Functional and Technological Aspects”. In: *AI & Society* 34 (2019), pp. 509–533. DOI: 10.1007/s00146-017-0758-8. URL: https://link.springer.com/article/10.1007/s00146-017-0758-8.

[114] Paul Neumann et al. ““"I Don’t Want Parents to Watch My Lessons" – Privacy Trade-offs in the Use of Telepresence Robots in Schools for Children with Long-term Illnesses”. In: *Proceedings of Mensch und Computer 2024*. 2024, pp. 448–454. DOI: 10.1145/3670653.3677509. URL: https://dl.acm.org/doi/10.1145/3670653.3677509.

[115] Tanja Heuer, Ina Schiering, and Reinhard Gerndt. “Privacy-Centered Design for Social Robots”. In: *Interaction Studies* 20.3 (2019), pp. 509–529. DOI: 10.1075/is.18063.heu. URL: https://benjamins.com/catalog/is.18063.heu.

[116] Bobbie Eicher, Lalith Polepeddi, and Ashok Goel. “Jill Watson Doesn’t Care if You’re Pregnant: Grounding AI Ethics in Empirical Studies”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* 2018 (2018), pp. 88–94. DOI: 10.1145/3278721.3278760. URL: https://doi.org/10.1145/3278721.3278760.

[117] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. “Transparent, Explainable, and Accountable AI for Robotics”. In: *Science Robotics* 2 (May 2017). DOI: 10.1126/scirobotics.aan6080.

[118] John-Stewart Gordon. “Building Moral Robots: Ethical Pitfalls and Challenges”. In: *Science and Engineering Ethics* 26 (Feb. 2020). DOI: 10.1007/s11948-019-00084-5.

[119] Sheshadri Chatterjee. “Impact of AI Regulation on Intention to Use Robots: From Citizens and Government Perspective”. In: *International Journal of Intelligent Unmanned Systems* ahead-of-print (Dec. 2019). DOI: 10.1108/IJIUS-09-2019-0051.

[120] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. “Meaningful human control and variable autonomy in human-robot teams for firefighting”. In: *Frontiers in Robotics and AI* 11 (2024). DOI: 10.3389/frobt.2024.1323980. URL: https://www.frontiersin.org/articles/10.3389/frobt.2024.1323980.

[121] Woodrow Barfield. “Liability for Autonomous and Artificially Intelligent Robots”. In: *Paladyn, Journal of Behavioral Robotics* 9.1 (2018), pp. 193–203. DOI: 10.1515/pjbr-2018-0018. URL: https://doi.org/10.1515/pjbr-2018-0018.

[122] Zsófia Tóth et al. “The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability”. In: *Journal of Business Ethics* 178 (2022), pp. 895–916. DOI: 10.1007/s10551-022-05050-z.

[123] Janina Loh. “Responsibility and Robot Ethics: A Critical Overview”. In: *Philosophies* 4.4 (2019), p. 58. DOI: 10.3390/philosophies4040058. URL: https://www.mdpi.com/2409-9287/4/4/58.

[124] Shane P. Saunderson and Goldie Nejat. “Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction”. In: *Science Robotics* 6.58 (2021), eabd5186. DOI: 10.1126/scirobotics.abd5186. URL: https://www.science.org/doi/abs/10.1126/scirobotics.abd5186.

[125] Laura Londono et al. “Fairness and Bias in Robot Learning”. In: *Proceedings of the IEEE* PP (2024), pp. 1–26. DOI: 10.1109/JPROC.2024.3403898.

[126] Laura Cesaro et al. “Gender biases in robots for education”. In: *Proceedings of the 3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2024)*. University of Padua, 2024. URL: https://hdl.handle.net/11577/3542939.

[127] Jie Zhu et al. “Fairness-Sensitive Policy-Gradient Reinforcement Learning for Reducing Bias in Robotic Assistance”. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2024, pp. 549–554. DOI: 10.1109/RO-MAN60168.2024.10731257.

[128] Stevienna de Saille et al. “Improving Inclusivity in Robotics Design: An Exploration of Methods for Upstream Co-Creation”. In: *Frontiers in Robotics and AI* 9 (2022), p. 731006. DOI: 10.3389/frobt.2022.731006.

[129] Anastasia K. Ostrowski et al. “Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions”. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2022, pp. 969–976. DOI: 10.1109/RO-MAN53752.2022.9900805.

[130] Adam Poulsen, Eduard Fosch-Villaronga, and Oliver Burmeister. “Cybersecurity, value sensing robots for LGBTIQ+ elderly, and the need for revised codes of conduct”. In: *Australasian Journal of Information Systems* 24 (2020). DOI: 10.3127/ajis.v24i0.2789.

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Rashad Gafarli

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Robot Assistants in Higher Education: A Study of Ethical and Cybersecurity Challenges", supervised by Fuad Budagov
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

04.01.2026

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Case Study 1: ACA Control Evaluation

This appendix contains the full Control Evaluation Matrix and related tables for the Attendance Check Application (ACA) robot assistant case study. The main text in Chapter 6 provides a narrative synthesis; here, all detailed per-control scores and comments are reported.

Table 19. Case Study (D.P.) Evaluation

Control (D.P.)	Requirement	Assessment	Observations
D.P.-1	End-to-End Encryption	3 - Fully Implemented	The system employs some encryption by virtue of underlying platforms, but not comprehensively. **T**he updated ACA system encrypts all sensitive data during both storage and transmission. Attendance results are encrypted before transmission to the institutional database, and ID photos are retrieved securely via the national ID card system. Live facial scans are processed locally, and neither image is stored beyond the immediate comparison, reducing data-at-rest exposure. These implementations reflect adherence to strong encryption standards at all stages, qualifying as fully implemented.
D.P.-2	Minimized Biometric Retention	3 - Fully Implemented	The ID-card integration significantly improves this aspect. Now, the reference photo comes from the ID card each time – the system can perform a live 1:1 match and does not need to store the student's face itself at all. Once the comparison is done and attendance marked, the biometric data (live scan and ID photo) can be discarded immediately. The robot can simply record "Student X present at time Y" with no face images retained. Assuming the system is designed to not keep copies of the ID photo or live image, it fulfills this control by minimizing biometric data retention. (It would be wise for the system designers to ensure any cached images are purged from memory/storage right after verification to fully meet this requirement.)
D.P.-3	Secure Transmission Enforcement	2 - Partially Fulfilled	In the pilot deployment, because data was kept local on the robot, transmission was minimal. However, the robot operates within any accessible network, including potentially open or less secure environments. While encryption (e.g., TLS) is assumed for ID card access and attendance logging, the system does not explicitly enforce or restrict insecure communication channels. There are no apparent safeguards against fallback to insecure protocols or unauthorized data extraction via Wi-Fi or peripheral connections. For example, if the robot were connected to an open network accidentally, do we know it would refuse to send sensitive data? Probably not explicitly handled. Thus, while the system probably uses secure protocols by default, it doesn't demonstrably enforce or audit against insecure comms, making this partially fulfilled.
D.P.-4	Purpose-Limited Data Use and Retention	3 - Fully Implemented	The updated deployment incorporates strict purpose-limitation principles. Data is only collected for attendance verification and not reused for unrelated functions such as behavioral profiling. Retention policies are clearly scoped to course duration, and records are deleted at term completion. Consent mechanisms ensure students are aware of these boundaries, and data reuse requires re-consent. These practices reflect strong alignment with GDPR and EU AI Act expectations for purpose-limited and justified data handling.

Table 20. Case Study (S.I.) Evaluation

Control (S.I.)	Requirement	Assessment	Observations
S.I.-1	Immutable Audit Logs	3 - Fully Implemented	The updated ACA deployment includes reliable audit logging. Attendance events are timestamped and linked to identity verification, and logs are secured through the institutional backend system. Although the robot may not implement cryptographic hash-chaining locally, the backend recordkeeping ensures traceability and integrity for academic workflows. These logs cannot be casually edited without triggering administrative review processes, qualifying the control as fully implemented in practice.
S.I.-2	Secure Update Pipeline	1 - Minimally Addresses	Updates to the Temi platform and ACA application are applied manually or automatically depending on the component, but lack strict signature verification or authenticated logging. While the manufacturer provides OS-level updates, the integration code for ID card matching and facial comparison appears maintained independently without a structured update pipeline. There's no known enforcement of cryptographic signature checks prior to application, and update logs are not systematically retained. This control is minimally addressed, primarily through vendor defaults rather than proactive institutional controls.
S.I.-3	Runtime Integrity Checks	0 - Not Addressed	The system has no known runtime integrity verification. For example, there are no watchdogs monitoring if the face recognition module was altered or if unauthorized processes are running on the robot. **Pilot:** This was a prototype application – no such defenses were built in. **New:** It's unlikely the new version includes integrity monitoring. The focus was on functionality (ID integration), not on adding intrusion detection on the robot. This is a gap; if malware or unauthorized changes occurred, the robot would not notice or report it.
S.I.-4	Isolation of Trusted Components	0 - Not Addressed	The face recognition and ID reading processes run as part of the application or OS with no special isolation. **Pilot:** The ACA app ran on the Temi's tablet environment (likely Android-based), which wasn't hardened. **New:** The ID card reading might involve a middleware or driver, but there's no indication of using secure enclaves or isolation. For instance, ideally the ID photo comparison would happen in a trusted execution environment to prevent spoofing/tampering, but the system likely just uses normal application space. Thus, a compromise of the robot could affect even the identity-matching module. This control is not implemented in the current design.

Table 21. Case Study (I.L.) Evaluation

Control (I.L.)	Requirement	Assessment	Observations
I.L.-1	Safe Defaults for Audio/Video	1 - Minimally Addresses	The robot's camera is used only for the specific purpose of face recognition during check-in. It is not continuously recording or streaming the classroom by default (no evidence of that). **New:** The ID verification process uses the camera only when a card is presented, so it's event-triggered, not constant surveillance. However, there weren't explicit user-controlled indicators. The robot appears to continuously listen for trigger phrases like a virtual assistant (e.g., similar to Siri-style passive listening), even outside attendance check-ins. It does not visibly indicate when it is actively recording or processing audio. There is no LED or on-screen indicator during microphone activation, and it's unclear whether processing is done locally or offloaded to a cloud server—especially when the robot is connected to the internet. Furthermore, the robot fails to respond if disconnected, suggesting it relies on remote systems for certain features. Without explicit confirmation of local-only processing and without user-facing controls, this control is minimally addressed and requires improvement.
I.L.-2	Session Expiry & Timeout Controls	2 - Partially Fulfills	The attendance check interactions are very short (about 10 seconds for returning users). After a student checks in, the robot resets for the next user. There is no persistent login state that stays open. This effectively means each "session" (each student's check-in) is isolated. **Pilot:** If a student walked away, the robot wouldn't keep any logged-in session active; it was ready for the next person. **New:** Similarly, once an ID card is removed and the check is done, that session ends. However, on a broader level, the robot itself might remain logged in to the ACA app or connected to networks. There is no mention of the robot locking its admin interface or timing out remote access sessions. So, for user-facing sessions it's fine (short by design), but for admin access there might not be robust timeouts. We mark this as partial – user interactions expire immediately, but the system itself may not enforce timeouts for maintenance interfaces.
I.L.-3	Output Filtering & Response Censorship	3 - Fulfills	The robot's output in this application is minimal and well-scoped. It typically greets the student and confirms attendance. It does not pull up or announce any sensitive personal information beyond perhaps the student's name. **Pilot:** There was no functionality where a student could query someone else's data. The robot would only address the student currently checking in. **New:** When using the ID card, the robot might read the student's name from the card to confirm (e.g., "Hello, [Name], you are marked present"). This is done one-on-one with the student at the robot and isn't broadcast to others (aside from nearby classmates overhearing a name). It does not reveal anything like personal ID numbers or other students' info. Because the system's dialogue and display are intentionally limited to the task at hand, it naturally avoids information leakage through its outputs. There is no chatty AI component that might spill data; it's a deterministic app, thus fulfilling this control.
I.L.-4	Retention & Disclosure Policies	2 - Partially Fulfills	Purpose-limited data use and minimization are largely practiced in implementation—biometric data is not stored persistently, and attendance records are presumably kept only for the academic term. However, there is no outward-facing documentation or policy provided to students explaining these data practices. Without a published retention policy or clear disclosure of what data is stored, for how long, and who has access, transparency remains lacking. Internally, the approach appears compliant with best practices, but externally the control is only partially fulfilled due to the absence of formal documentation and student communication.

Table 22. Case Study (S.A.) Evaluation

Control (S.A.)	Requirement	Assessment	Observations
S.A.-1	Role-Based Access Control (RBAC)	2 - Partially Fulfills	The system's primary user-facing function (attendance) doesn't involve user accounts (students aren't logging into the robot; they use their ID card for auth). However, the robot itself and the ACA app likely have admin credentials. Only authorized personnel have access to the system's internal database, and they can view, download, and edit data from the robot. There is no formal role-based access control system enforcing separation of duties or auditing access scopes. Access is likely based on shared assumptions of physical security and staff assignment rather than a structured, verifiable RBAC framework. Thus, the control is partially implemented through operational practice, not through robust technical enforcement.
S.A.-2	Credential and Interface Security Hardening	0 - Not Addressed	There's no evidence that the robot's admin interface (if one exists for the ACA or Temi) is protected with multi-factor authentication. **Pilot:** The researchers had direct physical access to the robot, so they didn't need remote admin. There's no evidence that the robot's admin interface is protected with strong credential policies. Admin accounts are created or removed as needed but lack centralized management, MFA, or enforcement of password complexity. The pilot considered this an acceptable risk due to its limited exposure and scope, but for production use, the lack of user lifecycle management is a clear gap. Credential and interface hardening were not performed.
S.A.-3	Access Attempt Logging & Anomaly Monitoring	2 - Partially Fulfills	**Pilot:** Not applicable in the user-facing sense since students don't log in, and any admin access was direct by researchers. **New:** The critical user authentication happens via the ID card – if a student fails the face match, the system likely notes an unsuccessful attempt to match, but that is not a "login" per se, and those events might not be centrally logged as security events.Access attempts—especially successful ones—are recorded, with details such as which user accessed the system and how many times. However, logging lacks granularity and security depth. Failed login attempts are not recorded, and the system doesn't log contextual details such as IP address, timestamp, or geolocation. The logs are useful for basic oversight but are insufficient for robust auditing or incident response.
S.A.-4	Privilege Escalation Controls and Re-Authentication	0 - Not Addressed	The attendance system is quite limited in scope, so it doesn't have the concept of an admin performing "sensitive actions" through the UI – the sensitive action would be editing the attendance record, which would typically be done in the university's system, not on the robot. If we consider maintenance tasks on the robot (like deleting all data or updating firmware), those are outside the ACA app's normal operations. The system lacks interfaces that would support re-authentication flows. Sensitive operations (e.g., wiping data, updating software) are performed directly at the OS or device level without prompting the user for credentials again. This control is not addressed in either the pilot or updated version.

Table 23. Case Study (V.P.) Evaluation

Control (V.P.)	Requirement	Assessment	Observations
V.P.-1	Scheduled Patch Application	3 - Fully Implemented	ACA was custom-developed for the research and actively maintained throughout the experiment by the lead developer. Updates were applied as needed across the pilot phases. The underlying Temi robot platform also receives automatic updates, ensuring the operating system and system components stay current. There is evidence that the ACA app was version-controlled and updated based on deployment needs. This control is fully met through both vendor-side automation and research team involvement.
V.P.-2	Periodic Vulnerability Scanning	0 - Not Addressed	There is no indication that the team or IT security performs vulnerability scans on the robot or the ACA software. **Pilot:** Not done – focus was on user study. **New:** Unless the robot is considered part of the IT asset inventory, it's likely not being regularly scanned by tools like Nessus. Many IoT/robot devices are introduced without being fully integrated into vulnerability management processes. Here, that seems to be the case – this control is not being observed.
V.P.-3	Secure Baseline Tracking	3 - Fully Implemented	The development team maintained version-controlled builds of the ACA system, and updates followed a tracked deployment schedule. Robot firmware versions, app installations, and configuration files were monitored during the research phases. This allows verification that unauthorized changes or version drift did not occur. As a result, the robot's configuration and software stack are traceable and managed against a known baseline.
V.P.-4	EOL and Unsupported Component Avoidance	3 - Fully Implemented	**Pilot:** The project used a fairly new robot (Temi) and a freshly developed app, so nothing was EOL at that time. **New:** The system uses Estonia's current ID card infrastructure, which is actively supported, and presumably the latest libraries to interface with it. All components used—Temi hardware, ACA software, and the Estonian ID card integration—are actively supported. There is clear documentation tracking compatibility and lifecycle support for the ID card infrastructure and application modules. No legacy dependencies or deprecated frameworks are in use. The developers maintain awareness of support timelines and apply updates to preempt obsolescence. As such, this control is fully met.

Table 24. Case Study (C.P.) Evaluation

Control (C.P.)	Requirement	Assessment	Observations
C.P.-1	Emergency Stop Access	3 - Fully Implemented	**Pilot:** The robot was stationed by the door and presumably not moving around the room during attendance. **New:** During ID scanning, Temi likely remains stationary.The latest generation of Temi robots (e.g., Temi Go) includes a dedicated emergency stop (E-stop) button for physical override. While the earlier pilot robot lacked this feature explicitly, updated deployments at TalTech are adopting models with clearly marked and accessible E-stop functionality. Additionally, software-based soft-stop commands are available through the supervisor app. Instructors are trained to intervene if needed. This meets safety expectations for physical override mechanisms.
C.P.-2	Motion Sandbox Testing	3 - Fully Implemented	The movement of the robot in this scenario is minimal (perhaps turning in place to face students). **Pilot:** The environment was a normal classroom; Temi's behavior was predictable and tested informally (they would have tried it in the classroom to ensure it can approach students or position properly). It wasn't a complex autonomous navigation scenario requiring extensive simulation. **New:** Because now the main function is reading ID and scanning face, the robot might not need to navigate at all (students come to it). Thus, the risk of erratic movement is low. In essence, by limiting movement in the use-case, they avoid many motion risks. The ACA application's robotic movement logic is conservative by design and tested prior to deployment. Movement routines such as pivoting or slight repositioning are validated in safe environments before being introduced to classrooms. The robot's limited mobility during attendance (usually fixed location) further mitigates risk. This pre-deployment validation aligns with sandbox testing principles, ensuring safety through constraint and verification.
C.P.-3	Restricted Control Interfaces	3 - Fully Implemented	**Pilot:** Students interacting with ACA were not given any control over the robot's movement beyond the app's guided interactions. However, if someone knew how to connect to Temi's control interface (for instance, via the Temi mobile app or a dev interface), could they drive it? Possibly, if not secured. The researchers didn't give out access, so by policy it was restricted. Movement control and actuator access are restricted to authorized personnel only. Remote commands and developer interfaces are disabled by default unless explicitly needed for maintenance. User interfaces for students or unprivileged users do not expose any motion functions. This segregation is enforced both via the application layer and at the device level using Temi's system settings. Unauthorized motion input is effectively blocked.
C.P.-4	Physical Activity Logging	3 - Fully Implemented	The Temi robot and ACA system log motion events and anomalies, such as failed navigation attempts or obstacle detection. These logs are accessible to system supervisors and used to monitor operational safety. If a collision or interruption occurs, associated telemetry (e.g., timestamp, sensor alert) is recorded for review. This enhances incident tracking and aligns with robotic safety logging best practices

Table 25. Case Study (N.C.) Evaluation

Control (N.C.)	Requirement	Assessment	Observations
N.C.-1	Encrypted Network Protocols	3 - Fully Implemented	The system appears to rely on inherently secure channels. **Pilot:** The data was mostly local; any internet connectivity (e.g., for the survey submission or remote monitoring) would use standard HTTPS. The university Wi-Fi and wired networks enforce encryption and authentication at the transport layer. **New:** The ID card integration likely requires the robot to communicate with the card reader over a USB or Bluetooth link; these are local connections. If any data goes to a server (like writing attendance to a database), it presumably uses HTTPS or a secure API channel. Given that Estonia's digital systems are mature, it's reasonable to assume proper use of secure protocols. We haven't identified any part of the workflow sending data in plaintext. This control is basically met by adhering to common secure communication practices.
N.C.-2	Network Segmentation & Firewalls	3 - Fully Implemented	**Pilot:** The Temi was likely on the university network along with other devices, not placed in a quarantined subnet. This wasn't a concern raised in the study. The robot operates on the university's managed network, which enforces segmentation and firewall controls. **New:** Devices like ACA-enabled Temi are enrolled under specific administrative units and have network profiles configured through IT. This ensures robots are logically separated from student and public traffic, with firewall policies defining which endpoints they can communicate with. This fulfills the segmentation control.
N.C.-3	Secure Interface Use	1 - Minimally Addresses	The integration of the ID card reader via USB means a new interface is used. Ideally, the robot should accept input only from that authorized reader device. A USB hub was added to extend connectivity, introducing potential new risks. While the ID card reader is the expected peripheral, there is no evidence of hardening (e.g., interface whitelisting, USB port lockdown). No protections against rogue USB device insertion were mentioned. The robot likely accepts connections from unknown peripherals by default, meaning control over physical ports is minimal. For example, if someone connected a laptop to the robot's port, could they extract data or exploit it? Possibly, if not locked down. Thus, external interfaces are a potential weak point not thoroughly secured (minimally addressed by default settings).
N.C.-4	Monitoring & Intrusion Detection	2 - Partially Fulfills	The university's broader IT infrastructure includes intrusion detection and anomaly monitoring for core systems. However, the robot itself is not monitored as a unique endpoint. While unusual behavior might be captured at the network level, no specialized monitoring tools (e.g., device-specific traffic alerts or IDS profiles) are tailored to the robot. Partial fulfillment is achieved through institutional baseline protections, but robot-specific attention is lacking.

Table 26. Case Study (I.R.) Evaluation

Control (I.R.)	Requirement	Assessment	Observations
I.R.-1	Mandatory IT Security Involvement	1 - Minimally Addresses	While the team handling the deployment includes technically skilled personnel, including those with IT backgrounds, there is no evidence of formal IT security reviews, threat modeling, or defined roles within the institution's security governance. Responsibility is implicitly carried by the technical implementers, but no structured security oversight or institutional approval process is in place. Security involvement appears incidental rather than mandated.
I.R.-2	Robot-Aware Policy Integration	3 - Fully Implemented	The university has taken steps to accommodate robotic deployments within its general IT governance framework. Acceptable use and privacy policies have been updated to account for AI systems and classroom automation tools. Where applicable, robot-specific responsibilities and privacy considerations are addressed in alignment with institutional and national frameworks. This control is fully satisfied given current capabilities.
I.R.-3	Incident Response Inclusion	2 - Partially Fulfills	**W**hile no robot-specific response plan exists, incidents such as data leaks or robot failure are now considered within the broader university risk management process. Pilot: If the robot failed, the researchers or instructor would simply intervene, but there was no defined escalation. Students can report issues (e.g., misrecorded attendance or perceived privacy harm), and such reports are handled by the IT support or ethics board. However, response pathways are not yet formally robot-specific, and drills or structured playbooks for robot-related incidents are still lacking

Table 27. Case Study (P.C.) Evaluation

Control (P.C.)	Requirement	Assessment	Observations
P.C.-1	Opt-in Consent for Data Collection	3 - Fully Implemented	Pilot: The process was explicitly opt-in, as students voluntarily registered their facial images after a transparent explanation by the instructor. **New:** Although the updated system relies on students presenting their ID card, this is still considered a voluntary action. Students are not forced to use the robot and may choose to consult the instructor directly, preserving the element of choice. Therefore, opt-in consent remains meaningfully implemented.
P.C.-2	Transparent Data Usage Notices	3 - Fully Implemented	Pilot: Students were informed about the purpose, scope, and research usage of the collected data. **New:** The use of the national ID card system came with a form of notification (students would be told that the robot will read their ID photo to verify attendance, alongside robot displaying disclaimer info). The interaction is clear in intent (attendance marking only), and no biometric data is retained. The limited scope and consistent use case help meet transparency expectations.
P.C.-3	Surveillance Indicators and Controls	1 - Minimally Addresses	The system lacks explicit surveillance cues like LED indicators or toggles. The robot obviously uses a camera to scan faces, but there is no specific indicator (like an LED or on-screen icon) showing "camera active" beyond the robot looking at the student. **Pilot:** Students engaged the robot directly, so it was evident when it was scanning them, but they did not have control to disable the camera – they simply could choose not to participate. **New:** The addition of the ID card means the camera activates when a card is inserted or presented. The student's action triggers scanning, implicitly consenting at that moment. While this gives some control (if you don't want to be scanned, you don't insert your ID), in practice if attendance is mandatory, that control is limited. There's no mention that students can toggle off any sensors; the system assumes compliance. Thus, explicit surveillance indicators or user controls are minimal.
P.C.-4	Users can view and delete personal data collected about them.	1 - Minimally Addresses	In both the pilot and new system, students do not have any interface to review what data the robot has stored about them. Students have no direct way to view or delete their data through the ACA system. **Pilot:** Research records were internal, and students couldn't access or manage their stored data. **New:** Attendance records become academic records. While students could raise requests through university DPOs or lecturers, no built-in feature exists to allow self-service data access or deletion. Some redress is possible via indirect channels, but not via the robot or ACA interface itself

Table 28. Case Study (T.E.) Evaluation

Control (T.E.)	Requirement	Assessment	Observations
T.E.-1	Disclosure of Non-Human Identity	3 - Fully Implemented	**Pilot:** The Temi robot's physical form and the instructor's introduction made it obvious that a machine – not a person – was checking attendance. The robot gave verbal greetings and process descriptions, reinforcing its role. **New:** Similarly, the robot is visibly a robot and presumably identifies itself when interacting. No deception is present – students know a robot (with ID card reader) is handling attendance.
T.E.-2	Explanation Request Interface	0 - Not Addressed	The system does not provide a user-facing "why" explanation for its decision. In the pilot, if a student was not recognized or marked absent, the robot did not offer an explanation beyond an error message. In the new system, if the face match fails, the robot likely just indicates failure to verify, without a detailed explanation of the reasoning (e.g. poor match or lighting). There is no interface or functionality where a user can request the reason for a failed match or verification outcome. There is no evidence of an on-demand explainability feature in the ACA.
T.E.-3	Documentation of Decision Logic	3 - Fully Implemented	The ACA application, developed for research at TalTech, includes documented decision logic regarding the attendance mechanism. While not necessarily open to public users, this documentation exists and was shared internally among the research and IT team. The biometric comparison process using the ID photo and live scan has been described in internal and collaborative documentation.
T.E.-4	Role Transparency in Educational Tasks	3 - Fully Implemented	Pilot: Instructor clearly communicated the robot's role and purpose as part of the voluntary research activity. **New:** The role of the robot as an attendance verifier is known to the students through usage context and instructor briefings. While routine use could benefit from formalization (e.g., course announcements or LMS notes), the implementation is strong enough to meet the transparency expectations, especially as no deceptive practices are present.

Table 29. Case Study (A.L.) Evaluation

Control (A.L.)	Requirement	Assessment	Observations
A.L.-1	Assigned Human Supervisor	3 - Fully Implemented	In practice, the lecturer or class instructor acts as the supervisor. **Pilot:** The lecturer was present during robot attendance, introduced the process, and could step in if needed. **New:** It's expected that an instructor or staff is responsible for the robot each class (to set it up and respond to issues). While not formally labeled a "supervisor," the human-in-charge role is fulfilled by the instructor, who monitors that attendance is recorded correctly and can address anomalies (e.g., a student not recognized).
A.L.-2	Defined Student Appeal Pathways	2 - Partially Fulfills	No formal appeal process exists (no online form or documented procedure specific to robot errors). However, there is an informal pathway: students can immediately tell the instructor if the robot misses them or makes an error. **Pilot:** Because it was voluntary and small-scale, a student could simply speak up if their attendance wasn't recorded. **New:** In a real deployment, if a student is marked absent due to a failed face match, they would likely approach the instructor or admin to correct the record. This addresses the need for appeal informally, but not through a predefined official channel. Thus, appeal is possible but not systematized (partial compliance).
A.L.-3	Source Attribution Logging	2 - Partially Fulfills	The ACA logs basic operational events (e.g., check-in timestamps, match status), but does not distinguish whether actions were autonomous or manually overridden. **Pilot:** Logging was rudimentary and focused on performance. **New:** There is some form of internal recordkeeping, but no robust audit trail mapping inputs to originators (human vs. robot). Logs are useful for general review but not comprehensive for accountability attribution.
A.L.-4	Institutional Responsibility Policies	1 - Minimally Addresses	There exists implied responsibility under IT or data governance frameworks, but there is no documentation outlining acceptable use, misuse consequences, or error redress procedures specific to robotic systems. The responsibility structure remains informal and fragmented. **Pilot:** It was a research trial, so standard research ethics applied, but no institutional policy beyond that. **New:** The deployment leverages national ID – an indication that some institutional decision was made – but there's no mention of a formal policy defining acceptable use, misuse consequences, or liability for errors. In case of a system failure or data breach, responsibilities would likely fall back on general IT policies or ad-hoc decisions, rather than a pre-defined charter for robot usage.

Table 30. Case Study (H.O.) Evaluation

Control (H.O.)	Requirement	Assessment	Observations
H.O.-1	Human Oversight for Critical Decisions	2 - Partially Fulfills	Marking attendance is a lower-stakes decision (compared to grading), but errors can affect a student's record. **Pilot:** The instructor was present and could verify the robot's output – for example, if a student was marked absent incorrectly, the instructor could manually note the correction. **New:** The system likely logs who was marked present based on ID verification. Instructors can review the attendance list after class and ensure it matches who they saw in class. There is no forced human approval for each attendance entry (that would defeat automation), but oversight exists in the form of instructor awareness and the ability to cross-check at any time. Thus, a human is not in the loop for every check-in, but oversight is inherently present, satisfying this control partially.
H.O.-2	Manual Override Capabilities	2 - Partially Fulfills	If the robot fails or misidentifies someone, the instructor can manually mark that student present through alternative means (either telling the system if possible, or more likely via the usual attendance reporting in the school's system). The robot application itself might not have a dedicated "override" button, but the instructor can take over the role of marking attendance if needed (effectively overriding the automation). Also, if the robot misbehaved (say, got stuck or repeatedly failed to scan), the instructor can stop using it and switch to roll-call. So override is possible, but not via a built-in feature – it relies on human intervention outside the system.
H.O.-3	Integrated Feedback and Error Reporting	1 - Minimally Addresses	There is no dedicated feedback interface in the ACA. If a student has an issue (e.g., "the robot didn't mark me present"), their channel is simply to speak to the instructor or IT. **Pilot:** Issues would be reported verbally to researchers or via the post-class survey (some students did note privacy concerns there). **New:** Likely still informal – a student might email the lecturer or IT if something consistently goes wrong. The system does log data, but it doesn't automatically package and send error reports. Any alerting to supervisors (instructor) is manual. Thus, while issues can be raised, the process isn't integrated into the system design (only minimally addressed by general class communication).

Table 31. Case Study (B.F.) Evaluation

Control (B.F.)	Requirement	Assessment	Observations
B.F.-1	Use-Case Specific Bias Evaluation	3 - Fully Implemented	The system has been evaluated across diverse student demographics during development and refinement. Ongoing research incorporates fairness-focused validation, with tests examining the system's reliability for users of different ethnicities, genders, and appearance profiles. The use of ID-based 1:1 matching also helps standardize recognition performance by relying on verified, state-issued photo IDs, reducing reliance on learned model assumptions. Bias evaluation is thus embedded into the deployment lifecycle.
B.F.-2	Inclusive Performance Audits	3 - Fully Implemented	Performance audits are actively conducted under TalTech's extended research supervision. These audits include student feedback segmentation and are designed to identify and mitigate any disparities in system accuracy or experience across subgroups. Regular reviews ensure detection of systemic biases and drive continuous improvement in ACA's biometric matching and user interface. This process is institutionalized and supported by the developer's research-driven methodology.
B.F.-3	Equity Monitoring Metrics	3 - Fully Implemented	The attendance system collects usage data that is anonymized and, when ethically permitted, analyzed for equity trends. Metrics such as error rates, failed matches, and feedback satisfaction are tracked across different demographic groups, especially during user trials. These analytics support fairness verification and adjustment of system thresholds or prompts to ensure consistent engagement across student populations.
B.F.-4	Inclusive Design Requirements	3 - Fully Implemented	The robot's interface in the pilot was in English (or the classroom's language) and presumably usable by all in that class. The ACA system and Temi platform have been updated with accessibility and inclusivity in mind. The deployment ensures usability across a broad student base, including international students and those with differing interaction preferences. The underlying app design prioritizes equitable access and was developed iteratively with student usability input, aligning with inclusive design principles.

Table 32. Case Study (E.A.) Evaluation

Control (E.A.)	Requirement	Assessment	Observations
E.A.-1	Inclusive Interface Design	3 - Fully Implemented	The ACA on the Temi robot largely relies on a touch screen and camera. The updated ACA system ensures usability for a wide range of student needs through simple, universal design. Interaction is streamlined via ID card presentation, requiring no advanced digital literacy or mobile devices. The robot interface accommodates students with limited mobility and has been tested for clarity and ease of use in diverse classroom contexts. While no formal assistive tech integration is mentioned (e.g., screen readers), the process itself was designed to be minimal and inclusive, thus effectively supporting participation without introducing barriers.
E.A.-2	Ensure remote or off-campus students have an equivalent way to check in	0 - Not Addressed	The attendance system is tied to physical presence – a robot in the classroom. **Pilot:** Only in-person students could use it; remote students (if any) were outside the system. **New:** The reliance on a physical ID card and face scan means there is no provision for virtual attendance check-ins. In pandemic or hybrid scenarios, this system offers no solution. It does not cater to satellite campuses or distance learning. Essentially, if you're not physically present with your ID, the system cannot mark you present. The control of providing a virtual equivalent or alternative is not met.
E.A.-3	Socioeconomic Inclusion Measures	3 - Fully Implemented	**Pilot:** Students did not need any personal gadget or paid service – the robot and its tablet were provided by the university. Even the feedback survey was via a QR code, which assumed a smartphone, but that was optional and could presumably be done on any device or with help. The attendance check itself took place on the robot, free for all. **New:** The national ID card is mandatory for all citizens and students in Estonia, so every student already has one at no extra cost. There is no dependency on a student-owned smartphone or app – just the ID card (and virtually every student has that, or the university would need to accommodate those who forget theirs). Therefore, the system does not disadvantage lower-income students; it's equally available to all enrolled students without additional expense.
E.A.-4	Institutional Monitoring for Disparities	0 - Not Addressed	The university does not appear to track disparities in robot usage. **New:** No mechanisms are in place to monitor adoption or non-use patterns by demographic or ability group. If students experience issues or avoid using the robot (due to disability, privacy concerns, or unfamiliarity), there is no formal audit or feedback loop to identify and respond to these gaps. Without data collection or review structures, institutions cannot detect or correct access inequalities in robot-assisted attendance.

Control Area Scoring Summary

The following table summarizes the control evaluation results across all cybersecurity and ethical domains. Each control domain contains multiple individual controls evaluated against the robot assistant-based attendance system. Scores reflect the average degree of fulfillment on a 0–3 scale, and the final column shows the percentage of controls in that domain which meet at least a partial threshold (score of 2 or higher).

Table 33. Detailed Control Scoring per Domain

Control Domain	**1**	**2**	**3**	**4**	**Avg. Score**	**% $\geq$ 2**
Data Privacy & Confidentiality (D.P.)	3	3	2	3	2.75	100%
System & Data Integrity (S.I.)	3	1	0	0	1	25%
Information Leakage & Misuse (I.L.)	1	2	3	2	2	75%
System Access & Control (S.A.)	2	0	2	0	1	50%
Vulnerability & Patch Management (V.P.)	3	0	3	3	2.25	75%
Cyber-Physical Safety (C.P.)	3	3	3	3	3	100%
Network & Communication Security (N.C.)	3	3	1	2	2.25	75%
Institutional Readiness & Governance (I.R.)	1	3	2	–	2	66.7%
Privacy, Consent (P.C.)	3	3	1	1	1	50%
Transparency & Explainability (T.E.)	3	0	3	3	3	75%
Accountability & Liability (A.L.)	3	2	2	1	1	75%
Autonomy & Human Oversight (H.O.)	3	2	1	–	2	66.7%
Bias, Fairness & Inclusion (B.F.)	3	3	3	3	3	100%
Equity & Accessibility (E.A.)	3	0	3	0	1.5	50%
Overall					**2.14**	**67.9%**

Appendix 3 – Case Study 2: Robot Assistant for Answering Student Questions

This appendix contains the full Control Evaluation Matrix and related tables for the robot teaching assistant used to answer student questions during computer networking lab sessions. The main text in Chapter 6 provides a narrative synthesis of this case study and high-level comparison with the other deployments; here, all detailed per-control scores and comments are reported.

The evaluated system consists of a Temi robot acting as a physical front-end to an Interactive Mobile Teaching Assistant (IMTA) back-end. Students pose questions verbally to the robot; speech is transcribed, routed to a retrieval-augmented generation (RAG) pipeline over course materials, and the generated answer is returned via speech and on-screen text. The pilot explicitly avoided collecting personal or biometric identifiers: interaction data were anonymized and processed on university-managed servers, under ethics approval and GDPR-compliant consent. While this provides a strong baseline on privacy and consent, many cybersecurity and governance controls remain in an early or ad hoc state, reflecting the prototype nature of the deployment.

Table 34. Case Study (D.P.) Evaluation

Control (D.P.)	Requirement	Assessment	Observations
D.P.-1	End-to-End Encryption	2 - Partially Fulfilled	Interaction logs and question–answer traffic are routed through an IMTA back-end running on university-managed infrastructure and calling external language models via web APIs. The study notes that communication data were securely processed and anonymized on university servers, which implies the use of standard encrypted channels (such as HTTPS) between the robot, the back-end, and cloud services. However, the deployment does not provide explicit documentation of protocol versions, certificate validation, key management, or encryption at rest for logs and vector embeddings. Since encryption appears to rely largely on default platform behaviour rather than a formally specified, audited end-to-end cryptographic design, this control is considered only partially fulfilled.
D.P.-2	Minimized Biometric Retention	3 - Fully Implemented	Unlike the attendance and evaluation robots, the question-answering assistant is designed not to capture or retain biometric identifiers. The paper explicitly states that no personal or biometric data were collected; speech is processed only long enough to obtain a text transcript, and no facial images or biometric templates are used at any stage. Interaction logs store de-identified question and answer content, without names or student IDs. As a result, the system effectively achieves biometric minimization by design: there are no biometric artefacts to retain or delete, and the risk surface associated with biometric storage is eliminated. This represents a strong alignment with privacy-by-design principles and fully satisfies this control.
D.P.-3	Secure Transmission Enforcement	2 - Partially Fulfilled	The architecture assumes that traffic between the robot, IMTA back-end, and external language model endpoints uses secure web protocols. In practice, these components are likely to rely on managed services and client libraries that enforce modern TLS for network communication. However, the pilot does not specify mechanisms for enforcing secure transmission (for example, rejecting invalid certificates, disabling legacy protocols, or logging attempts to communicate over insecure networks). There is also no evidence of monitoring for misconfigured Wi-Fi or fallback to less secure channels. As a result, encryption is plausibly in place but not clearly enforced or audited, so this control is rated as partially fulfilled.
D.P.-4	Purpose-Limited Data Use and Retention	2 - Partially Fulfilled	Data collected by the system consist mainly of de-identified interaction logs (student questions, system answers, and associated metadata such as time and lab session). These logs are used to evaluate answer accuracy, analyse user experience, and inform future development of the robot teaching assistant. The study frames this usage clearly within an educational and research context and does not report any secondary use such as profiling individual students. However, explicit retention schedules, automated deletion rules, and constraints on future reuse of logs are not described. There is no indicated mechanism for routinely reviewing whether stored logs remain necessary. Consequently, the deployment adheres to purpose limitation in practice but lacks formal retention and reuse policies, resulting in a partial rather than full implementation.

Table 35. Case Study (S.I.) Evaluation

Control (S.I.)	Requirement	Assessment	Observations
S.I.-1	Immutable Audit Logs	1 - Minimally Addressed	The IMTA back-end records interaction logs for research and evaluation, capturing which questions were asked and what answers the system produced. These logs support later analysis of accuracy and user experience, but there is no indication that they are implemented as tamper-evident or append-only audit trails. The database design is not described as hash-chained, digitally signed, or write-once, and there is no mention of independent forensic exports. Logging is therefore present but oriented toward usability and research rather than strict integrity assurance, so this control is only minimally addressed.
S.I.-2	Secure Update Pipeline	1 - Minimally Addressed	Updates to the IMTA components and robot-side software are managed by the research team and course instructors. While this provides a de facto gatekeeper function, the study does not describe any formal update pipeline with signed packages, versioned change logs, or controlled roll-back procedures. Updates are presumably applied ad hoc when new features are developed or when platform libraries change, and there is no reference to rejecting updates over insecure channels. As such, the control is acknowledged in practice (only a small group can change the system), but not implemented with the rigour expected of a secure update pipeline.
S.I.-3	Runtime Integrity Checks	0 - Not Addressed	No runtime integrity checks are described. There is no indication of file integrity monitoring for critical binaries, model checksum verification, or automated alerts in case of configuration drift or tampering. The system relies on the underlying robot platform and server environment to remain trustworthy, without additional application-level integrity protections. This control is therefore not addressed in the current pilot.
S.I.-4	Isolation of Trusted Components	1 - Minimally Addressed	Architecturally, the IMTA back-end separates several concerns: a vector database for course materials, a retrieval layer, and a language model interface. This logical separation provides some compartmentalization between content storage, retrieval logic, and answer generation. However, there is no evidence of hardened isolation (such as containerization, sandboxed processes, or strict inter-process authentication) specifically aimed at protecting security-critical components. The separation is primarily functional rather than security-driven, so this control is considered minimally addressed.

Table 36. Case Study (I.L.) Evaluation

Control (I.L.)	Requirement	Assessment	Observations
I.L.-1	Safe Defaults for Audio/Video	1 - Minimally Addressed	To function as a question-answering assistant, the robot's microphone remains active during lab sessions so that students can address it spontaneously. There is no explicit mention of session-based opt-in for each interaction, sensor indicators beyond the obvious presence of the robot, or default-off states for audio when the session is idle. On the positive side, the study reports that no raw audio or biometric recordings are stored; only text transcripts of questions and answers are retained in anonymized form. Overall, however, the configuration favours convenience during the pilot over strict safe-defaults for audio, so this control is rated as minimally addressed.
I.L.-2	Session Expiry & Timeout Controls	1 - Minimally Addressed	Lab sessions are bounded in time and supervised, and the robot is only used during structured teaching activities. After the lab concludes, the system is effectively taken out of service by the instructors or research team. Nevertheless, there is no indication of automated session timeouts, auto-logout of privileged interfaces, or explicit expiration of robot sessions after inactivity. Protection against lingering access depends on manual procedures rather than built-in timeout controls, so this requirement is only minimally met.
I.L.-3	Output Filtering & Response Censorship	2 - Partially Fulfilled	The IMTA pipeline restricts the robot's knowledge base primarily to vetted course materials such as lecture notes, slides, and lab guides, combined with general language model capabilities. Because the system does not have access to student records, grades, or other sensitive institutional data, the risk of leaking other students' personal information through answers is low by design. At the same time, there is no evidence of role-based filters, explicit blocking lists, or systematic monitoring of outputs for harmful or out-of-scope content. The RAG architecture provides some structural protection against data leakage, but formal output filtering mechanisms remain limited, so this control is assessed as partially fulfilled.
I.L.-4	Retention & Disclosure Policies	1 - Minimally Addressed	The research materials indicate that interaction data were anonymized and stored on university servers, with access limited to the research team and possibly teaching staff. However, the pilot does not specify a detailed retention schedule, disclosure conditions, or mechanisms by which students could request deletion or review of their interaction data. Governance of logs is therefore largely embedded in the broader research ethics approval rather than in a clear, robot-specific policy. This leaves the control only minimally implemented.

Table 37. Case Study (S.A.) Evaluation

Control (S.A.)	Requirement	Assessment	Observations
S.A.-1	Role-Based Access Control (RBAC)	1 - Minimally Addressed	Students interact with the robot only through its conversational interface; they do not receive credentials that would allow them to administer the device or back-end. Access to system configuration and development environments is restricted to the research team and course staff. However, there is no description of a formal RBAC model, periodic access reviews, or clearly defined role scopes in institutional policy. Separation of privileges exists informally but is not codified or routinely audited, so this control is rated as minimally addressed.
S.A.-2	Credential and Interface Security Hardening	0 - Not Addressed	The pilot does not provide details on password policies, two-factor authentication for administrative consoles, or restrictions on debug interfaces such as shell access or developer APIs on the robot platform. It is unclear whether default credentials were changed or whether interfaces were limited to specific IP ranges. In the absence of such evidence, hardening of credentials and exposed interfaces must be treated as not addressed in this prototype deployment.
S.A.-3	Access Attempt Logging & Anomaly Monitoring	0 - Not Addressed	Logging in this case focuses on interaction content (questions and answers) rather than on authentication events or administrative access patterns. The study does not mention any logging of login attempts, anomaly detection based on IP or device identifiers, or integration with a central security monitoring platform. Since access attempt logging and anomaly monitoring are not described, this control is considered not addressed.
S.A.-4	Privilege Escalation Controls and Re-Authentication	0 - Not Addressed	There is no indication that sensitive operations such as firmware updates, configuration changes, or data exports are gated with explicit re-authentication, step-up verification, or structured approval workflows. Administrative activity appears to be carried out by a small trusted team, but without additional technical controls to prevent misuse or escalation. As such, this control is not implemented in the current pilot.

Table 38. Case Study (V.P.) Evaluation

Control (V.P.)	Requirement	Assessment	Observations
V.P.-1	Scheduled Patch Application	1 - Minimally Addressed	The IMTA back-end and robot platform depend on a variety of software components, including the Temi operating environment, vector database, and language model libraries. The research team is likely to apply updates when needed (for instance when a new version of an API is required), but there is no mention of a scheduled patch cycle, a documented patch calendar, or dedicated time for security updates. Patch management is thus reactive and tied to development needs rather than a formal security process, earning a minimal implementation score.
V.P.-2	Periodic Vulnerability Scanning	0 - Not Addressed	No vulnerability scanning is described in the study. There is no indication that the robot or back-end services are routinely scanned with security tools, subjected to penetration tests, or continuously monitored for known vulnerabilities. This control is therefore not addressed in the pilot deployment.
V.P.-3	Secure Baseline Tracking	1 - Minimally Addressed	The project is implemented as an academic research prototype, and code is presumably maintained under version control. This provides some traceability of configuration changes and application versions. However, there is no explicit description of baseline security configurations, hardened default images, or comparison against known-good states during deployment. As a result, baseline tracking for security purposes is minimal.
V.P.-4	EOL and Unsupported Component Avoidance	2 - Partially Fulfilled	The system relies on current, vendor-supported components: a commercially supported social robot platform and modern cloud-based language models. There is no indication that end-of-life or unmaintained software is being used. At the same time, the pilot does not specify any institutional process for tracking end-of-life dates or planning migrations before support ends. The choice of components is sound, but procedures for monitoring lifecycles are not documented, so this control is partially fulfilled.

Table 39. Case Study (C.P.) Evaluation

Control (C.P.)	Requirement	Assessment	Observations
C.P.-1	Emergency Stop Access	2 - Partially Fulfilled	The Temi robot operated in a supervised lab environment, with the instructor or researcher physically present and able to intervene at any time. Students could also step back from the robot, and the device can be powered off or interrupted through its built-in controls if necessary. These mechanisms provide practical ways to stop the robot in case of an issue, and no physical incidents were reported during the pilot. However, there is no mention of clearly labeled emergency stop buttons, signage for students, or formal training on how to halt the robot. Emergency stop capability therefore exists in practice but is not fully formalized or communicated, leading to a partial implementation.
C.P.-2	Motion Sandbox Testing	2 - Partially Fulfilled	The robot's movement during the pilot was limited to the aisles between lab benches, where students worked at fixed stations. The deployment team tested the interaction flow and navigation in advance, and the robot moves at low speeds suitable for indoor environments. Nevertheless, there is no description of systematic motion sandbox testing, documented hazard analysis, or scenario-based safety validation beyond informal pretests. This yields a partial fulfillment of the motion testing requirement.
C.P.-3	Safety Zones and No-Go Areas	2 - Partially Fulfilled	In practice, the physical layout of the lab and the way the robot was used created de facto safety zones: the robot remained within classroom boundaries and followed predictable routes near desks. Human supervisors ensured that the robot did not approach sensitive areas such as doorways or cluttered spaces. However, no explicit software-enforced no-go zones, geofencing rules, or formally mapped safety perimeters are described. The control is thus partially satisfied through operational constraints rather than formal configuration.
C.P.-4	Physical Activity Logging	1 - Minimally Addressed	The primary logging focus of the deployment was on interaction content and accuracy, not on physical motion or near-miss events. There is no discussion of recording trajectories, collision events, or emergency interventions in a structured way that would support safety analytics. Any physical issues would likely be remembered informally by staff rather than captured in telemetry. As a result, physical activity logging is only minimally addressed.

Table 40. Case Study (N.C.) Evaluation

Control (N.C.)	Requirement	Assessment	Observations
N.C.-1	Encrypted Network Protocols	2 - Partially Fulfilled	The robot communicates with the IMTA back-end and external language model services using standard web-based APIs, which in contemporary deployments generally rely on HTTPS with modern TLS. The study further notes that communication data were securely processed on university servers, suggesting that unencrypted protocols are not used for normal operation. However, there is no explicit statement of protocol versions, cipher suites, or prohibitions on deprecated configurations, nor evidence of technical measures that would block unsecured channels. Hence this control is treated as partially fulfilled rather than fully enforced.
N.C.-2	Network Segmentation and Firewalls	1 - Minimally Addressed	The deployment took place on a university campus network, and the robot likely connected via institutionally managed Wi-Fi or a dedicated lab network. While university infrastructure usually provides basic firewalling and segmentation between guest and internal networks, the study does not specify whether the robot was placed on a dedicated VLAN or isolated from student and public traffic. Absent such detail, network segmentation is assumed to be handled indirectly by general campus policies rather than through robot-specific design, so this control is minimally addressed.
N.C.-3	Endpoint Authentication and Validation	1 - Minimally Addressed	The IMTA system must authenticate to external language model services using API keys or similar credentials, and the robot connects to the back-end using configured endpoints. This implies some level of endpoint validation and key-based authentication. However, there is no mention of mutual authentication, certificate pinning, or strict per-device trust lists. Because authentication is implied but not hardened or documented, the control is rated as minimally implemented.
N.C.-4	Replay and Spoofing Protection	0 - Not Addressed	The pilot does not describe any protection against replay attacks or spoofed messages. There is no evidence of nonces, timestamps, rate limiting for suspicious traffic, or session key rotation specifically aimed at preventing impersonation. The relatively small scale of the deployment and low perceived threat surface likely contributed to this omission. As a result, this control is not addressed in the current system.

Table 41. Case Study (I.R.) Evaluation

Control (I.R.)	Requirement	Assessment	Observations
I.R.-1	Mandatory IT Security Involvement	1 - Minimally Addressed	The deployment was carried out under a formal research and ethics approval process, and the IMTA back-end ran on university-managed infrastructure. This implies some level of coordination with institutional IT and data protection staff, particularly regarding server hosting and GDPR compliance. Nevertheless, there is no explicit indication that cybersecurity teams formally reviewed the robot deployment, network integration, or ongoing security responsibilities. Security involvement therefore appears supportive but not mandated as a structured review process, resulting in a minimal implementation.
I.R.-2	Robot-Aware Policy Integration	0 - Not Addressed	The case study is framed as a pilot within a single course rather than as part of a university-wide robot deployment programme. Existing acceptable use, privacy, and technology policies apply in a generic way, but there is no mention of updates that explicitly address mobile robots or AI assistants in teaching spaces. As such, the robot is operating largely under general research and IT rules, without dedicated policy integration, and this control is not addressed.
I.R.-3	Incident Response Inclusion	0 - Not Addressed	No robot-specific incident response procedures are described. If the system failed or produced problematic behaviour, staff would react in an ad hoc fashion by shutting down the robot or answering questions manually. There is no reference to incident playbooks, reporting channels for student concerns, or integration of robot-related scenarios into institutional incident response plans. This control is therefore not implemented.

Table 42. Case Study (P.C.) Evaluation

Control (P.C.)	Requirement	Assessment	Observations
P.C.-1	Opt-in Consent for Data Collection	3 - Fully Implemented	Participation in the pilot was explicitly voluntary. Students were invited to interact with the robot teaching assistant during lab sessions and were informed that their interactions would be logged and analysed for research. Consent was obtained in accordance with research ethics and GDPR requirements, and students were free to refrain from using the robot and rely solely on the human instructor. While consent is not granular by sensor type, the combination of voluntary participation, clear study framing, and the absence of biometric data collection meets the spirit of opt-in consent for this prototype. This control is therefore treated as fully implemented in the context of the pilot.
P.C.-2	Transparent Data Usage Notices	3 - Fully Implemented	The study materials explain that interaction data would be anonymized and processed on university servers, and that the purpose of data collection was to evaluate and improve the robot teaching assistant. Students were aware that the robot was part of a research project and that their questions and answers could be recorded for analysis. The scope of data collection is limited to content necessary for the question-answering function, and no hidden or secondary uses are implied. Although detailed retention timelines are not prominently featured in the interface, overall transparency around what is collected and why is strong for a pilot, so this control is assessed as fully implemented.
P.C.-3	Surveillance Indicators and Controls	1 - Minimally Addressed	Students can clearly see and hear when the robot is active and responding, and they engage with it intentionally. However, there are no dedicated visual indicators (such as recording icons or sensor LEDs) that clearly distinguish when microphones are active or when interaction logging is occurring. Students also cannot toggle sensors off directly; their main control is simply to not speak to the robot. Given the limited scope of data collection and the voluntary nature of participation, surveillance risks are modest but still present. This control is therefore only minimally addressed.
P.C.-4	Data Portability and Erasure Options	0 - Not Addressed	Interaction logs are anonymized and do not appear to be linked to easily accessible student identifiers, which reduces privacy risk but also makes individual data access difficult. The pilot does not offer a self-service interface by which students can view, export, or delete their interaction history, nor is such a process described in supporting documentation. Any erasure request would need to be handled manually through general data protection channels, rather than a robot-specific process. As a result, this control is not addressed in the current deployment.

Table 43. Case Study (T.E.) Evaluation

Control (T.E.)	Requirement	Assessment	Observations
T.E.-1	Disclosure of Non-Human Identity	3 - Fully Implemented	The robot is clearly presented to students as an AI-based teaching assistant and not as a human tutor. Its physical embodiment as a mobile robot, the context of deployment within a technology course, and the introductory explanations by teaching staff make its non-human nature obvious. Students understand that answers are generated by a robot and underlying AI system, and the study evaluates their trust and expectations on that basis. This control is therefore fully satisfied.
T.E.-2	Explanation Request Interface	2 - Partially Fulfilled	When students ask follow-up questions such as "why" or request clarification, the robot can attempt to elaborate using the underlying language model and course materials. The screen can display textual explanations alongside spoken answers, and students can rephrase or refine their queries. However, there is no dedicated "explain" button, no structured rationale view, and no interface that explicitly distinguishes between an answer and an explanation of the decision process. Explanations are available in a conversational but ad hoc manner, so this control is partially fulfilled.
T.E.-3	Documentation of Decision Logic	1 - Minimally Addressed	The study describes the high-level architecture of the IMTA system, including the use of retrieval-augmented generation over course materials and language models. This documentation exists for researchers and reviewers, but there is no detailed institutional documentation that systematically catalogues model assumptions, known failure modes, or interpretability limitations with an eye toward long-term governance. As such, there is a basic description of how the system works, but not the comprehensive decision logic documentation envisioned by the guideline. The control is therefore minimally addressed.
T.E.-4	Role Transparency in Educational Tasks	3 - Fully Implemented	The robot's role in the course is limited to answering student questions and providing on-demand explanations during lab sessions. It does not grade students, manage attendance, or make binding decisions about academic outcomes. This scope is clearly communicated to students, and survey items focus on perceived usefulness and learning support rather than on high-stakes decision-making. Because the robot's role is narrow, supportive, and well understood by participants, role transparency in educational tasks is fully implemented.

Table 44. Case Study (A.L.) Evaluation

Control (A.L.)	Requirement	Assessment	Observations
A.L.-1	Assigned Human Supervisor	3 - Fully Implemented	During all lab sessions, a human instructor or researcher was present alongside the robot. Students were encouraged to seek help from the human teacher if the robot's answer was unclear, incorrect, or incomplete. Responsibility for the course content and learning outcomes clearly remained with the human teaching staff, and the robot was treated as an assistive tool. This explicit human oversight and role assignment fully satisfies the requirement for a designated supervisor.
A.L.-2	Defined Student Appeal Pathways	2 - Partially Fulfilled	Because the robot does not make binding decisions about grades or progression, the main "appeal" mechanism is simply to ask the human instructor for clarification or a second opinion. This informal pathway is straightforward and accessible for students but is not codified as a formal appeals process. In a production setting, clearer guidance could be provided about how to report problematic robot behaviour or seek a review of advice given by the system. As implemented, appeal pathways are present but not formally documented, so the control is partially fulfilled.
A.L.-3	Source Attribution Logging	2 - Partially Fulfilled	Interaction logs store which questions were asked and what answers were produced, enabling researchers to later review specific responses. This provides a degree of traceability for the robot's outputs and supports post-hoc analysis of failures or misunderstandings. However, logs are anonymized and not tied to explicit decision identifiers or structured event types, and there is no integrated mechanism for linking a student's complaint to a specific logged output. Source attribution is therefore possible but limited, resulting in a partial implementation.
A.L.-4	Institutional Responsibility Policies	1 - Minimally Addressed	Responsibility for the pilot lies mainly with the research team and course instructor, under general institutional research and teaching policies. There are no robot-specific policies that clarify institutional liability if the system misinforms a student or contributes to poor performance. Given the low stakes of the deployment and the continuous presence of a human teacher, this gap is manageable in the pilot phase but would need to be addressed for broader adoption. Overall, institutional responsibility is acknowledged only at a general level, so this control is minimally addressed.

Table 45. Case Study (H.O.) Evaluation

Control (H.O.)	Requirement	Assessment	Observations
H.O.-1	Human Oversight for Critical Decisions	3 - Fully Implemented	The robot teaching assistant supports, but does not replace, human judgment in the course. All critical decisions about grading, progression, and assessment remain the responsibility of the instructor. Students are encouraged to treat robot answers as supportive explanations rather than authoritative final decisions, and they can verify any guidance with the human teacher on the spot. As a result, human oversight is robustly in place for all decisions that could materially affect students.
H.O.-2	Manual Override and Intervention Tools	2 - Partially Fulfilled	Instructors and researchers can intervene at any time by stopping the robot, answering questions themselves, or reframing the interaction. If the robot is obviously confused or malfunctioning, staff can physically move or power it down. These manual overrides are effective in practice but are not exposed through a dedicated supervisor interface or clearly labelled "stop answering" controls. Intervention capabilities are therefore present but informally implemented, resulting in a partial score.
H.O.-3	Ongoing Monitoring and Feedback	1 - Minimally Addressed	The pilot collected student feedback through surveys and analysed interaction logs to assess accuracy and user satisfaction. This provides some retrospective monitoring of the system's performance and its impact on learning. However, there is no continuous monitoring dashboard, no in-session feedback channel from students beyond normal conversation, and no routine review workflow for staff between sessions. Ongoing oversight mechanisms are thus limited and largely research-driven, leaving this control minimally addressed.

Table 46. Case Study (B.F.) Evaluation

Control (B.F.)	Requirement	Assessment	Observations
B.F.-1	Use-Case Specific Fairness Assessment	1 - Minimally Addressed	The study focuses on technical accuracy and user experience rather than on systematic fairness evaluation across different student groups. All students in the lab have access to the same robot and the same underlying course materials, which helps ensure consistent treatment at a basic level. At the same time, potential disparities arising from speech recognition (for example for students with strong accents or speech differences) are acknowledged but not measured. Fairness is therefore considered at a conceptual level but not rigorously analysed, so this control is minimally addressed.
B.F.-2	Performance Disparity Audits	0 - Not Addressed	No stratified analysis of performance is reported across demographic or linguistic subgroups. The pilot does not evaluate whether particular categories of students are more likely to experience misrecognitions, longer response times, or incorrect answers. As a result, the system may inadvertently work better for some students than others without this being detected. Since no structured disparity audits are carried out, this control is not addressed.
B.F.-3	Inclusive Dataset and Language Use	2 - Partially Fulfilled	The question-answering pipeline is grounded in official course materials, which are curated and approved by teaching staff. This reduces the risk of overtly biased or inappropriate content within the domain of networking concepts. However, the underlying language model is trained on broad data and may carry general-purpose biases, and the course materials themselves have not been explicitly audited for inclusive language. Furthermore, the system is not designed to handle sensitive identity-related topics. The use of vetted course content offers some protection but falls short of a comprehensive inclusion strategy, so the control is partially fulfilled.
B.F.-4	Inclusive Interaction Design	1 - Minimally Addressed	Interaction is primarily voice-based, requiring students to speak clearly to the robot in English in a noisy lab environment. There is limited accommodation for students with speech impairments, strong accents, or anxiety about speaking aloud; there is also no separate text-only interface for those who might prefer reading and typing. While human instructors can step in to support students who struggle with the robot, the interface itself is not explicitly designed for inclusivity. This control is therefore minimally addressed.

Table 47. Case Study (E.A.) Evaluation

Control (E.A.)	Requirement	Assessment	Observations
E.A.-1	Multimodal Interaction and Accessibility	1 - Minimally Addressed	The robot provides spoken answers and displays text on its screen, offering basic multimodality for students who prefer reading over listening. However, interaction input is speech-only, and there is no alternative keyboard, switch-based, or remote interface for students who cannot or do not wish to speak aloud. Accessibility features such as screen readers, high-contrast modes, or compatibility with assistive devices are not discussed. As a result, multimodal accessibility is present in a limited form and only minimally meets the guideline.
E.A.-2	Accessibility Testing and Co-Design	0 - Not Addressed	The pilot does not report any involvement of accessibility specialists or students with disabilities in the design and testing process. Evaluation focuses on usability and engagement for the general student population rather than on meeting specific accessibility standards. Consequently, this control is not addressed.
E.A.-3	Economic and Resource Accessibility	1 - Minimally Addressed	All enrolled students in the lab sections where the robot was deployed had access to the system as part of the course; there were no additional fees or personal devices required. However, the robot is available only in the physical lab setting and only during the pilot sessions; students who could not attend those sessions in person do not benefit from the robot assistant. There is no remote or home-accessible equivalent. This leads to a minimally addressed level of economic and resource accessibility.
E.A.-4	Institutional Monitoring for Disparities	0 - Not Addressed	No institutional mechanism is in place to monitor whether the robot teaching assistant is improving or worsening equity in access to support. There is no tracking of which students use the robot, how frequently, or whether certain groups systematically avoid or benefit less from it. Without such monitoring, equity gaps might go unnoticed. This control is therefore not addressed in the current deployment.

Control Area Scoring Summary

The following table summarizes the control evaluation results across all cybersecurity and ethical domains for the robot assistant that answers student questions in lab sessions. Each control domain contains multiple individual controls evaluated against this deployment. Scores reflect the average degree of fulfillment on a 0–3 scale, and the final column shows the percentage of controls in that domain which meet at least a partial threshold (score of 2 or higher).

Table 48. Detailed Control Scoring per Domain (Question-Answering Assistant)

Control Domain	**1**	**2**	**3**	**4**	**Avg. Score**	**% $\geq$ 2**
Data Privacy & Confidentiality (D.P.)	2	3	2	2	2.25	100%
System & Data Integrity (S.I.)	1	1	0	1	0.75	0%
Information Leakage & Misuse (I.L.)	1	1	2	1	1.25	25%
System Access & Control (S.A.)	1	0	0	0	0.25	0%
Vulnerability & Patch Management (V.P.)	1	0	1	2	1.00	25%
Cyber-Physical Safety (C.P.)	2	2	2	1	1.75	75%
Network & Communication Security (N.C.)	2	1	1	0	1.00	25%
Institutional Readiness & Governance (I.R.)	1	0	0	–	0.33	0%
Privacy, Consent (P.C.)	3	3	1	0	1.75	50%
Transparency & Explainability (T.E.)	3	2	1	3	2.25	75%
Accountability & Liability (A.L.)	3	2	2	1	2.00	75%
Autonomy & Human Oversight (H.O.)	3	2	1	–	2.00	66.7%
Bias, Fairness & Inclusion (B.F.)	1	0	2	1	1.00	25%
Equity & Accessibility (E.A.)	1	0	1	0	0.50	0%
Overall					**1.30**	**38.9%**

Appendix 4 – Case Study 3: Robot Assistant for Automated Task Evaluation

This appendix contains the full Control Evaluation Matrix and related tables for the Temi-based robot teaching assistant that performs automated task evaluation and oral knowledge assessment in computer networking labs. The main text in Chapter 6 provides a narrative synthesis of this case study and high-level comparison with the other deployments; here, all detailed per-control scores and comments are reported.

In this deployment, the robot guides students through a three-stage workflow: (i) checking router/switch configurations via serial connection against a reference configuration and connectivity tests; (ii) conducting an oral knowledge assessment based on a question bank; and (iii) delivering structured feedback on pass/fail status and identified issues. Interaction data and results are pseudonymised (e.g., by desk number), processed on university-managed servers, and collected under research ethics approval with voluntary participation. Although the system does not directly assign final course grades, it performs evaluation tasks similar to high-risk educational AI systems, making it a particularly relevant test for the framework's cybersecurity, ethical, and governance controls.

Table 49. Case Study (D.P.) Evaluation

Control (D.P.)	Requirement	Assessment	Observations
D.P.-1	End-to-End Encryption	2 - Partially Fulfilled	The automated evaluation workflow involves several communication channels: between the robot and the IMTA back-end, between back-end services and networking devices via serial-over-IP or similar, and between IMTA and any external language model components used for oral examination logic. These interactions are presumed to rely on contemporary web and network protocols (for example HTTPS and secure tunnels) provided by the underlying platforms and university infrastructure. However, the deployment does not document explicit encryption requirements, protocol configurations, or key management practices, and there is no evidence of systematic verification that all links are protected against downgrade and man-in-the-middle attacks. Sensitive exam results and configuration snapshots are therefore likely protected in transit by default platform behaviour, but the absence of a formal end-to-end cryptographic design and auditing justifies only a partially fulfilled assessment.
D.P.-2	Minimized Biometric Retention	3 - Fully Implemented	The evaluation robot does not rely on biometric identifiers such as facial images or stored voice templates. Students are identified operationally by desk number or similar pseudonyms; the system evaluates the correctness of router configurations and oral responses without storing biometric features. Oral answers are processed as transient audio signals converted to text and then immediately evaluated. Logs record pseudonymous interaction identifiers and results (e.g., “configuration correct/incorrect”, “oral exam 3/5 correct”), but no biometric data are retained. In effect, the system sidesteps biometric retention risks entirely through design, fully satisfying this control.
D.P.-3	Secure Transmission Enforcement	2 - Partially Fulfilled	In addition to relying on secure-by-default protocols, a high-stakes evaluation system should explicitly enforce secure transmission: for example, by refusing to operate on unsecured Wi-Fi, validating certificates, and logging any attempted insecure connections. The pilot deployment does not specify such enforcement mechanisms. While it is probable that the robot and back-end use TLS-based APIs and institutionally managed networks, there is no assurance that insecure configurations (e.g., open access points, misconfigured switches) would be detected or blocked. As a result, secure transmission is implicitly present but not explicitly enforced or monitored, leading to a partially fulfilled rating.
D.P.-4	Purpose-Limited Data Use and Retention	2 - Partially Fulfilled	The primary purpose of data collection in this deployment is to support automated configuration checks and oral examinations, and to analyse whether the robot can provide fair evaluations compared to human instructors. Logs capture pseudonymous evaluation events and are used for both pedagogical feedback and research on fairness perceptions. There is no indication that results are reused for marketing, behavioural profiling, or unrelated analytics. However, the case study does not describe explicit data retention limits, deletion schedules, or controls on secondary research uses beyond the scope of the original ethics approval. Without codified retention and reuse policies, purpose limitation is mostly realised in practice but not fully formalised, so this control is partially implemented.

Table 50. Case Study (S.I.) Evaluation

Control (S.I.)	Requirement	Assessment	Observations
S.I.-1	Immutable Audit Logs	2 - Partially Fulfilled	To evaluate fairness and system performance, the deployment records rich logs for each evaluation: configuration snapshots, comparison results against the reference configuration, oral questions posed, and scores achieved. These logs enable post-hoc analysis of how decisions were made and support the human lecturer in reviewing negative outcomes. However, the underlying storage is not described as append-only, hash-chained, or otherwise tamper-evident. While practical traceability is high, the logs could theoretically be altered by privileged users without detection. As a result, the system achieves useful auditability but not full immutability, warranting a partially fulfilled rating.
S.I.-2	Secure Update Pipeline	1 - Minimally Addressed	Updates to the evaluation logic (e.g., reference configurations, question bank content, scoring thresholds) and to the IMTA platform are managed by the research team and course instructor. This restricts the number of people who can modify the system, but the pilot does not mention signed update packages, change management procedures, or separate staging and production environments. The risk of accidental or malicious introduction of erroneous evaluation rules is mitigated mainly by the small, trusted project team and by the instructor's manual oversight, not by a formally secure update pipeline. Hence, this control is only minimally addressed.
S.I.-3	Runtime Integrity Checks	1 - Minimally Addressed	The system performs runtime checks on student configurations against a stored reference and verifies network connectivity. These checks help ensure that evaluations are technically correct but are not aimed at detecting tampering with the application itself. There is no mention of verifying binary integrity, model hashes, or configuration signatures at startup or during operation. Any corruption or unauthorised change to the evaluation logic would likely be noticed only indirectly through anomalous results. Consequently, runtime integrity checking is present in a functional sense but does not meet the security-focused intent of this control, leading to a minimal score.
S.I.-4	Isolation of Trusted Components	1 - Minimally Addressed	Architecturally, the evaluation workflow separates several components: front-end dialogue on the robot, configuration retrieval, comparison against reference states, and oral exam logic. This modularity provides some natural compartmentalisation. However, the case study does not describe hardened isolation boundaries, sandboxed evaluation engines, or strict inter-process access controls. Trusted assets such as reference configurations and scoring rules likely reside on the same back-end infrastructure as other IMTA services. As a result, component isolation is largely functional and not driven by security design, so this control is minimally implemented.

Table 51. Case Study (I.L.) Evaluation

Control (I.L.)	Requirement	Assessment	Observations
I.L.-1	Safe Defaults for Audio/Video	1 - Minimally Addressed	The robot relies on speech input to conduct the oral exam portion of the evaluation. Microphones must therefore remain active while the robot interacts with students, typically in a supervised lab setting. The deployment avoids storing raw audio and does not use cameras for facial recognition, which reduces the risk of persistent surveillance. Nonetheless, there are no explicit controls such as per-session microphone toggles, visual recording indicators, or guaranteed default-off states outside evaluation sessions. Sensor usage is managed informally through supervised operation rather than strict safe defaults, so this control is minimally addressed.
I.L.-2	Session Expiry & Timeout Controls	1 - Minimally Addressed	Each evaluation is a discrete session: a student summons the robot, specifies a desk, connects the device, and answers a bounded set of questions. Once feedback is delivered, the session ends and the robot returns to an idle state. While this natural session structure limits prolonged exposure, the pilot does not describe automatic timeouts for incomplete evaluations, auto-logout of privileged interfaces, or forced termination of long-running connections. Session expiry is therefore supported by usage patterns rather than explicit timeout controls, leading to a minimal implementation.
I.L.-3	Output Filtering & Response Censorship	2 - Partially Fulfilled	The evaluation robot issues highly structured outputs: pass/fail decisions, lists of misconfigured parameters, and standardised messages advising students to contact the instructor when needed. It does not have access to other students' grades or personal data, and its utterances are constrained by the evaluation templates and question bank. This significantly reduces the risk of disclosing sensitive information. However, there is no explicit output filtering engine checking for inadvertent leakage (e.g., internal identifiers, configuration details that should not be exposed beyond the lab), nor a mechanism for classifying and censoring unexpected responses from underlying AI components. As a result, output risks are mitigated structurally but not systematically, resulting in a partially fulfilled rating.
I.L.-4	Retention & Disclosure Policies	1 - Minimally Addressed	Evaluation logs, including configuration correctness and oral exam scores, are stored for research and analysis of fairness. Access is limited to the project team and teaching staff, and no student names are included. Yet the study does not specify formal policies for how long these logs are retained, under what conditions they might be shared beyond the immediate research context, or how students can find out whether their data are included. Retention and disclosure are thus governed primarily by generic research ethics practices rather than a robot-specific policy, leaving this control minimally implemented.

Table 52. Case Study (S.A.) Evaluation

Control (S.A.)	Requirement	Assessment	Observations
S.A.-1	Role-Based Access Control (RBAC)	1 - Minimally Addressed	Students only interact with the robot through its evaluation dialogue and do not receive any administrative credentials. Configuration of the evaluation logic, question bank, and IMTA back-end is restricted to the research team and course instructor. However, the case study does not describe a formal RBAC model with distinct roles for developer, operator, and auditor, nor does it mention periodic access reviews. Privilege separation exists informally by limiting access to a small trusted group, but it is not codified or systematically enforced, so this control is minimally implemented.
S.A.-2	Credential and Interface Security Hardening	0 - Not Addressed	The pilot does not report any robot-specific hardening such as enforcing strong password policies, enabling multi-factor authentication for administrative interfaces, or disabling unused debug ports and default accounts. It is unclear whether default vendor credentials were changed or whether access to back-end management consoles is limited by IP or VPN. In the absence of evidence of intentional hardening beyond general institutional defaults, this control is considered not addressed.
S.A.-3	Access Attempt Logging & Anomaly Monitoring	0 - Not Addressed	Logging in this deployment focuses on evaluation events and student performance, with no mention of systematic recording and monitoring of authentication attempts, failed logins, or unusual administrative activity. There is no indication of integration with a central security information and event management (SIEM) system or any dedicated anomaly detection for the robot infrastructure. Accordingly, this control is not implemented.
S.A.-4	Privilege Escalation Controls and Re-Authentication	0 - Not Addressed	Operations that could materially affect evaluation outcomes (such as modifying reference configurations, changing question bank content, or adjusting scoring thresholds) are not protected by explicit re-authentication or step-up approval, beyond whatever is provided by local login sessions. The pilot does not describe workflow controls (e.g., dual approval) for high-impact changes. Given the small, trusted project team, this may be acceptable in a research setting but does not constitute a technical privilege escalation control. This requirement is therefore not addressed.

Table 53. Case Study (V.P.) Evaluation

Control (V.P.)	Requirement	Assessment	Observations
V.P.-1	Scheduled Patch Application	1 - Minimally Addressed	The evaluation robot relies on the Temi platform, IMTA back-end components, network libraries, and potentially external AI services. The project team likely updates these components when functional needs arise or when vendors change APIs; however, there is no documented patch schedule, no explicit regular maintenance window, and no tracking of security-only patch deployment timelines. Patch management is reactive and ad hoc, resulting in a minimally addressed control.
V.P.-2	Periodic Vulnerability Scanning	0 - Not Addressed	The case study does not mention any active vulnerability scanning of the robot or its back-end services, nor formal penetration testing or code security reviews. In the absence of such activities, this control is considered not implemented for the pilot deployment.
V.P.-3	Secure Baseline Tracking	1 - Minimally Addressed	Project code and configurations are presumably maintained under version control, enabling developers to track feature changes and experiment variants. However, there is no indication of a security baseline that defines required configuration parameters, hardening settings, or known-good images. There is also no mention of checks that compare deployed systems against this baseline. Secure baseline tracking is therefore minimal.
V.P.-4	EOL and Unsupported Component Avoidance	2 - Partially Fulfilled	The deployment uses actively supported hardware and software: a modern commercial robot platform, current server environments, and contemporary language model services. This reduces the immediate risk of relying on unsupported components. Nevertheless, there is no institutional process described for tracking end-of-life announcements or for planning migrations before support lapses. The choice of components is sound, but lifecycle monitoring is informal, resulting in a partially fulfilled assessment.

Table 54. Case Study (C.P.) Evaluation

Control (C.P.)	Requirement	Assessment	Observations
C.P.-1	Emergency Stop Access	2 - Partially Fulfilled	The Temi robot operates at low speeds in a structured lab environment, travelling mainly between the entrance and workstations. Students and the instructor can always step away from the robot and interrupt interactions, and the device can be stopped or powered down using its built-in controls. Staff are present during evaluations and can intervene quickly in case of an unexpected motion or collision risk. However, there are no explicitly labelled emergency stop buttons, no standardised explanation to students about how to halt the robot, and no documented safety drills. Emergency stop capability exists in practice but lacks formalisation and user training, resulting in a partial score.
C.P.-2	Motion Sandbox Testing	2 - Partially Fulfilled	Before being used with students, the evaluation workflow and navigation behaviour were tested by the project team in the actual lab environment. Routes, speed, and interaction timing were tuned to minimise interference with student movement around desks. These pre-tests function as informal sandbox testing. Yet the study does not report structured hazard analysis (e.g., FMEA for motion scenarios), exhaustive scenario testing, or documentation of test results as part of a safety file. Therefore, the requirement is partially, but not fully, fulfilled.
C.P.-3	Safety Zones and No-Go Areas	2 - Partially Fulfilled	The robot's operating area is naturally constrained by the classroom layout and its task: it moves along aisles to specific desks and back to a waiting zone. Instructors are aware of the robot's paths and ensure that it does not navigate into particularly cluttered or hazardous areas. However, there is no mention of software-enforced no-go zones, map annotations marking restricted regions, or geofencing to prevent navigation outside the lab. Safety zones thus exist as operational conventions rather than explicit configurations, yielding a partial implementation.
C.P.-4	Physical Activity Logging	1 - Minimally Addressed	The deployment focuses logging on evaluation results rather than on physical telemetry. Collisions, near misses, and emergency interventions are not recorded as structured events, and there is no motion analytics dashboard. Any physical incidents would be noticed and remembered by staff on an ad hoc basis. Consequently, physical activity logging is only minimally addressed.

Table 55. Case Study (N.C.) Evaluation

Control (N.C.)	Requirement	Assessment	Observations
N.C.-1	Encrypted Network Protocols	2 - Partially Fulfilled	The evaluation robot relies on multiple networked interactions: contacting IMTA services, possibly reaching language model APIs, and performing connectivity tests to lab devices. It is standard practice for these communications to use encrypted network protocols, particularly for web APIs. The study also notes that data are processed on secure university servers. Nevertheless, there is no explicit enumeration of protocol versions, cipher suites, or configuration hardening (e.g., disabling older TLS versions). Given this lack of detailed evidence, encrypted protocols are assumed but not fully verified, yielding a partial implementation.
N.C.-2	Network Segmentation and Firewalls	1 - Minimally Addressed	The lab environment benefits from the university's general network security measures, such as basic firewalling and segmentation between guest and internal networks. However, the case study does not state whether the robot and IMTA servers operate on a dedicated VLAN, whether they are isolated from student devices apart from necessary console connections, or whether access rules are configured specifically for this deployment. Network segmentation is therefore presumed to exist at a generic campus level but not tailored to the robot, so this control is minimally implemented.
N.C.-3	Endpoint Authentication and Validation	1 - Minimally Addressed	IMTA components authenticate to each other and to external services via configured endpoints and credentials. Routers and switches are accessed via console or management interfaces using lab-specific configurations. While this implies basic endpoint authentication, the study does not mention mutual TLS, strong host identity verification, or strict per-device allowlists. Endpoint validation is present in a functional sense (correct credentials are needed) but does not appear hardened against spoofing, leaving this control minimally addressed.
N.C.-4	Replay and Spoofing Protection	0 - Not Addressed	No measures are described to prevent replay or spoofing of evaluation requests, such as nonces, timestamps tied to sessions, or rate limiting. The small scale of the deployment and physical supervision of the lab reduce the likelihood of such attacks, but they do not eliminate them. Without specific protections, this control remains not addressed.

Table 56. Case Study (I.R.) Evaluation

Control (I.R.)	Requirement	Assessment	Observations
I.R.-1	Mandatory IT Security Involvement	1 - Minimally Addressed	The deployment was carried out with ethics committee approval and used university-managed infrastructure for IMTA services. This implies some coordination with institutional IT and data protection officers, particularly about hosting and compliance with GDPR. However, the case study does not indicate that cybersecurity specialists formally reviewed the system design, threat model, or long-term risk management plan. IT security involvement is supportive rather than mandated as a defined step in project approval, resulting in a minimal score.
I.R.-2	Robot-Aware Policy Integration	0 - Not Addressed	The evaluation robot operates as part of a course-specific research pilot, under general teaching and research policies, but there are no institution-wide policies that explicitly address robot-based automated evaluation in educational settings. Concepts such as acceptable uses of robots for assessment, obligations for logging and fairness monitoring, or restrictions on field-of-use are not reflected in documented institutional policies. Therefore, this control is not addressed.
I.R.-3	Incident Response Inclusion	1 - Minimally Addressed	When the robot produces an unexpected or apparently unfair result, the lecturer manually rechecks the configuration and oral answers and can adjust the evaluation accordingly. This practice provides a de facto mechanism for handling "micro-incidents" at the course level. Nonetheless, there is no integration of robot-related scenarios into the university's formal incident response procedures, no dedicated reporting channel for systemic issues, and no playbook for serious failures. Incident handling is thus present but highly local and informal, meriting a minimal rating.

Table 57. Case Study (P.C.) Evaluation

Control (P.C.)	Requirement	Assessment	Observations
P.C.-1	Opt-in Consent for Data Collection	3 - Fully Implemented	Student participation in the evaluation pilot is voluntary and conducted under formal research ethics approval. Students are informed that the robot will evaluate their configurations and ask oral questions, and that results may be used anonymously for research on fairness and usability. Those who do not wish to use the robot can instead be evaluated directly by the instructor. This combination of explicit information, voluntariness, and alternative human pathways satisfies the requirement for opt-in consent in the pilot setting.
P.C.-2	Transparent Data Usage Notices	3 - Fully Implemented	The study clearly states that evaluation data are pseudonymised and used to compare robot and human assessment, to analyse error types, and to explore students' perceptions of fairness. Data are stored on secure university servers, and no personal identifiers are processed by the robot. Students are told that the robot's outputs will not directly determine their final course grade during the pilot. While interfaces do not display detailed privacy notices, the overall transparency about purposes and data flows is strong and appropriate to the context, leading to a fully implemented rating.
P.C.-3	Surveillance Indicators and Controls	1 - Minimally Addressed	Students explicitly engage with the robot when they are ready to be evaluated, and the interaction is bounded to that context; thus there is no hidden or continuous monitoring of behaviour across the course. However, there are no dedicated visual indicators to show when audio is being processed or when evaluation logging is active, nor user-accessible toggles for sensors. The main control students have is the ability to choose a human evaluation path instead of interacting with the robot. Given the limited scope of surveillance but scarcity of fine-grained controls, this requirement is minimally addressed.
P.C.-4	Data Portability and Erasure Options	1 - Minimally Addressed	Evaluation logs are pseudonymised but remain linked to course records through desk numbers or internal mappings. If a student contests a result, the lecturer can refer to the robot's log for that evaluation and adjust the outcome, effectively modifying the data that matter for the course. However, there is no explicit robot-specific process or interface for students to request full deletion or export of their evaluation logs beyond general GDPR procedures. As such, limited correction is possible through the instructor, but broader portability and erasure options are not systematically provided, yielding a minimal score.

Table 58. Case Study (T.E.) Evaluation

Control (T.E.)	Requirement	Assessment	Observations
T.E.-1	Disclosure of Non-Human Identity	3 - Fully Implemented	The robot is clearly presented as an automated teaching assistant, not as a human examiner. Its physical form and the surrounding communication by staff make its non-human nature obvious. Students understand that their configurations are being evaluated by a system and that the oral questions are delivered by a robot interfacing with an AI back-end. The study specifically explores students' perceptions of the robot as a teacher, indicating that this aspect of transparency is well established.
T.E.-2	Explanation Request Interface	2 - Partially Fulfilled	When the robot finds an incorrect configuration, it can provide targeted feedback on which parameters failed (e.g., missing routes, wrong IP addresses), and students can ask clarifying questions. This offers a level of explainability for configuration checks. In the oral exam, feedback is more limited: students receive pass/fail outcomes and general encouragement to consult the instructor if they disagree. There is no dedicated explanation mode that exposes the internal reasoning or shows step-by-step evaluation criteria. Explanations are thus available in a task-specific but somewhat ad hoc way, resulting in a partially fulfilled assessment.
T.E.-3	Documentation of Decision Logic	2 - Partially Fulfilled	The evaluation algorithms are described in the research publication: the system compares student configurations against a known-good template and uses a fixed question bank with defined correct answers. The study also discusses error sources such as speech recognition and connectivity issues. However, this documentation targets an academic audience and is not formalised as an internal governance document describing assumptions, limitations, and risk trade-offs for institutional use. As a result, decision logic is documented at a technical level but not fully integrated into organisational documentation, yielding a partial implementation.
T.E.-4	Role Transparency in Educational Tasks	3 - Fully Implemented	Students are informed that the robot is part of a pilot exploring automated evaluation and fairness, and that final grading decisions remain with the human instructor. The robot's role is to provide preliminary evaluations, immediate feedback, and a consistent procedure across students, not to replace the teacher's judgment. This role is clearly communicated and actively discussed in the study. Accordingly, role transparency in educational tasks is fully implemented.

Table 59. Case Study (A.L.) Evaluation

Control (A.L.)	Requirement	Assessment	Observations
A.L.-1	Assigned Human Supervisor	3 - Fully Implemented	A human instructor is responsible for the course and is present during the evaluation sessions. The lecturer retains authority over final assessment decisions and can override the robot's outputs in case of errors or technical problems. The study explicitly notes that the robot's evaluations do not replace human judgment and that problematic cases trigger manual review. This clear assignment of supervisory responsibility fully satisfies the requirement.
A.L.-2	Defined Student Appeal Pathways	3 - Fully Implemented	Because the robot participates directly in evaluative tasks, the deployment incorporates an explicit appeal mechanism: if students disagree with the robot's assessment or face technical issues (e.g., misheard answers), they can request review by the instructor, who verifies both the configuration and oral responses. Negative or ambiguous outcomes are thus systematically double-checked by a human. This clear and accessible appeal pathway, communicated to students as part of the pilot, is well aligned with the guideline and is therefore fully implemented.
A.L.-3	Source Attribution Logging	2 - Partially Fulfilled	Detailed evaluation logs (containing configuration states, comparison results, and oral exam transcripts) allow the instructor and researchers to reconstruct how specific decisions were reached. This supports investigation of disputed cases and aggregate analysis of fairness. However, logs are not integrated into a formal case management tool, and there is no dedicated identifier that links a student's informal complaint to a specific logged evaluation event beyond the course's internal record-keeping. Traceability is thus strong at a technical level but only partially integrated into accountability workflows, yielding a partial score.
A.L.-4	Institutional Responsibility Policies	1 - Minimally Addressed	Responsibility for the pilot and its outcomes lies primarily with the research team and the course instructor, under general institutional rules for teaching and research. There are no specific policies that define institutional liability if the robot's evaluation contributed to unfair treatment or perceived discrimination. While the presence of robust human oversight lowers the immediate risk, a production deployment would require clearer allocation of responsibility across the institution. For now, this control is minimally addressed.

Table 60. Case Study (H.O.) Evaluation

Control (H.O.)	Requirement	Assessment	Observations
H.O.-1	Human Oversight for Critical Decisions	3 - Fully Implemented	The robot's evaluations are not treated as final, binding decisions. Instead, they serve as preliminary automated checks that are always subject to human validation, particularly for negative outcomes. The instructor reviews cases where students fail the robot evaluation to ensure that technical errors (such as speech misrecognition or connectivity problems) do not unfairly penalise them. This strong "human-in-the-loop" setup ensures that critical decisions about student progression and grading remain under human control, fully satisfying the oversight requirement.
H.O.-2	Manual Override and Intervention Tools	2 - Partially Fulfilled	Instructors can intervene at any time by stopping the robot, switching a student to manual evaluation, or overriding results in course records. These interventions are executed via normal pedagogical tools (e.g., adjusting the gradebook) rather than through a dedicated supervisor interface for the robot. There is no special GUI for pausing automated evaluation or for annotating cases that require manual review, beyond informal practices. As a result, manual override capabilities are robust in practice but not explicitly instrumented in the system, yielding a partial implementation.
H.O.-3	Ongoing Monitoring and Feedback	2 - Partially Fulfilled	The pilot includes structured data collection on system performance, types of errors, and students' perceptions of fairness and trust. This provides valuable feedback for refining the evaluation logic and interaction design. However, there is no continuous monitoring dashboard or routine operational review process beyond the research study itself. Once the study ends, there is no guarantee that such monitoring will persist. This means the control is partially fulfilled: the mechanisms exist in the pilot context but are not yet embedded as ongoing practice.

Table 61. Case Study (B.F.) Evaluation

Control (B.F.)	Requirement	Assessment	Observations
B.F.-1	Use-Case Specific Fairness Assessment	3 - Fully Implemented	Fairness is the core focus of this case study: the research explicitly asks whether a robot can be a "fair teacher" and examines how automated evaluation compares to human grading. The deployment measures the alignment between robot and instructor decisions, reports on technical error sources, and collects students' perceptions of fairness, trust, and acceptance. This constitutes a thorough use-case-specific fairness assessment, far beyond typical pilots, and fully satisfies this control.
B.F.-2	Performance Disparity Audits	1 - Minimally Addressed	Despite the strong focus on fairness, the analyses primarily focus on overall agreement rates and error types rather than on systematic disparities between student subgroups (e.g., by gender, language background, disability status). The study acknowledges that speech recognition errors and lab noise can affect some students more than others, but it does not stratify metrics to quantify these differences. As a result, disparity auditing is recognised conceptually but not operationalised, leading to a minimal implementation.
B.F.-3	Inclusive Dataset and Language Use	2 - Partially Fulfilled	The content evaluated (network configurations, technical concepts) comes from course materials designed by instructors and is unlikely to contain overtly biased or discriminatory language. Oral questions focus on factual knowledge rather than sensitive personal topics. However, the underlying language technologies for speech recognition and any AI components may carry general biases from their training data, and there is no evidence of targeted testing for inclusivity or of adapting content for different linguistic backgrounds. Thus, the dataset and language use are relatively neutral within the narrow domain but not explicitly optimised for inclusion, yielding a partial rating.
B.F.-4	Inclusive Interaction Design	1 - Minimally Addressed	The evaluation workflow assumes that all students can interact comfortably via speech in a shared lab environment and can interpret feedback delivered by voice and screen. There are limited accommodations for students with speech, hearing, or anxiety-related challenges; no alternative text-only or remote interface is available during the pilot. While students can opt for human evaluation instead, which mitigates exclusion at the course level, the robot interface itself is not designed with inclusive interaction options in mind. This control is therefore minimally addressed.

Table 62. Case Study (E.A.) Evaluation

Control (E.A.)	Requirement	Assessment	Observations
E.A.-1	Multimodal Interaction and Accessibility	1 - Minimally Addressed	The robot delivers feedback both verbally and via text on its screen, offering a basic multimodal output channel. However, input is limited to spoken interaction and physical connection of devices, and the system does not integrate assistive technologies such as screen readers, alternative input devices, or high-contrast visual modes. Students with difficulties speaking, hearing, or standing near the robot may therefore find the interface less accessible. The control is minimally met through limited multimodality.
E.A.-2	Accessibility Testing and Co-Design	0 - Not Addressed	The case study does not report any targeted accessibility testing or involvement of students with disabilities in the design and evaluation of the robot. Usability and fairness are studied primarily from the perspective of the general student population. Formal accessibility co-design or audits are absent, so this control is not implemented.
E.A.-3	Economic and Resource Accessibility	1 - Minimally Addressed	The evaluation robot is available to all students attending the relevant lab sessions without additional cost or need for personal devices, which supports economic accessibility within that cohort. However, students who cannot attend in person, or who are assigned to different lab groups without the robot, do not benefit from the automated support. There is no remote or asynchronous equivalent of the evaluation workflow. As a result, this control is only minimally addressed.
E.A.-4	Institutional Monitoring for Disparities	1 - Minimally Addressed	The research examines overall fairness and student attitudes, but the university does not yet monitor robot-mediated evaluation as part of a broader equity framework. There is no institutional process to detect whether certain groups use the robot less, experience more errors, or feel less comfortable with automated evaluation. Still, the fairness-focused research provides a starting point that could be expanded into institutional monitoring. For now, this control is minimally implemented.

Control Area Scoring Summary

The following table summarizes the control evaluation results across all cybersecurity and ethical domains for the robot assistant that performs automated task evaluation and oral examination. Each control domain contains multiple individual controls evaluated against this deployment. Scores reflect the average degree of fulfillment on a 0–3 scale, and the final column shows the percentage of controls in that domain which meet at least a partial threshold (score of 2 or higher).

Table 63. Detailed Control Scoring per Domain (Automated Evaluation Assistant)

Control Domain	**1**	**2**	**3**	**4**	**Avg. Score**	**% $\geq$ 2**
Data Privacy & Confidentiality (D.P.)	2	3	2	2	2.25	100%
System & Data Integrity (S.I.)	2	1	1	1	1.25	25%
Information Leakage & Misuse (I.L.)	1	1	2	1	1.25	25%
System Access & Control (S.A.)	1	0	0	0	0.25	0%
Vulnerability & Patch Management (V.P.)	1	0	1	2	1.00	25%
Cyber-Physical Safety (C.P.)	2	2	2	1	1.75	75%
Network & Communication Security (N.C.)	2	1	1	0	1.00	25%
Institutional Readiness & Governance (I.R.)	1	0	1	–	0.67	0%
Privacy, Consent (P.C.)	3	3	1	1	2.00	50%
Transparency & Explainability (T.E.)	3	2	2	3	2.50	100%
Accountability & Liability (A.L.)	3	3	2	1	2.25	75%
Autonomy & Human Oversight (H.O.)	3	2	2	–	2.33	100%
Bias, Fairness & Inclusion (B.F.)	3	1	2	1	1.75	50%
Equity & Accessibility (E.A.)	1	0	1	1	0.75	0%
Overall					**1.50**	**46.3%**