AI and Ethics (2026) 6:91
https://doi.org/10.1007/s43681-025-00935-x

RESEARCH

From AI security to ethical AI security: a comparative risk-mitigation framework for classical and hybrid AI governance

Ludovica Ilari[1] · **Simona Tiribelli**[2] · **Filippo Caruso**[1]

Received: 26 May 2025 / Accepted: 27 November 2025 / Published online: 8 January 2026

Abstract
As Artificial Intelligence (AI) systems evolve from classical to hybrid classical-quantum architectures, traditional notions of security—mainly centered on technical robustness—are no longer sufficient. This study aims to provide an integrated security ethics compliance framework that bridges technical and ethical dimensions across the AI lifecycle. By adopting a security ethics-by-design approach, the framework introduces mitigation measures in relation to key ethical principles capable of addressing emerging risks and considering AI governance needs in the initial AI design and development phases. This study proposes a novel framework, currently absent from the literature, to address security ethics challenges in both classical and hybrid systems. Key contributions include the integration of post-quantum and quantum cryptography, particularly homomorphic encryption, to ensure long-term privacy and security in hybrid AI. The framework also includes bias testing and explainable AI techniques to promote fairness and explainability, and to prevent safety-related vulnerabilities—such as algorithmic bias—from serving as vectors for malicious, discriminatory attacks. Ultimately, it provides a preliminary roadmap for embedding ethical security considerations throughout the lifecycle of classical and hybrid AI systems.

Keywords Hybrid classical-quantum machine learning · Security and ethics by design · AI ethics and governance · Post-quantum and quantum cryptography · Homomorphic encryption · Trustworthy AI

1 Introduction

The rapid evolution of Artificial Intelligence (AI) systems-from classical architectures to increasingly sophisticated hybrid classical–quantum models-is profoundly transforming data analysis, decision-making, and problem-solving across multiple domains [1–3]. Hybrid AI systems harness principles of quantum computing, such as superposition and entanglement, to offer unprecedented computational power for optimization, complex pattern recognition, and large-scale data processing [4, 5]. These advancements open new frontiers, but they also expose critical vulnerabilities and ethical challenges typical of quantum processing that traditional security frameworks are ill-equipped to address [6–8].

Traditionally, ethical AI security frameworks have primarily addressed security as a technical issue-such as protection against malware, data breaches, and adversarial attacks-as reflected in the Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the European Commission's High-Level Expert Group on AI (AI HLEG), and in the WHO's guidelines on the ethics and governance of artificial intelligence for health [9, 10]. However, the growing complexity and societal impact of security in AI systems demand a broader and more integrated approach-one that explicitly incorporates ethics.

In particular, the field of *security ethics* has emerged to address the specific ethical issues and challenges that arise from the specific practices and techniques of security more

✉ Ludovica Ilari
ludovica.ilari@unifi.it

✉ Simona Tiribelli
simona.tiribelli@unimc.it

✉ Filippo Caruso
filippo.caruso@unifi.it

[1] Department of Physics and Astronomy, University of Florence, Via Sansone 1, 50019 Sesto Fiorentino, Italy

[2] Department of Political Sciences, Communication, and International Relations, University of Macerata, Via Don Giovanni Minzoni 22/A, 62100 Macerata, Italy

broadly [11]. For example: Is it right to pay the ransom in a ransomware attack (where data are encrypted) in the hope of getting the data back in clear? Would this security practice of payment ethically legitimize and/or incentivize further attacks? Is it possible that attackers might publish the data infringing individuals' privacy? Is this risk ethically admissible? And if they only publish the data of certain individuals, wouldn't there be a high risk of discrimination? These are some of the questions security ethics deal with and it is in this specific field we position our research contribution, which entails adopting an ethical perspective on security practices. What is missing in the literature is an explicit ethical foundation for AI security: current literature lacks a framework that integrates technical safeguards with ethical reasoning throughout the AI system lifecycle. These concerns are pressing in practice. Recent cases-such as deepfakes, AI-driven disinformation, and biased predictive analytics-demonstrate that security failures in AI systems can threaten democratic processes, individual freedoms, and public trust [12, 13]. In healthcare, adversarial attacks on AI diagnostic tools can lead to harmful misdiagnoses [14], while biased prediction systems may leak sensitive data or exacerbate health disparities [15, 16]. This dual focus on compliance and moral reflection represents a distinctive contribution to the field of AI and cybersecurity ethics.

This paper addresses that gap by proposing a comprehensive security ethics framework tailored to both classical and hybrid classical-quantum AI.

In doing so, we also pursue a novel constructive approach that differs from existing literature.

- It aligns with the well-known "ethics by design" approach [17–22], embedding ethical principles from the earliest stages of system design. This approach is lacking in security ethics.
- It supports legal, ethical, and technical compliance by aligning mitigation measures and ethical principles with existing regulatory and normative frameworks. In this regard, it offers a novel contribution to the field of security ethics by providing an integrated and overarching structure that helps avoid the duplication of policy efforts.
- Furthermore, unlike previous approaches to security ethics [11]—which typically follow either top-down (from moral theory to practice) [23–26], bottom-up (from cases to principles) [27–29], or pragmatic (based on expert consensus) methods [30–33]—our approach introduces a hybrid direction. We begin by examining the expert regulatory context or pragmatic approach-focusing on established regulations and mitigation measures—and proceed with a top-down ethical analysis to identify the underlying moral principles rooted in bioethics and technology ethics that justify the use of such techniques in relation to specific ethical risks and challenges. This structure partially mirrors the logic of the German Standard Data Protection Model, developed by a subgroup of the Data Protection Conference (DSK), which translates the legal requirements of the General Data Protection Regulation (GDPR) into concrete technical and organizational measures. These take the form of seven Data Protection Goals: data minimization, availability, integrity, confidentiality, unlinkability, transparency, and intervenability.

Moreover, we adopt a broad view of security that integrates both *safety* (accidental errors) and *security* (malicious attacks), recognizing that these dimensions often overlap. This is particularly evident in scenarios involving malicious attacks that exploit accidental errors such as bias to their advantage, as in the case of adversarial attacks. For instance, facial recognition systems often perform worse on darker-skinned individuals due to biased training data [34, 35]. Adversarial attackers can exploit this by using patterns or makeup to evade detection, taking advantage of both technical flaws and existing bias. To sum up, we refer to "security ethics" to indicate a range of key ethical principles that should be implemented to ensure both the security of AI systems and the ethicality of security practices and techniques themselves in security at risk scenarios.

Overall, this framework is intended to guide stakeholders-policymakers, researchers, and industry leaders-in the ethical and secure governance of AI technologies [36, 37]. In this regard, this research offers practical guidelines to ensure that advances in AI align with ethical standards and societal expectations. In doing so, it promotes the trustworthiness of classical and hybrid AI systems, particularly as quantum-enhanced technologies become increasingly influential in critical sectors such as healthcare, finance, and national security [38–41].

Thus, this paper addresses an emerging research field that has gained increasing relevance in recent years, offering a novel contribution through the development of a technical and ethical security framework for classical and hybrid AI currently missing in the scholarship on ethics of AI and quantum AI.

This paper introduces the first ethical security framework for both classical and hybrid AI, aimed at paving the way for future efforts to support stakeholders (academia, industry, and institutions) in governing the entire AI lifecycle in compliance with current regulations, from both technical and ethical perspectives.

1.1 Related works

AI security has traditionally been framed in terms of technical robustness, as outlined in the Ethics Guidelines for Trustworthy AI (2019) by the AI HLEG [17]. These guidelines define security through the concept of resilience to attack and security, emphasizing the protection of AI systems from vulnerabilities that could expose them to cyber threats. The ALTAI framework [9], developed by HLEG, adopts a risk-preventive approach and emphasizes the importance of implementing mitigation measures to safeguard AI models, data, and infrastructures. Key threats-such as data poisoning, model evasion, and model inversion-require proactive countermeasures, including penetration testing, red teaming, and continuous security updates.

Beyond technical security, the guidelines also emphasize privacy and data governance, which are addressed through regulatory frameworks such as the General Data Protection Regulation (GDPR) [42]. These regulations establish strict rules on data protection, ensuring that AI systems process information ethically and transparently. However, while existing regulations provide a strong foundation for data protection and AI security, they do not address the broader concept of security ethics, which involves evaluating the ethical implications of security measures themselves-such as their impact on transparency, individual freedoms, and governance.

Another approach to AI security is presented by Brundage et al. [6], which categorizes AI security risks into digital, physical, and political security. Digital security focuses on AI-driven cyber threats, such as automated hacking and adversarial attacks. Physical security pertains to AI-enabled real-world threats, including autonomous drones and cyber-physical sabotage. Political security concerns the use of AI for information manipulation, mass surveillance, and targeted propaganda. While this framework helps in understanding AI-related security risks, its primary focus remains on threat prevention and mitigation, rather than exploring the ethical dimension of security strategies.

A further gap in the literature is the lack of specific frameworks for security in hybrid AI systems. Some studies, such as those by Possati [43], examine the ethics of post-quantum cryptography (PQC), addressing dilemmas arising from transitioning to quantum-secure architectures. Current security strategies against quantum threats include reinforcing classical cryptographic protocols with post-quantum algorithms and migrating entirely to quantum-based security models. However, both approaches raise ethical concerns, such as inequality in digital security access, increased governmental control over encrypted communications, and new risks of surveillance and exploitation. While these studies highlight the intersection of security and ethics, they remain focused on technical and regulatory aspects, lacking a structured ethical framework for AI and hybrid AI security.

The literature outlines three main approaches to security ethics [11]: bottom-up [27], pragmatist [30], and top-down [23, 25]. Bottom-up methods derive insights from case studies but often lack generalizability. Pragmatist models emphasize operational principles such as confidentiality, integrity, and availability (CIA), yet tend to overlook broader ethical concerns. Top-down approaches offer strong normative foundations but face challenges in translating abstract values into actionable security policies. Among the leading proponents of the top-down approach are Formosa et al. [23], who adopt a principle-based ethical framework incorporating five key principles from the AI4People framework: beneficence, non-maleficence, autonomy, justice, and explicability. These principles are then applied in various cybersecurity scenarios, such as ransomware and denial-of-service attacks. While such frameworks provide valuable ethical insights, they are not specifically designed to address the distinctive security challenges posed by AI systems and hybrid AI models.

Although each approach offers important perspectives, none is sufficient in isolation to capture the full complexity of contemporary security. A unified ethical framework that integrates these viewpoints may offer a more comprehensive and actionable foundation for guiding ethical decision-making in security-critical AI systems.

The literature reveals multiple gaps. While classical AI guidelines provide high-level recommendations, and some ethical frameworks include a measurement phase, neither clearly identifies ethical principles related to security, nor is there a framework addressing hybrid AI (Table 1).

In the domain of hybrid AI, some ethical considerations are beginning to emerge in the context of PQC and quantum cryptography implementations. However, these are typically limited in scope and remain without technical suggestion. Among existing contributions, only Formosa et al. propose a framework for the ethics of cybersecurity applied across various cybersecurity scenarios. Yet, this framework is not specifically tailored to the AI context, let alone to the hybrid context [8]. From this analysis, it becomes evident that, although several frameworks address AI security and some studies explore the challenges of hybrid AI, there is not a structured framework dedicated to the ethics of AI security-nor for hybrid AI security. This paper aims to bridge that gap by proposing an integrated framework that combines technical security measures with ethical considerations, applicable to both classical and hybrid AI systems.

Table 1 Overview of AI security and ethics frameworks and their limitations

Contribution	Type	Principles	Context	Limitations
Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) [17]	Guidelines	Resilience to attacks; Privacy and data governance	AI	Lack of ethical principles in security (to address ethical challenges posed by security techniques); no guidelines for hybrid AI; no ethical risk metrics
ALTAI (AI HLEG, 2019) [9]	Framework	Resilience to attacks; Privacy and data governance	AI	No ethical grounding in security; lacks framework for hybrid AI
Avin et al. (2018) [6]	Framework	Digital, physical, and political security	AI	No explicit ethics in security; no risk metrics; no hybrid AI model
Possati et al. (2023) [43]	Framework	Inequality in digital security; Encrypted communication control; Surveillance risk	Quantum and post-quantum	No structured ethics for hybrid AI; no classical AI extension; no risk metrics
Formosa et al. (2021) [23]	Framework	Beneficence; Autonomy; Non-maleficence; Justice; Explicability	General	Lack of an ethics-by-design framework for classical and hybrid AI

1.2 Paper organization

This paper is organized as follows. In Sect. 2, we describe the methods employed and the sources used to construct the framework. In Sect. 3, we present the proposed framework and outline the key differences between AI security ethics in classical and hybrid systems, which are further discussed in Sect. 4. In Sect. 5, we discuss limitations and directions for future work, while in Sect. 6, we provide practical recommendations for implementation by organizations, policymakers, and regulators. Finally, in Sect. 7, we present our conclusions. In Appendix A, we provide supporting tables for both the general and the tailored frameworks in classical and quantum AI.

2 Methods

In developing a security ethics framework for both classical and hybrid AI systems, we have deliberately chosen to transition from a purely technical mitigation approach to an ethical perspective. Unlike traditional frameworks-such as ALTAI [9], the NIST AI Risk Management Framework [44], or ISO/IEC 42001 [45]—which incorporate organizational and human-centric mitigation strategies from the outset (e.g., governance structures, audit mechanisms, or access policies on the organizational side; and user training, human-in-the-loop oversight, or risk awareness programs on the human side), we have opted to initially exclude these aspects. Instead, we focus on establishing a technical-ethical foundation, which will later be expanded to include organizational and human factors, ensuring a more comprehensive understanding of AI security vulnerabilities.

In Fig. 1, we present a model that outlines the steps leading to the development of the classical framework, which subsequently served as the foundation for constructing the hybrid framework. This framework was developed in response to the limitations identified in existing security ethics approaches [11], emphasizing the need for a hybrid method that draws on top-down ethical theory, bottom-up case analysis, or pragmatic, consensus-based strategies.

Our approach is structured around the lifecycle of an AI system, following the principles of security by design. This means that rather than treating security as an afterthought, we integrate mitigation measures at each stage of the AI lifecycle.

To construct this framework, we first identified the key phases in the AI lifecycle, including data collection and preparation, training, validation and testing, deployment, inference and operations, maintenance and updates, and decommissioning. For each of these phases, we mapped out the relevant security mitigation measures from technical regulations and standards, adopting a pragmatic approach.

Furthermore, to identify the ethical principles associated with the proposed measures, we draw on bioethics and the ethics of technology, adopting a top-down ethical approach.

Our framework positions itself within a hybrid approach, aiming to integrate the strengths of both pragmatist and top-down approaches, while also considering the future inclusion of a bottom-up approach grounded in concrete cases derived from potential attack scenarios (the cyber-attack domain).

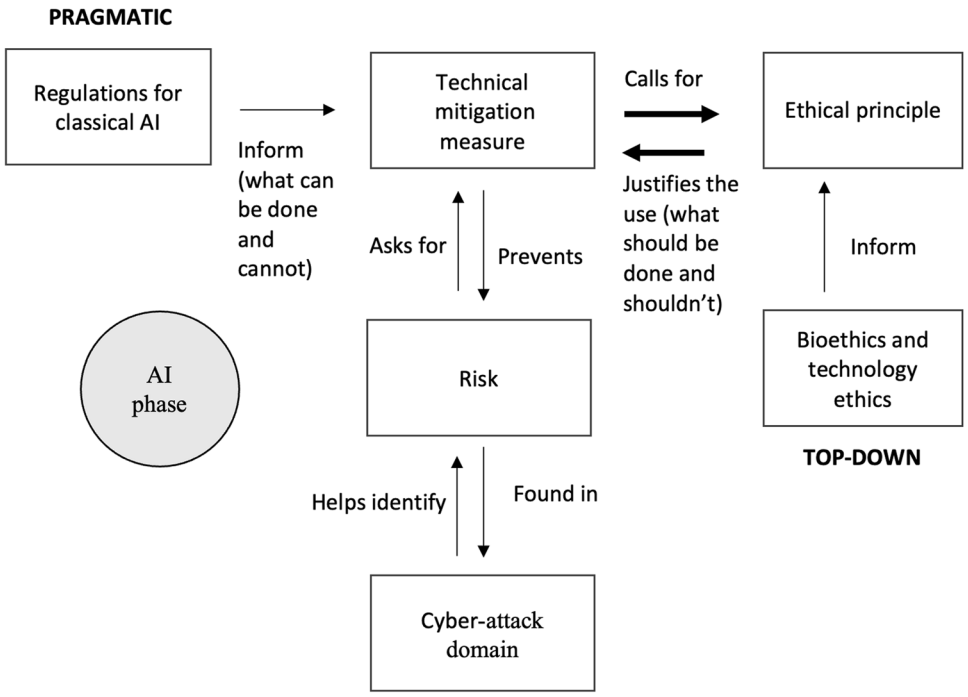

Fig. 1 Framework workflow

In this regard, this framework is developed in response to the limitations identified by [11] emphasizing the need for an hybrid approach that considers top-down ethical theory, bottom-up case analysis, and pragmatic consensus-based methods. Top-down approaches provide valuable normative direction but frequently struggle to translate abstract ethical principles into concrete, actionable security measures-especially in complex and fast-evolving domains like healthcare [23, 25]. Bottom-up approaches, although rooted in real-world practice, often lack generalizability and may be ill-equipped to anticipate emerging threats [27]. Pragmatic approaches, on the other hand, offer strong technical grounding but tend to overlook broader ethical concerns [30].

A fundamental distinction in our method lies in its departure from conventional risk assessment models, which typically follow a linear process: identifying risks, evaluating risks, and then applying technical and organizational countermeasures. Instead, our approach reverses this sequence. We begin by analyzing existing mitigation strategies and, from there, bridge the underlying ethical principles (Fig. 1).

Applying this method consistently across both classical and hybrid AI systems enables a direct comparative analysis. The same security measures and ethical principles are assessed within both paradigms, providing insight into how hybrid AI systems introduce new risks and challenges that may not be as prevalent in classical AI. While the overall structure remains the same, certain aspects require specific attention, such as the different nature of AI lifecycle phases in hybrid models (explored in Sect. 2.1), the role of cryptographic techniques in securing these systems (detailed in Sects. 2.5–2.7), and the unique security risks and attacks introduced by quantum computing (examined in Sects. 2.3, 2.4) with respect to classical risks (Sects. 2.2, 2.4). Given the interdisciplinary nature of this research, our framework is informed by an extensive review of regulatory frameworks, cybersecurity guidelines, and ethical AI governance models. Key sources include the GDPR, the AI Act [46], the Cybersecurity Act [30], the NIS2 Directive [47], and the Digital Operational Resilience Act (DORA) [48], alongside cybersecurity standards from NIST [19, 44, 49], and ISO/IEC [45, 50–57] (Fig. 1).

Additionally, we incorporate insights from academic literature on AI security, ethical governance, and emerging threats posed by quantum computing. As quantum computing and hybrid AI technologies continue to evolve, ongoing adaptation of this framework will be necessary to address new security threats and ethical dilemmas that emerge over time.

While this study is fundamentally theoretical, it establishes a practical foundation for future empirical research. Specifically, we acknowledge that real-world case studies and experimental validation are essential to fully assess the effectiveness of our proposed framework. Future research should integrate these elements to test the applicability of ethical security measures in actual AI governance.

At this stage, we propose a hypothetical scenario to illustrate how the framework could be applied to Federated Quantum Machine Learning in the context of rare disease diagnostics [58, 59].

A network of hospitals adopts a Federated Quantum Machine Learning (FQML) framework to develop advanced predictive models while keeping clinical data on-premises. Each hospital locally trains a parameterized quantum circuit on sensitive medical records, sharing only encrypted model parameters via TLS 1.3 with AES-256 and a hybrid X25519+Kyber handshake. The models undergo resilience testing against data poisoning and are continuously monitored for bias and fairness. During inference, circuits are executed on a quantum processing unit (QPU)-either local or cloud-based-and the resulting predictions (class and probability) are digitally signed using Dilithium [60] and encrypted with AES-256. An Explainable AI module, based on feature importance techniques, supports clinicians in interpreting the model's outputs. The system includes automated procedures for verifying cryptographic dependencies and securely retiring obsolete keys and parameters during maintenance and decommissioning phases. Governance is managed by an ethical-technical oversight committee, which evaluates the framework's implementation and its impact on privacy, fairness, and explainability. Each operational cycle produces non-compliance reports that inform iterative reviews and model updates, ensuring alignment with evolving regulatory and technological advancements in quantum computing.

Together with this hypothetical scenario, we outline possible techniques for validating the framework.

To validate the framework, we propose the following technical contributions, organized as a three-level approach:

1. Controlled simulations in a laboratory environment using synthetic healthcare datasets, measuring both technical metrics (e.g., intrusion detection time, false positive rate) and ethical indicators (e.g., degree of privacy preservation, model fairness across demographic groups). This phase completes the framework construction by assessing security and ethical risks.
2. Empirical benchmarking by comparing our solutions with established datasets and benchmarks-such as FedML Bench and FedSecurity [61]—evaluating practical performance (accuracy, latency, scalability) alongside ethical compliance (audit-trail completeness, level of transparency).
3. Pilot studies in real-world settings, in collaboration with healthcare facilities, to test protocol implementation and gather direct feedback from clinicians, administrators, and patients.

2.1 Hybrid classical-quantum artificial intelligence: integrating classical and quantum approaches

In hybrid classical-quantum machine learning (ML), data processing occurs across both classical and quantum domains, combining the strengths of traditional data handling methods with the advanced computational capabilities offered by quantum mechanics. In these classical-quantum hybrid systems, the data collected from real-world sources is inherently classical, meaning it exists in conventional digital formats. However, there are also emerging quantum technologies, such as quantum sensing, that collect intrinsically quantum data. In such cases, the configuration of hybrid machine learning could be quantum-classical instead of classical-quantum. These classical data undergo preliminary steps such as data cleaning, preparation, feature selection or extraction, and are then encoded into quantum-compatible representations through quantum encoding techniques (Fig. 2). In this phase of classical pre-processing, the goal is to transform traditional datasets into quantum state vectors represented by qubits, effectively bridging the gap between the classical and quantum computational paradigms. However, this step-commonly referred to as data encoding or quantum feature mapping-often represents a significant bottleneck. In many cases, the encoding process is computationally costly and not always efficient, thereby limiting the overall performance and scalability of hybrid classical-quantum machine learning pipelines.

Once classical data have been mapped onto quantum states, the system enters the quantum processing phase, which involves executing a quantum algorithm tailored to the encoded representation (Fig. 2). This transition is non-trivial and depends on both the encoding strategy and the capabilities of the quantum hardware. Here, the quantum-encoded data are processed using quantum circuits specifically designed to exploit the unique properties of quantum mechanics, such as superposition and entanglement. Quantum algorithms are executed within these circuits, enabling computations that are potentially faster and more efficient for certain complex tasks compared to their classical counterparts, often using significantly fewer parameters. This quantum phase forms the core computational advantage of hybrid classical-quantum ML systems, providing capabilities particularly suitable for solving problems in optimization, pattern recognition, and complex data analysis.

After completing quantum processing, the system transitions back into the classical realm through a measurement process, initiating the post-processing phase (Fig. 2). Quantum states, which exist in probabilistic superpositions during computation, collapse into definite classical outcomes upon measurement. These measurement results are then interpreted and analyzed using classical methods. This phase

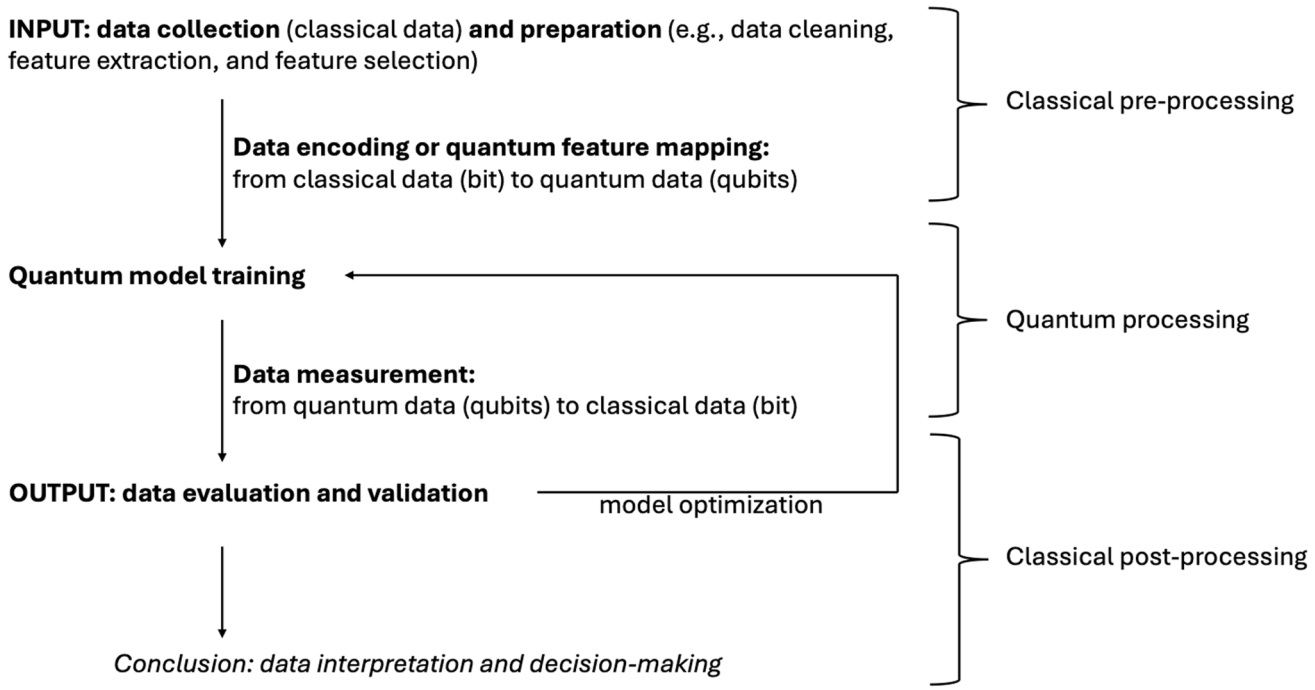

Fig. 2 Hybrid classical-quantum AI workflow

involves evaluating and validating the quantum-derived results, refining or optimizing the learning model, and ultimately drawing conclusions from the quantum-enhanced analysis. Thus, the hybrid system leverages quantum computational power for processing while relying on classical techniques for final interpretation and decision-making, effectively integrating the best of both paradigms.

Having outlined the technical aspects of hybrid classical-quantum ML systems, it becomes essential to recognize the novel cybersecurity risks these innovations introduce. Quantum ML (QML) systems face unique vulnerabilities compared to purely classical systems, raising new questions about their exposure to potential cyber-attacks (Sect. 2.3). It is therefore essential to examine not only how traditional cyber threats might adapt or evolve in quantum contexts but also to anticipate entirely new types of attacks that exploit the quantum properties of these systems. Future protection techniques—including PQC, designed to withstand quantum-enabled attacks, and quantum cryptography, such as quantum key distribution (QKD), which leverages quantum mechanics for secure communication—must be integrated into these systems to ensure robust cybersecurity defenses.

As we envision developing and deploying a hybrid classical-quantum ML system, we must critically assess the potential cyber threats that could affect its operations, considering how these threats may differ fundamentally from those known in classical computing environments. Moreover, ethical frameworks established for cybersecurity in classical contexts must be reevaluated and possibly reshaped to address the ethical implications arising from quantum-enabled capabilities. This evaluation is necessary because quantum computing could disrupt the balance of privacy, transparency, and fairness traditionally maintained in cybersecurity policies. For instance, quantum algorithms may compromise privacy by breaking widely adopted encryption schemes, exposing sensitive data previously considered secure. Transparency, typically ensured through logging mechanisms and Explainable AI (XAI) techniques, is challenged by the inherent complexity and opacity of quantum computations, which hinder traceability and auditability. Fairness, often maintained through classical bias mitigation tools, may also be at risk, as quantum data encoding and optimization processes can introduce new, less detectable forms of algorithmic bias that existing governance mechanisms are not yet equipped to address. Thus, there is an urgent need to evolve existing ethical frameworks for cybersecurity to encompass the novel risks and protective measures associated with hybrid classical-quantum ML systems, ensuring they uphold fundamental ethical principles while fostering responsible and secure technological development.

2.2 Cyber risks, attacks and mitigation measures in classical artificial intelligence

Securing hybrid classical-quantum ML systems involves addressing a spectrum of cyber threats, both classical and quantum-specific, making the cybersecurity landscape

notably more complex. Before delving into quantum-specific vulnerabilities, it is relevant to provide an overview of traditional cyber risks typically encountered in classical AI algorithms and classical data contexts. These foundational risks persist even in hybrid ML environments, serving as baseline threats that must be understood and mitigated to ensure the overall security of advanced AI systems.

Among the common classical cyber threats are malware attacks, referring to any malicious software designed to harm or exploit systems against users' interests. Malware can range from spyware designed to extract sensitive data stealthily, to destructive viruses that compromise system functionality. Similarly, man-in-the-middle (MITM) attacks represent a significant threat, involving the interception or alteration of communications between two parties without their knowledge. MITM attacks are particularly insidious because they can persist undetected, silently compromising data integrity and confidentiality over extended periods.

Another prevalent risk is the denial of service (DoS) or the more potent variant known as distributed denial of service (DDoS) attacks. These attacks aim to overwhelm system resources, rendering services unavailable to legitimate users by flooding systems with excessive requests. Unlike traditional DoS attacks originating from a single source, DDoS attacks leverage multiple compromised devices-often forming a botnet-to generate overwhelming traffic, making them especially challenging to defend against due to their distributed nature.

Brute force attacks represent yet another classic cyber risk. These attacks systematically attempt all possible password combinations to gain unauthorized access to systems or services. Although simple in concept, brute force attacks can succeed due to weak or predictable passwords, especially in environments lacking robust authentication protocols.

Zero-day exploits present unique challenges because they target previously unknown software vulnerabilities, often leaving organizations defenseless until patches or mitigations become available. Such attacks exploit the window between the discovery of a vulnerability and its remediation, enabling attackers to cause substantial damage or exfiltrate sensitive data undetected.

Cyber attackers frequently exploit human factors through social engineering techniques, manipulating individuals' behaviors to gain access to confidential information or sensitive systems. These attacks often bypass traditional cybersecurity measures by tricking authorized personnel into inadvertently compromising security, highlighting the critical need for ongoing user awareness training alongside technical defenses.

Additionally, classical cyber threats include web application attacks, which exploit vulnerabilities in web services. These attacks can lead to unauthorized control of servers, data theft, and malware infections. With the proliferation of cloud-based services and online applications, web vulnerabilities represent an increasingly significant risk vector for classical AI systems that rely on web-based interfaces or services.

Within the realm of ML specifically, classical AI systems also face unique threats known as adversarial attacks [44]. Poisoning attacks, for instance, compromise the training data or the model itself during the training phase. Attackers insert malicious data points designed to manipulate or degrade the performance of ML models. Similarly, evasion attacks involve subtly altering input data during testing or inference to create adversarial examples that deceive models into making incorrect predictions. These adversarial manipulations are concerning because they can bypass conventional security measures designed for traditional software systems, exploiting vulnerabilities unique to ML algorithms.

Finally, privacy attacks represent another critical category of cyber risk targeting ML models. Examples include membership inference attacks, which aim to determine whether specific data points were included in a model's training set, and data reconstruction attacks, which attempt to recover sensitive data directly from model outputs. These privacy breaches pose severe ethical and regulatory concerns, particularly given the heightened sensitivity surrounding data privacy in regulatory frameworks like the GDPR.

To address these classical cybersecurity threats effectively, several regulatory frameworks and technical standards have been established. These include the GDPR, which sets stringent rules for data privacy and protection-particularly effective in countering social engineering and malware attacks through requirements for access control, encryption (e.g., TLS, AES-256), and data minimization. The AI Act provides regulatory standards for AI applications and promotes resilience against adversarial attacks and bias exploitation through adversarial training, bias testing, XAI, and robustness evaluations. The Cybersecurity Act introduces an EU-wide certification framework for digital products, services, and processes, supporting protections against MITM and zero-day exploits via secure protocol enforcement, patching policies, and certified cryptographic solutions. Additionally, the NIS2 Directive aims to secure critical infrastructure sectors, including energy, healthcare, and transport, by mandating robust cybersecurity measures that help mitigate DoS and brute force attacks through requirements such as anomaly detection, system redundancy, and continuous monitoring. Internationally recognized standards, including NIST SP 800-53 [62] and ISO/IEC 27001/27002 [51, 52], offer operational guidance on implementing and managing such security controls,

including measures like endpoint protection, intrusion detection, and vulnerability management. These regulatory frameworks and technical standards collectively establish a comprehensive defense posture and form the backbone of cybersecurity practices, ensuring that classical ML systems maintain high standards of security, privacy, and ethical accountability. By outlining classical cyber risks, attacks, and corresponding mitigation measures embedded within regulatory frameworks and technical standards, we establish a foundation for understanding how traditional threats continue to impact hybrid classical-quantum ML systems. Recognizing these baseline vulnerabilities is essential as we explore new quantum-specific threats, requiring advanced, innovative approaches to cybersecurity in the rapidly evolving landscape of hybrid classical-quantum ML.

2.3 Cyber risks, attacks and mitigation measures in quantum artificial intelligence

In QML, unique cybersecurity risks arise due to the distinct nature of quantum computation.

Figure 3 illustrates quantum threats to classical models (including man-in-the-middle, brute-force, web application, and privacy attacks), as well as vulnerabilities specific to quantum-native models models-such as noise, fault injection, scaling pitfalls, compilation processes, and privacy concerns-which we analyze below.

In addition, QML introduces vulnerabilities that require novel protective measures. These are shown in Fig. 3 and examined in this subsection.

Recent studies have highlighted various security concerns, shedding light on emerging attack vectors unique to quantum environments. One of these studies, by Kundu et al. [7], explores several potential vulnerabilities in QML, focusing primarily on hardware, compilation processes, and fault injection attacks.

A key issue identified is the identical coupling hardware problem. Major quantum cloud providers like IBM, QuEra, Rigetti, D-Wave, and IonQ utilize various hardware platforms of differing quality. Since users often cannot differentiate between coupling maps, they may unknowingly execute algorithms on subpar hardware, adversely affecting QML performance. This risk underscores the need for mechanisms to verify hardware quality, such as the proposed Quantum Physically Unclonable Function (QuPUF), which authenticates hardware identity prior to executing quantum programs [63].

Another critical concern involves the compilation process. Compilers convert high-level quantum programs into hardware-specific quantum gate sets. Although reputable providers like IBM and Rigetti supply their own compilers, third-party compilers may offer performance optimizations at the expense of security. These external compilers carry risks such as intellectual property theft or malicious code insertion. The situation worsens when users embed

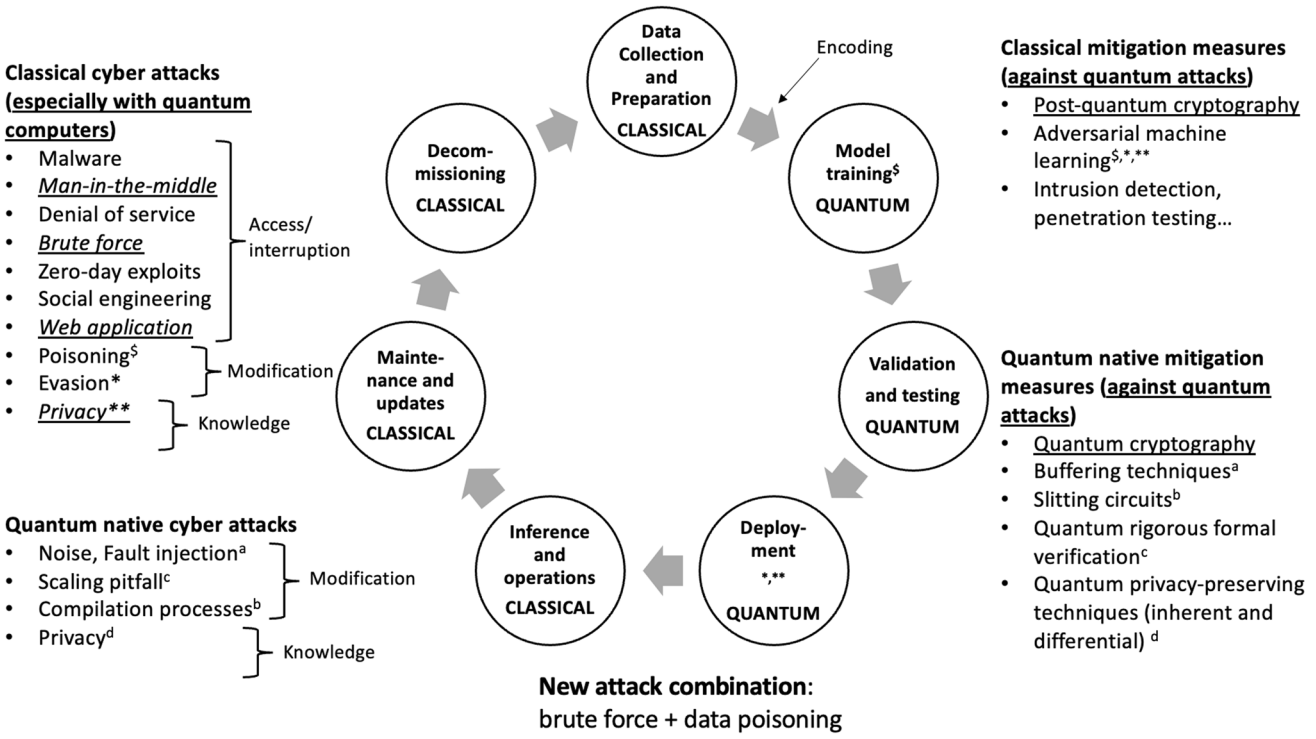

Fig. 3 Hybrid AI lifecycle with classical and quantum interactions, attack points, and mitigation mechanisms

sensitive information, like proprietary algorithms or confidential financial data, into quantum circuits sent to untrusted compilers. To mitigate these risks, a technique such as splitting quantum circuits and distributing them among multiple compilers has been suggested [64]. This measure aims to complicate reverse engineering attempts, thereby enhancing security. By fragmenting the quantum circuit across distinct compilation environments, it becomes significantly more difficult for an adversary to reconstruct the full architecture or logic of the quantum algorithm, even if partial access is gained. This approach introduces a form of security through distribution, limiting the exposure of sensitive quantum operations and protecting intellectual property.

Fault injection attacks represent another significant threat in QML environments. Quantum hardware often experiences crosstalk errors and is vulnerable to deliberate fault injection by external adversaries, leading to degraded model performance. Notably, in shared cloud-based quantum computing platforms, it has been shown that an attacker can indirectly affect another user's program by executing circuits that drive their allocated qubits-e.g., via repeated CNOT operations-thereby increasing error rates on adjacent qubits due to crosstalk. This can significantly reduce the prediction or classification accuracy of a quantum machine learning model. To address this, buffer or isolation qubits are employed to separate concurrent user programs running on the same QPU [65]. By preventing simultaneous operations on neighboring qubits, this approach has been shown to improve the system's reliability by up to 1.87×, albeit at the cost of reduced qubit availability [65]. These mitigation strategies collectively enhance the resilience of QML systems, protecting them from various forms of malicious interference.

Expanding on these concerns, Franco et al. [8] highlight further vulnerabilities and attack vectors that uniquely target QML models. Their study categorizes these threats into quantum-specific attack vectors and challenges related to scaling quantum systems.

For instance, fault injection attacks in the quantum context can take the form of "quantum Trojan" viruses. These viruses covertly insert additional quantum gates into Quantum Neural Networks (QNNs), embedding hidden backdoors that remain dormant until triggered, at which point they drastically alter model behavior. Such attacks exploit subtle differences between compiled and synthesized QNN circuits, making detection difficult.

Attackers may also leverage quantum hardware's intrinsic noise characteristics. In shared quantum computing environments, adversaries can intentionally induce hardware errors-such as crosstalk in superconducting systems or repeated shuttling operations in ion-trap architectures-not through physical access, but by executing carefully crafted quantum circuits. These software-level attacks exploit the hardware's physical behavior to degrade model performance or even lead to denial-of-service scenarios. Defenses against such threats include recognizing malicious circuit patterns or implementing buffering techniques that isolate user programs and limit quantum interference.

Another major concern is the scaling pitfall. As quantum systems grow larger and more intricate, they become exceedingly sensitive to even minor perturbations. This heightened sensitivity jeopardizes the reliability of quantum classifiers, requiring more substantial verification resources. Unfortunately, this additional overhead can negate some of the quantum computational advantages initially sought. Moreover, research indicates that adversarial perturbations become more potent in high-dimensional quantum systems, exacerbating security concerns as qubit counts increase.

Recognizing these novel risks, Franco et al. [8] also discuss defensive strategies categorized into three primary domains: adversarial training, data privacy, and formal verification. Adversarial training enhances model robustness by deliberately exposing quantum classifiers to adversarial inputs, familiarizing them with potential malicious manipulations. Traditional classical methods like the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD), and Momentum Iterative Method (MIM) have been adapted for quantum contexts. Encouragingly, studies show that quantum classifiers trained with adversarial methods can outperform classical counterparts in robustness, especially at larger scales. Recent innovations include quantum adversarial metric learning and random unitary encoding, which strengthen adversarial defenses further.

In parallel, ensuring data privacy has become paramount, given the rising frequency of data breaches. Two leading privacy-preserving strategies in QML are differential privacy and inherent privacy. Differential privacy introduces carefully calibrated noise into computational processes, protecting individual data points without significantly compromising model utility [66]. For example, research has shown that applying quantum rotation noise during training can notably improve model robustness by obscuring sensitive training [67, 68]. Inherent privacy, on the other hand, does not arise solely from the fundamental principles of quantum mechanics (such as superposition or the no-cloning theorem), but rather from the structural and computational complexity of overparameterized quantum models [69, 70]. Specifically, the vast solution space and intricate parameter interactions make it exceedingly difficult for an adversary to reconstruct sensitive inputs or gradients, even when partial information is exposed. For instance, in federated quantum learning settings, where many small quantum models are trained across distributed nodes, the inherent

overparameterization makes it extremely challenging for attackers to carry out gradient inversion attacks or reconstruct user-specific data from shared updates. This natural form of privacy is particularly valuable when explicit encryption mechanisms are impractical or would introduce prohibitive computational overhead.

Lastly, formal verification plays a fundamental role in ensuring that quantum models stay secure, even against unexpected or unknown attacks. This involves using rigorous mathematical methods to test and guarantee a model's robustness [71]. Examples include developing generative models that simulate possible adversarial attacks [72, 73], turning the challenge of checking model robustness into optimization problems solvable by quantum algorithms [71], or using mathematical techniques like Lipschitz continuity to estimate how sensitive a model is to small input changes [74]. Together, these formal tools offer practical and reliable ways to certify the security of quantum models.

Altogether, these defense strategies represent the foundation of a comprehensive framework to secure quantum machine learning systems against evolving adversarial threats. By integrating adversarial training, privacy-preserving techniques, and rigorous formal verification, the QML community aims to bolster model resilience, protect sensitive data, and safeguard the integrity of emerging quantum-enhanced applications. While conceptually promising, many of these approaches remain at a preliminary stage and are not yet fully feasible in current experimental quantum computing platforms. Their practical implementation is constrained by hardware limitations, noise sensitivity, and the lack of standardized evaluation protocols. As a result, significant research and engineering efforts are still required to transition these defense mechanisms from theoretical models to experimentally validated solutions.

2.4 Classical (quantum-enhanced) vs quantum-native cyber attacks

As quantum computing advances, its impact on cybersecurity can be classified into two broad categories: the transformation of classical cyber attacks-particularly those that rely on cryptography-and the emergence of quantum-native attack vectors that exploit the unique physical and computational properties of quantum systems (Fig. 3). This subsection provides a structured taxonomy and analysis of both.

Quantum computing either *breaks* widely adopted classical public-key (asymmetric) schemes-via Shor's algorithm-or *speeds up* brute-force against symmetric schemes without breaking them-via Grover's algorithm-thereby challenging established security assumptions [75–78]. Consequently, cryptography-based attacks—e.g., MITM through compromised public-key cryptography and brute-force on symmetric encryption-are significantly enhanced (or, in the public-key case, effectively enabled) under quantum capabilities. To counter Shor, new techniques are needed (Sect. 2.6), whereas Grover is mitigated by using longer keys [77, 78].

Other attack vectors, such as network-level and saturation attacks (e.g., denial-of-service, poisoning), psychological attacks (e.g., social engineering), and logical exploits (e.g., malware, zero-days, web application vulnerabilities), remain largely unaffected by quantum computing. These operate at different layers of the system and do not depend on cryptographic primitives that quantum algorithms threaten.

A further class of concern includes AI-oriented attacks, such as data poisoning, evasion, model inversion, and privacy breaches. While not directly reliant on cryptographic weakness, these threats may be amplified when combined with quantum-accelerated attacks. For example, a brute-force attack may be used to gain unauthorized access to a system, which then allows an attacker to carry out a data poisoning attack by injecting manipulated or malicious data into the training set. Similarly, an evasion attack that tricks a model into misclassifying inputs at inference time can be further exacerbated by existing dataset biases, amplifying the risk of unfair outcomes. In such combinations, the first attack often acts as a vector or enabler that creates the conditions necessary for the second, more damaging attack to be executed effectively.

Although the structure of classical threats remains largely intact, the defensive landscape must be re-engineered-particularly where cryptographic integrity is foundational. Strategies for post-quantum encryption and mitigation will be discussed in the following subsection.

Quantum-native attacks, by contrast, target vulnerabilities specific to the design and operation of quantum technologies. These require an entirely new class of security mechanisms.

At the hardware level, adversaries may perform noise injection or amplification attacks, deliberately manipulating environmental conditions (such as cross-talk or electromagnetic interference) to degrade qubit performance. Fault injection techniques, using radiation or targeted pulses, can similarly destabilize circuit execution.

Compiler-level threats emerge from the complexity of transpilation and circuit mapping. Attackers might introduce subtle manipulations-such as hidden gate insertions or altered topologies-that create undetectable backdoors without changing the intended high-level computation.

Scalability presents another challenge: as quantum systems increase in qubit count and circuit depth, the attack surface expands non-linearly. This creates new opportunities for interference, leakage, or instability across entangled subsystems.

In quantum machine learning contexts, inference and privacy attacks are especially relevant. Techniques such as membership inference, readout falsification, or model extraction through quantum tomography can compromise both the confidentiality and integrity of quantum-based AI systems.

Finally, shared infrastructures-such as quantum cloud services or federated quantum learning-introduce risks related to side-channel exposure and cross-user contamination. Multi-tenant environments amplify concerns about isolation, data leakage, and service-level compromise.

Just as classical threats demand classical defenses, defending against quantum-native attacks requires a form of cybersecurity native to the quantum domain-that is, quantum cryptography. This, in turn, necessitates the design of entirely new security paradigms grounded in the physics and architecture of quantum computing.

2.5 Classical cryptography

Cryptography plays a fundamental role in mitigating cybersecurity risks. It involves various methods and algorithms that secure communications, protect sensitive data, and maintain privacy within digital systems. Traditionally, cryptographic techniques are categorized into two primary approaches: symmetric key cryptography and asymmetric key cryptography. Alongside these, other essential methods include hash functions and digital signatures, each serving distinct yet complementary roles in maintaining data security.

Symmetric key cryptography, often referred to as secret key cryptography, involves the use of a single key shared privately between the communicating parties. The same key is employed for both encrypting and decrypting information, necessitating secure methods of key distribution to prevent unauthorized access. Among the widely adopted techniques in this category is the Advanced Encryption Standard (AES), a robust and widely implemented block cipher. AES encrypts data in fixed-size blocks of 128 bits, using keys of lengths 128, 192, or 256 bits, providing varying levels of security based on key length. Historically significant methods such as the Data Encryption Standard (DES), though now considered insecure due to its shorter key length (56-bit), illustrate the evolution of symmetric encryption. To address DES's vulnerabilities, Triple DES (3DES) was developed, applying the DES algorithm three times consecutively to enhance security. Nevertheless, due to its limitations, AES remains the current industry standard for symmetric encryption, favored for its strong security properties and computational efficiency.

On the other hand, asymmetric key cryptography, or public key cryptography, revolutionized secure communication by employing two distinct keys-a public key for encryption and a private key for decryption. This approach eliminates the need to secretly exchange keys beforehand, significantly simplifying secure interactions over unsecured channels. The most famous asymmetric method is the RSA (Rivest–Shamir–Adleman) cryptosystem, which relies on the computational complexity of factoring large prime numbers to ensure security. Elliptic Curve Cryptography (ECC) represents a modern evolution in this domain, leveraging the mathematical properties of elliptic curves to offer comparable security to RSA but with significantly smaller key sizes. Other fundamental methods like the Diffie-Hellman (DH) key exchange facilitate secure key negotiation over insecure channels, while ElGamal encryption, based on Diffie-Hellman principles, is integral to various cryptographic protocols.

Complementing these encryption methods are hash functions, which play a pivotal role in ensuring data integrity rather than confidentiality. Hash functions convert input data of arbitrary length into a fixed-size output called a hash or digest, making them invaluable for verifying data integrity. Commonly utilized hash functions include the SHA-2 family, featuring algorithms like SHA-256 and SHA-512, widely integrated into cryptographic applications. The more recent SHA-3, based on the Keccak algorithm, represents the latest advancement in secure hashing. In contrast, older methods such as MD5 (Message Digest Algorithm 5), while historically significant, are now deemed insecure due to vulnerabilities allowing hash collisions, thereby underscoring the continuous need for advancements in cryptographic hashing.

Moreover, ensuring authentication and integrity of digital communications necessitates the use of digital signatures, which confirm that messages have not been altered and verify the identity of senders. Digital signatures rely heavily on asymmetric cryptographic principles, with RSA-based digital signatures being among the most widely deployed, ensuring both message integrity and sender authentication. The Elliptic Curve Digital Signature Algorithm (ECDSA) offers a modern alternative, combining the strengths of ECC with the practicality of smaller key sizes. Similarly, the Digital Signature Algorithm (DSA), a U.S. federal standard based on discrete logarithms, remains a critical tool for secure document verification.

Together, these classical cryptographic techniques form the backbone of secure communications across diverse sectors, including online financial transactions, secure messaging applications, authentication systems, and broader data protection efforts. However, the reliability and effectiveness of these cryptographic solutions are not solely defined by technical robustness. Their integrity is inherently linked to the ethical principles guiding their development and use

(Sect. 3.1). As cyber threats grow increasingly sophisticated, purely technical approaches to cybersecurity reveal their limitations, underscoring the necessity of a holistic approach integrating ethical considerations into cybersecurity practices.

In Sect. 3.1 we will delve deeper into how ethical frameworks provide essential guidance for navigating the complex interactions among technological innovation, individual rights, and societal values. Such a comprehensive approach ensures that advancements in cybersecurity not only enhance technical security but also contribute meaningfully to creating a more trustworthy digital environment.

2.6 Post-quantum cryptography vs quantum cryptography

Having previously examined classical cryptographic techniques, it is now essential to consider the new approaches required to secure classical systems in the era of quantum computing. As quantum computers become more powerful, traditional cryptographic methods-such as RSA, ElGamal, Diffie-Hellman, and elliptic curve cryptography (ECC)-are increasingly vulnerable. These classical methods rely on the computational complexity of certain mathematical problems, like integer factorization and discrete logarithms, which have historically been infeasible to solve within practical timeframes using classical computers. However, quantum algorithms such as Shor's algorithms (1994) [75, 79] alter this landscape by drastically reducing the complexity and time required to solve these problems, rendering many classical cryptographic systems insecure in a quantum computing environment.

In response to this emerging threat, two distinct but complementary approaches have emerged: quantum cryptography and PQC (also known as quantum-safe cryptography) [80–83]. Quantum cryptography directly harnesses principles of quantum mechanics to ensure security, with QKD being the most notable example. QKD enables two parties to securely generate a shared secret key by exploiting fundamental quantum properties, such as the quantum superposition of states and the no-cloning theorem. According to this theorem, quantum states cannot be perfectly copied; measuring or intercepting qubits (quantum bits) inevitably alters their state. This inherent property ensures that any eavesdropping attempts on the key exchange are immediately detectable, making QKD exceptionally secure. In fact, QKD is often considered the second "perfect cipher" after the one-time pad (OTP). The OTP, originally proposed by Gilbert Vernam in 1919 [84], achieves information-theoretic security by using a key that is truly random, as long as the message, and used only once. Its perfect secrecy was formally proven by Claude Shannon in 1949 [85], who demonstrated that, under these conditions, the ciphertext reveals no information about the plaintext.

Despite its promise, quantum cryptography faces significant practical limitations. According to the National Security Agency (NSA) and other cybersecurity authorities, including the NIST, quantum cryptography presents challenges such as the high cost of quantum devices, extensive infrastructure requirements, and vulnerability to classical attacks like denial-of-service. Given these constraints, it is currently considered less practical for widespread deployment, particularly in securing large-scale communication networks and critical infrastructures. However, the field is rapidly evolving, with ongoing advancements in miniaturization, integration, and deployment strategies. Notably, satellite-based QKD has emerged as a promising approach to overcome terrestrial distance limitations and enable global-scale secure communication, as demonstrated by projects like China's Micius satellite [86]. These developments indicate that, while still facing obstacles, quantum cryptography is progressing toward more scalable and practical implementations.

In contrast, PQC aims to provide security against quantum computing attacks without relying on quantum technologies themselves. PQC leverages new mathematical problems that are believed to remain computationally difficult even for quantum computers. However, unlike the information-theoretic guarantees of the one-time pad or the physics-based security of QKD, the hardness assumptions underlying PQC have not been formally proven. The mathematical foundation of these schemes is still under active investigation, and their long-term security remains an open question, especially in light of potential advances in quantum algorithms and classical cryptanalysis.

Unlike traditional public-key cryptography vulnerable to Shor's algorithms [75], PQC methods are designed explicitly to resist quantum-enabled attacks. For instance, while classical RSA relies on the difficulty of factoring large integers-now easily solvable by quantum computers-PQC introduces alternative mathematical constructs, such as lattice-based cryptography, multivariate cryptography, hash-based signatures, and code-based cryptography. These constructions are currently regarded as promising candidates for post-quantum security, aiming to provide cryptographic primitives that are resilient to both classical and quantum computational threats-although rigorous, long-term validation is still an open challenge.

Given these advantages, PQC is widely recognized as the most feasible and practical solution for the post-quantum era. This view is supported by major international cybersecurity agencies, including the NIST, France's Agence nationale de la sécurité des systèmes d'information (ANSSI), Germany's Federal Office for Information Security (BSI),

the Netherlands' National Communications Security Agency (NLNCSA), and Sweden's National Cyber Security Centre (SNCSA/SAF). In line with this consensus, since 2016, NIST has spearheaded an initiative to develop and standardize one or more post-quantum public-key cryptographic algorithms, aiming to provide robust, long-term security for digital systems against future quantum threats.

Thus, the transition from classical to PQC represents a proactive approach to cybersecurity, anticipating and addressing the risks posed by quantum computing. However, beyond the technical aspects, this transition also raises critical ethical considerations. Decisions about adopting quantum or post-quantum cryptographic approaches must be guided by ethical frameworks that consider the broader societal impacts. Issues such as the transparency of these new technologies and accountability for potential failures must be integral to any cybersecurity strategy. In Sect. 3.1 we will explore these ethical considerations in greater depth, evaluating how ethical frameworks can guide the responsible integration of these advanced cryptographic methods into digital infrastructures, ensuring they contribute positively to a trustworthy digital environment.

2.7 Fully homomorphic encryption and quantum fully homomorphic encryption

In response to the growing security challenges posed by quantum computing, particularly those that threaten traditional cryptographic methods, researchers have turned to advanced techniques such as Fully Homomorphic Encryption (FHE). Falling under the broader category of PQC, FHE emerges as a powerful strategy for protecting data collected and pre-processed within hybrid classical-quantum ML systems. As the pre-processing phase typically handles sensitive information, privacy-preserving techniques like FHE are particularly relevant for safeguarding confidentiality and ensuring secure computations in the face of quantum-enhanced adversaries [87].

An FHE scheme operates in three main phases: first, data is encrypted; then, computations are performed directly on the encrypted data; finally, the output is decrypted. Crucially, the decrypted result is identical to the result that would have been obtained if the operations had been performed on the plaintext-this is the core property of homomorphism. In ML pipelines, this allows sensitive data to be processed by untrusted systems without revealing its contents.

However, integrating FHE into quantum contexts presents new challenges. In hybrid classical–quantum systems, the transition between classical and quantum processing phases introduces additional constraints. A key issue is that classical data encrypted via FHE cannot be trivially converted into qubits and then processed homomorphically in a quantum system. The noise introduced by FHE encryption renders the data incompatible with quantum circuits, particularly those used in QML models. Thus, to preserve homomorphism across the entire workflow, a more structured encryption path must be followed: first, applying FHE on the pre-processed classical data, then encoding the output into a quantum representation, and finally processing it through a FHE scheme tailored to quantum inputs. This layered approach is necessary to maintain homomorphic properties throughout and ensure that the final decrypted quantum output corresponds to the correct transformed data.

A specialized concept within this domain, Quantum Fully Homomorphic Encryption (QFHE), has been proposed to extend homomorphic encryption principles into the quantum realm [88]. QFHE allows secure quantum computations while preserving the confidentiality of quantum inputs-whether these are directly sensed quantum states (e.g., from quantum sensors) or data encoded into qubits during hybrid ML processing. This is particularly relevant for QML applications that involve highly sensitive or confidential data, such as healthcare records or genomic information. Nonetheless, applying FHE techniques in quantum contexts is not without difficulties. For instance, verifying quantum computations performed on encrypted data is complicated by quantum mechanical constraints like the no-cloning theorem. Recent advancements such as verifiable QFHE (vQFHE) address this issue by enabling non-interactive delegation and verification of quantum computations [88].

Despite promising theoretical foundations, integrating FHE and QFHE into practical QML pipelines remains an open research challenge. Issues such as performance overhead, quantum circuit compatibility, and the complexity of maintaining encryption consistency across classical and quantum domains are still being addressed [89]. Moreover, applying FHE to complex ML models like Convolutional Neural Networks (CNNs) or XGBoost may require significant architectural adaptations, which could impact efficiency and scalability [87, 90, 91].

Emerging frameworks aim to bridge these gaps, enabling privacy-preserving training and inference without interactive decryption. While these frameworks are currently designed for classical ML, their adaptation to quantum algorithms is under exploration. In this evolving landscape, achieving a balance between security, model accuracy, and computational efficiency remains a key hurdle.

In conclusion, although FHE and its quantum counterpart QFHE offer compelling solutions for securing hybrid classical-quantum ML systems, their practical implementation remains at an early stage. Continued research is essential to overcome current limitations and fully realize their potential, particularly in privacy-critical sectors such as healthcare, genomics, and finance. Addressing these limitations

can contribute not only to ensuring security and privacy principles but also to promoting the responsible deployment and use of such technologies. As discussed in Sect. 3.1, integrating these ethical dimensions into the technical security framework can help ensure the trustworthiness of hybrid classical-quantum AI systems.

3 Framework demonstration

Starting from the possible attacks, risks, and technical mitigation measures discussed in Sects. 2.2 and 2.3 for both classical and hybrid classical–quantum AI, we now examine how this knowledge shapes the construction of integrated security frameworks that explicitly connect technical requirements to ethical dimensions.

In particular, we demonstrate how security frameworks are built upon specific technical mitigation measures-derived from regulatory frameworks and technical standards-and how, from these measures, it becomes possible to construct a corresponding ethical framework grounded in principles of bioethics such as non-maleficence and justice, and technology ethics such as transparency and responsibility. This approach moves beyond a purely technical perspective, establishing a coherent and ethically informed framework that integrates security ethics across the entire lifecycle of AI systems.

Mitigation measures thus serve as the essential first line of defense against targeted cyberattacks, ensuring the protection of data, networks, and systems throughout all phases of classical and hybrid AI operations. Compliance with key regulatory frameworks-including the GDPR, AI Act, Cybersecurity Act, NIS2 Directive, DORA, and NIST/ISO/IEC standards [19, 30, 42, 44–48, 50, 52, 54–57]—not only strengthens these technical defenses but also reinforces the broader commitment to security, privacy, and ethical accountability.

Among the measures, special attention is given to encryption in hybrid AI systems due to its critical role in data security. We discuss scenarios employing symmetric and asymmetric cryptography, differentiating between local pre-processing, processing, and post-processing versus local pre-processing coupled with non-local processing and post-processing. This distinction highlights the increased security achievable with non-local approaches, particularly through homomorphic cryptography, which, with the evolution of quantum computing, may become practically accessible in the future-especially for reducing the computational overhead typical of classical homomorphic encryption implementations.

Table 8 presents a unified technical-and-ethical security framework for classical and hybrid AI systems.

To bridge technical security measures and ethical oversight, it is essential to clarify the ethical foundations that underpin the framework. Understanding where each principle originates allows us to trace how technical security measures translate into moral responsibilities that guide responsible AI governance. The ethical principles referenced in the framework (privacy, security, fairness, reliability, responsibility, non-maleficence, transparency, explainability, intellectual property, responsibility, justice) draw from two main traditions: bioethics and technology ethics.

Principles such as non-maleficence and justice have their roots in bioethics—particularly the well-known framework articulated by Beauchamp and Childress in Principles of Biomedical Ethics (1979)—which identifies non-maleficence, beneficence, justice, and autonomy as foundational ethical pillars in the life sciences, developed to safeguard individual well-being, ensure fair treatment, and uphold personal rights [92, 93].

In parallel, principles like transparency, explainability, fairness, responsibility, and intellectual property emerge from the field of technology ethics, which addresses the moral implications of designing, developing, and deploying technological systems [94, 95]. The inclusion of intellectual property as an ethical principle reflects the recognition, within technology ethics, of the need to respect and fairly attribute innovation, ensuring that creators' rights are protected and that unjust appropriation or misuse of technological outputs is avoided—aligning with broader concerns of fairness and justice in digital environments. These principles align with international AI guidelines, such as those outlined by the OECD AI Principles, the Ethics Guidelines for Trustworthy AI by the European Commission's HLEG-AI (2019), and the WHO's guidance on AI ethics in health (2021), all of which emphasize fairness, transparency, and explainability in AI system design and governance [10, 17, 22].

Some principles, such as security, reliability, and privacy, bridge both domains—reflecting not only technical safeguards but also fundamental ethical duties to protect individuals, communities, and societal trust in technological systems [96, 97].

Table 2 summarizes the meaning of each principle in relation to its ethical tradition.

To justify the ethical principles presented in the framework, it is necessary to connect them to: the purpose of the adopted technical measures (i.e., the concrete why behind security or protection—what ethical principle justifies them? What values and principles the risk and measures in order to mitigate them can infringe?); the risks they aim to mitigate; the regulatory or philosophical/ethical-legal references that proposed those principles in AI contexts (Table 8).

Table 2 Mapping of ethical principles to foundational ethical bases

Ethical principle	Foundational ethical basis
Privacy	Rooted in respect for autonomy and informed consent, privacy ensures that individuals have the right to control their personal information. This includes safeguarding the confidentiality (preventing unauthorized disclosure), integrity (preventing unauthorized modification), and availability (ensuring timely and reliable access) of personal data
Security	Grounded in the principle of non-maleficence, security ensures that systems protect individuals, data, and infrastructure not only from accidental failures or harm, but also from intentional, malevolent actions such as cyberattacks, unauthorized access, and exploitation
Fairness	Based on the principle of justice, ensuring equitable treatment and avoiding discrimination or bias
Reliability	Linked to beneficence and non-maleficence, ensuring systems work properly and do not cause harm through errors or malfunctions
Responsibility	Connected to professional ethics; actors are answerable for the system's impacts and decisions
Transparency	Related to respect for persons and the right to explanation, allowing people to understand system decisions
Non-maleficence	Directly from bioethics, the obligation to avoid causing harm
Explainability	An extension of transparency, ensuring that decisions are interpretable and supporting autonomy and responsibility
Intellectual Property	Anchored in justice and fairness, recognizing the rights and contributions of creators
Justice	Emphasizes fair distribution of benefits and risks, avoiding discrimination or favoritism

Table 3 Technical metrics associated with ethical principles and regulatory references

Ethical value	Technical metrics	Regulations/standards
Privacy	Encryption level, presence of DP, access logging	GDPR, ISO/IEC 29134
Fairness	Equal opportunity diff, statistical parity, fairness accuracy	AI Act, ISO/IEC 24027
Explainability	Presence of XAI, % of decisions explained, user feedback	ISO/IEC 24029-1, AI Act
Non-maleficence	Residual risk assessment + simulated attacks	AI Act art. 9-15, NIST AI RMF
Responsibility	Active logging, audit system, traceability	GDPR art. 5(2), DORA
Reliability	Stress test, output consistency, failover test	ISO/IEC 25010, NIST RMF

Moreover, we introduce an additional rationale related to the measurement phase of the assessment. While ethical validation is often regarded as measurable only through qualitative or semi-quantitative methods-being rooted in argumentation and value judgment-anchoring it to technical measures and regulatory criteria allows for the possibility of quantitative assessment which is an emerging area in AI and ethics scholarship (Table 3).

3.1 An overarching security framework for classical and hybrid artificial intelligence systems

A classical AI system follows a structured lifecycle, from data collection to decommissioning (Table 8). Each phase presents specific security challenges and requires technical measures to mitigate associated risks. Without appropriate protections, AI models can become targets of cyberattacks that undermine ethical principles and fundamental rights. These security practices align with regulatory frameworks such as the GDPR, the AI Act, the Cybersecurity Act, the NIS2 Directive, and DORA, as well as standards by NIST and ISO/IEC [19, 30, 42, 44–48, 50, 52, 54–57]. This subsection examines each phase of the AI framework, analyzing the implemented security measures, associated risks, potential cyber threats, and the ethical principles that support the secure and responsible design, development, and use of AI systems (Table 9).

3.1.1 Data collection and preparation

This phase is decisive as it determines the quality and security of the data used to train the AI model. Without proper protections, data can be stolen, manipulated, or corrupted, leading to biased or unreliable AI behavior.

One of the primary security measures in this phase is *encryption*, which secures data during storage and transmission. Encryption prevents unauthorized access to data in transit and at rest, mitigating the risk of interception and exfiltration; to mitigate MITM attacks (where attackers intercept and manipulate communications), encryption must be combined with proper peer authentication (e.g., server certificate validation or mTLS). When using AEAD (Authenticated Encryption with Associated Data) schemes (e.g., AES-GCM or ChaCha20-Poly1309), encryption ensures both confidentiality and integrity of personal data, thereby protecting the principles of **Privacy** and **Security**.

Another relevant measure is *differential privacy*, which limits the possibility of identifying individuals within a dataset (privacy violation). Without this protection, attackers could exploit re-identification or inference attacks to extract sensitive personal information. By masking individual data points, minimizing re-identification risks and providing

uniform privacy guarantees across demographic groups, differential privacy protects **Privacy** and can support **Fairness** (although it does not by itself ensure fair outcomes).

To prevent data manipulation before training, *data validation* ensures that only legitimate and accurate data are used. Without validation, attackers could inject malicious or poisoned data (and, where unsafe loaders or executable formats are present, even malware), corrupt the dataset and influence the model's behavior. By ensuring that only accurate data are processed and supporting reliable AI outcomes, data validation protects the principle of **Reliability**.

Access control is another essential security measure that limits who can view or modify sensitive data. Without access control, AI models are vulnerable to credential theft and account takeover (via social engineering, malware, MITM, or brute-force attacks), allowing unauthorized users to manipulate data. By restricting system access to authorized users, thereby enforcing responsibility and preventing potential data misuse, access control upholds the principle of **Responsibility**.

Finally, *data sanitization* ensures that the collected data are free from malicious code or unauthorized content. If this measure is ignored, attackers could embed malicious payloads (e.g., through malware) within the dataset, introducing security vulnerabilities into the AI system. By removing malicious or corrupt data prior to processing and avoiding the generation of harmful or misleading outputs, data sanitization protects the principle of **Non-maleficence**.

3.1.2 Model training

During this phase, the AI system learns from data, making it particularly vulnerable to manipulation and cyberattacks.

To prevent unauthorized access, *environmental isolation* ensures that the training environment is protected from external physical and logical interference/noise (for instance, by separating data computation from data storage and segmenting networks/environments). If compromised, an attacker could infiltrate the training infrastructure and modify the learning process; this is possible in case of *account takeover* from social engineering, malware, MITM, or brute-force attacks. By preventing external intrusions and preserving system integrity, environmental isolation upholds the principle of **Security**.

Another important security measure is *model watermarking*, which is intended to deter and detect the unauthorized copying of AI models. Without this, attackers could perform reverse engineering to duplicate the model and use it illicitly (malware can facilitate such theft, but is not strictly required). By helping track ownership and supporting equitable protection of proprietary innovations, watermarking protects the principles of **Intellectual Property** and **Justice**.

To enhance AI robustness, *adversarial training* prepares the model to recognize and resist adversarial attacks, where manipulated inputs are designed to deceive AI systems. If this security measure is missing, attackers could exploit adversarial vulnerabilities, causing AI to produce incorrect or misleading results; this also occurs in the presence of data and model poisoning attacks. By strengthening the model against manipulative data and parameters and reducing the likelihood of harmful or erroneous decisions, adversarial training protects the principles of **Non-maleficence** and **Reliability**.

Logging and monitoring play a key role in detecting anomalies during training. If left unchecked, unauthorized modifications to training data or process could go unnoticed, potentially affecting AI performance and security. For instance, this includes cases where untrusted content induces prompt-like manipulations during data ingestion. By providing clear traceability of system behavior and enabling oversight by stakeholders, logging and monitoring protect the principles of **Transparency** and **Responsibility**.

To protect sensitive data during training, *homomorphic encryption* allows computation on encrypted data without decryption, thereby preserving the confidentiality of sensitive information throughout the training phase and minimizing the risk of exposure under privacy attacks (noting that FHE currently incurs significant overhead and is applied to targeted computations). By enabling processing in encrypted form, homomorphic encryption protects the principles of **Privacy** and **Security**.

Finally, *quality control of training data* prevents data poisoning attacks, where malicious actors inject harmful data to manipulate AI behavior. By verifying data accuracy against possible manipulations and supporting honest system behavior and ethical responsibility, data quality control protects the principles of **Responsibility** and **Transparency**.

3.1.3 Validation and testing

Before deployment, the AI model must be rigorously tested to ensure it operates fairly, securely, and effectively.

One of the most significant concerns in AI systems is bias. Bias testing is essential to identify hidden prejudices in the model and the data. While bias can arise from non-malicious causes (e.g., sampling or labeling practices), it can also be *exacerbated or deliberately induced* by adversaries via *data or model poisoning*, injecting biased data or parameters to favor or discriminate against specific groups. By helping to identify and mitigate systemic discrimination and promoting equal treatment regardless of background, bias testing protects the principle of **Fairness**.

To assess the model's resilience to adversarial behavior, *adversarial attack simulations* are conducted. These

tests evaluate whether the model can resist *evasion attacks*, where attackers submit manipulated inputs to deceive the model at inference time. By verifying robustness under realistic threats and adversarial conditions, attack simulation protects the principle of **Reliability**.

To ensure that validation activities are not compromised, *isolated test environments* are used. Isolation helps prevent *account takeover* (e.g., via social engineering) and unauthorized changes to test assets, and reduces the risk of *logical tampering* or *physical interference* that could alter test data or results. By protecting the system from alterations and enabling a responsible rollout of AI models, isolated testing protects the principles of **Security** and **Responsibility**.

Another key aspect is the use of *XAI* techniques, which make model decision-making more intelligible. Without explainability, there is a risk of decision opacity (a "black box" effect) that impedes interpretation and oversight, and can hinder the timely detection of issues such as data or model poisoning. By making decisions intelligible and fostering trust and informed use of the technology, explainable AI techniques protect the principles of **Explainability** and **Responsibility**.

3.1.4 Deployment

Once the model is validated, it is deployed for real-world use, requiring security measures to prevent theft and misuse.

To protect AI models from tampering, *model encryption* should be complemented by *code/artefact signing* and, where feasible, *runtime attestation* (e.g., TEE). Encryption protects confidentiality (and, with AEAD, ciphertext integrity), while *signatures and attestation* ensure the model has not been altered or replaced. In transit, *TLS 1.3 with certificate validation (or mTLS)* mitigates MITM; at rest, encryption limits the impact of exfiltration when only encrypted media are stolen. By preventing unauthorized interception and by enabling integrity verification against tampering, these controls protect the principles of **Security** and **Reliability**.

To prevent unauthorized modifications to the AI system, *access control* remains critical. Without it, *admin account takeover* (via social engineering, malware, MITM, or brute-force attacks) could allow attackers to change the model's behavior. By limiting modification rights so that only authorized actors can influence the deployed system and by creating clear responsibility, access control protects the principle of **Responsibility**.

Lastly, *persistent watermarking* helps deter illicit use of AI models after deployment and supports ownership tracking. Without it, attackers could perform unauthorized use or attempt reverse engineering (malware can facilitate such theft). By deterring unlicensed deployment and protecting rights through traceability and attribution, persistent watermarking protects the principles of **Intellectual Property** and **Justice**.

3.1.5 Inference and operations

After the model is deployed, the inference phase begins. This is when the AI processes new data to generate outputs. This phase is particularly vulnerable to adversarial attacks, which can manipulate inputs to distort the AI's responses outputs.

One of the main security measures implemented is *input validation*, which filters and verifies incoming data before the model processes it. Without this protection, an attacker could send malicious or specially crafted inputs designed to confuse the model (poisoning attacks), causing it to make incorrect decisions and compromising system reliability. By filtering problematic inputs and helping prevent harmful errors (manipulated results) while supporting consistent model performance, input validation protects the principles of **Non-maleficence** and **Reliability**.

To monitor abnormal behavior during inference, *anomaly detection* is used. This technique helps identify suspicious patterns or attempts to manipulate inference through crafted inputs (e.g., prompt injection via hostile external content). If the model starts producing unexpected or inappropriate results, anomaly detection enables timely intervention. By identifying unusual or malicious behaviors in real time and supporting secure and consistent AI use, anomaly detection protects the principle of **Reliability**.

Another significant risk is the exposure of sensitive information in the AI-generated outputs. To protect user privacy, *encryption of output data* (e.g., TLS 1.3 in transit and encryption at rest) prevents third parties from intercepting and analyzing data produced by the model. The primary threat here is unauthorized exposure of generated data, which could reveal confidential information. By protecting the results of sensitive analyses against interception and unauthorized access, output encryption protects the principle of **Privacy and Security**. Notably, encryption does not prevent leakage if the model itself emits sensitive content; apply access control, redaction, and privacy filters as needed.

To prevent misuse or persistent attacks, *continuous monitoring* of the deployed model is implemented. This ensures observability (logging, metrics, tracing) and rapid response to abuse, such as using the model to generate harmful content (e.g., via prompt injection) or operating it in unauthorized scenarios. By enabling continuous oversight, ensuring ethical use, and supporting rapid intervention in case of misuse, monitoring protects the principle of **Responsibility**.

Finally, to maintain clarity in system operations, *XAI* techniques are adopted, enabling stakeholders to understand

the model's decision-making process. Without XAI techniques the system risks becoming a "black box," leading to decision opacity and making it difficult to identify potential biases or errors stemming from poisoning attacks or data issues. By clarifying model outputs so that stakeholders can assess whether decisions are equitable and grounded in sound reasoning, explainable AI techniques protect the principles of **Explainability** and **Fairness**.

3.1.6 Maintenance and updates

AI models require periodic updates to remain secure and functional.

Security patches address known vulnerabilities, reducing the risk of 1-day and n-day exploits on outdated components. By closing vulnerabilities that could be exploited and preserving the system's secure operation, security patches protect the principle of **Security**.

Before production rollout, *pre-deploy testing* ensures that updates do not introduce errors or misconfigurations. If skipped, there is a risk of faulty updates degrading AI performance or introducing new vulnerabilities that could later be exploited (including zero-day until disclosed). By validating updates before deployment and maintaining responsibility for functional integrity, pre-deploy testing protects the principle of **Responsibility**.

Regular audits help detect persistent attacks or unnoticed vulnerabilities/configuration drifts. Without audits, attackers may conduct ongoing cyberattacks on AI systems (including stealthy exploitation). By identifying missed threats and ensuring reliable, responsible system operation, regular audits protect the principles of **Responsibility** and **Reliability**.

Version control (version management) ensures that different AI iterations remain fully traceable. Without it, there is a risk of losing model traceability and unintentionally retaining outdated versions that are more exposed to 1-day or n-day exploits. By enabling end-to-end tracking of all modifications and supporting transparency in model evolution, version control protects the principle of **Transparency**.

3.1.7 Decommissioning

When an AI model is no longer in use, it must be securely retired to prevent unauthorized access.

Secure model destruction ensures that the model cannot be recovered or misused after termination. The primary risk is unauthorized model retrieval or reuse from leftover artefacts (e.g., backups, snapshots, containers, logs, temporary files), which an attacker could leverage or repurpose (malware may facilitate such theft but is not required). By ensuring the model cannot be retrieved or misused after system termination, secure destruction protects the principle of **Non-maleficence**.

Revoking cryptographic material and access prevents continued access to AI systems and data. Beyond cryptographic keys, this includes revoking or rotating certificates, API/refresh tokens, and disabling or deleting accounts and roles. If not revoked, compromised or leftover credentials can enable unauthorized access (e.g., following account takeover via social engineering, malware, or brute-force attacks). By preventing ongoing access after decommissioning and preserving the system's secure operation, revocation protects the principle of **Security**.

Final audit ensures that all security measures have been correctly applied. If neglected, residual vulnerabilities or leftover artifacts could remain and be leveraged by attackers (e.g., via 1-day/n-day exploits). By confirming that any remaining risks have been addressed by the appropriate parties-such as key and account revocation, removal of DNS/endpoints, and secure data/media sanitization-the final audit ensures responsible system decommissioning and protects the principle of **Responsibility**.

3.2 A comparative analysis of security frameworks in classical and hybrid artificial intelligence systems

AI systems-whether classical or hybrid-follow the same lifecycle composed of the following phases: data collection, training, validation, deployment, inference, maintenance, and decommissioning (Tables 9, 10). Each phase presents distinct security risks and ethical implications, requiring targeted mitigation measures. While both classical and hybrid systems share foundational principles such as privacy, fairness, and reliability, hybrid AI introduces additional ethical challenges and countermeasures due to the integration of quantum technologies-whose physical sensitivity (e.g., decoherence) can increase impact and ethical salience.

3.2.1 Data collection and preparation (classical vs hybrid artificial intelligence)

Both classical and hybrid systems require secure data collection and preparation through data validation, data sanitization, access control, and differential privacy. Classical systems typically use encryption (e.g., TLS 1.3, AES-256). Hybrid systems should additionally adopt *PQC* for key establishment and signatures (e.g., ML-KEM/Kyber, ML-DSA/Dilithium); *QKD* may be considered only in specialized, hardware-enabled settings. From an ethical point of view, both emphasize privacy, security, reliability, and non-maleficence; hybrid systems add future-proof confidentiality against "store-now, decrypt-later" threats.

3.2.2 Model training (classical vs hybrid artificial intelligence)

In classical AI, training relies on environmental isolation, watermarking, adversarial training, logging/monitoring, data quality control, and-in niche use-homomorphic encryption for specific computations. Hybrid systems build on these with *classical–quantum isolation*, calibrated transpilation/mapping, and stricter provenance. Integrity of artefacts/logs should be enforced with *post-quantum digital signatures*. *QFHE* can be referenced as a *forward-looking, experimental option* for privacy-preserving computation during training-useful for research pilots but currently constrained by high overhead and limited practicality. From an ethical point of view, both systems stress security, privacy, responsibility, intellectual property, justice, reliability, non-maleficence, and transparency. In hybrid systems these principles are more *salient* due to quantum hardware sensitivity (e.g., physical noise channels) and stricter interface isolation.

3.2.3 Validation and testing (classical vs hybrid artificial intelligence)

Bias testing, attack simulations, isolated test environments, and XAI techniques are used in both paradigms. Hybrid systems may add *quantum-aware attack simulations* (e.g., robustness under calibration drift) and *quantum-aware test isolation* to prevent result corruption. Ethical priorities remain fairness, reliability, security, responsibility, and explainability, with stronger guarantees sought due to probabilistic behavior and added complexity.

3.2.4 Deployment (classical vs hybrid artificial intelligence)

Both systems implement artefact *signing/attestation*, encryption in transit/at rest, watermarking, and access control. Hybrid systems should employ *PQC-hybrid key exchange* in transport (e.g., X25519+Kyber) while continuing to use symmetric AEAD (AES-GCM/ChaCha20-Poly1305). From an ethical point of view, this phase links to security, reliability, responsibility, intellectual property, and justice; hybrid deployments underscore these due to additional interfaces and dependencies.

3.2.5 Inference and operations (classical vs hybrid artificial intelligence)

Input validation, anomaly detection, output encryption, monitoring, and XAI techniques remain fundamental. Hybrid systems should protect channels and artefacts with PQC-backed key establishment (and, where dedicated quantum links are available, QKD-derived keys) to resist long-term decryption risks; data confidentiality and integrity are then enforced with symmetric AEAD. Ethical principles include non-maleficence, reliability, privacy, responsibility, explainability, and fairness; hybrid contexts stress explainability under quantum uncertainty and complexity.

3.2.6 Maintenance and updates (classical vs hybrid artificial intelligence)

Routine updates, audits, and version control are necessary in both settings. Hybrid AI adds coordinated versioning across classical and quantum components and timely updates to cryptographic stacks (e.g., enabling PQC, rotating certificates). This phase reflects commitments to security, responsibility, reliability, and transparency.

3.2.7 Decommissioning (classical vs hybrid artificial intelligence)

Secure destruction of models and artefacts, revocation/rotation of cryptographic material (keys, certificates, tokens), and a final audit are required to prevent post-use access, model reuse, and residual vulnerabilities. Ethical priorities-non-maleficence, responsibility, and security-remain constant.

While classical and hybrid AI systems share a common lifecycle and overlapping security principles, hybrid systems require advanced protections-*primarily PQC* (and, in niche cases, QKD)-to address quantum-relevant threats. *QFHE* is retained as forward-looking, experimental options for privacy during computation in select scenarios. These differences extend beyond technical implementation to ethics, requiring not only protection against current vulnerabilities but also foresight for future quantum capabilities. A unified framework accommodating both paradigms supports ethical, secure, and trustworthy AI across computing architectures, noting that the quantum stack's physical characteristics broaden the attack surface.

3.3 Different encryption scenarios

Let's resume the hybrid AI system with the different numbered phases. These phases are grouped into pre-processing (whose data are classical), processing (whose data are quantum) and post-processing (whose data are classical). The data are initially collected and prepared for processing (using techniques such as data cleaning, feature extraction, and other preprocessing techniques). Subsequently, the data, through encoding, will pass from classical data to quantum data and will be ready to be fed to a supervised QML model. The model formed by quantum gate-based circuits will allow to apply the principles of superposition

and entanglement to quantum data. In this way, the model will learn to perform a certain task, such as classification (recognizing a certain class of objects, such as patients suffering or not suffering from a certain pathology). At the end of the learning, the data will be converted back into classics through the measurement of the probabilistic quantum state. In this way, the measurements will be repeated by creating probability distributions of a given state, with the consequent analysis of the distributions. Hence, the most likely outcome of the state will be the expected final result. So, going back to the initial example, we will have that the trained model will be able to recognize a certain patient having a certain probability of suffering from a certain pathology with a margin of error, given by the inaccuracy and non-precision of the trained model. This measurement process will lead to the calculation of the goodness of the model in recognizing a given situation from the input features, returning to the training phase if necessary, and then repeating the measurement phase later. This will be followed by the phase of collecting new data to validate the model (trying to optimize the performance or goodness of the model) and test the model (to know the behavior of the model itself on the most varied data). To validate the model, the measurement will have to be carried out again, and then understand whether to retrain the model again in case of non-goodness of the model or whether to move on to the testing phase. Also from testing, the measurement phase and the possible retraining of the model will follow in case of non-goodness of the model. In the case of model goodness, the phases of model deployment, inference and operations, maintenance and updating and decommissioning will follow to complete the model life cycle.

For the various processing phases, we imagine three different encryption scenarios differentiated by current and future scenarios and by the type of processing (local or not). The first scenario is the current one, with classic cryptography. Pre-processing, processing and post-processing is carried out locally. Data collection includes security given by the TLS cryptographic protocol necessary for data transfer and storage with AES symmetric encryption. The dataset is prepared locally, so the data from AES encryption can be decrypted to have it in plaintext. Quantum model training does not require encryption, because the data is inherently protected theoretically (ideal case). The subsequent security phases, from model deployment to decommissioning, will have AES encryption for data storage and TLS protocol in the case of data transfer. The second scenario is the near-term scenario with the availability of PQC or even current with quantum cryptography, i.e., QKD. As in the previous scenario, the pre-processing, processing and post-processing are carried out locally. The present scenario is similar to the previous one in the different phases, but it uses encryption that protects against quantum attacks. Therefore, the longer key AES (AES-256) and the TLS protocol with post-quantum asymmetric encryption or QKD are used for key exchange. The third scenario is inherently forward-looking (future-oriented scenario), particularly in the context of PQC and quantum cryptography. In this case, data are encrypted using FHE during the preprocessing and post-processing phases, while QFHE is employed during the quantum processing phase. This multilayered encryption approach is especially relevant when all computation phases are executed remotely in a cloud environment, where data confidentiality must be preserved end-to-end.

3.4 Case study: quantum-enhanced medical diagnosis system

Q-MedAI is an AI system enhanced with quantum algorithms, designed to improve diagnostic accuracy in the analysis of radiological images (e.g., CT scans and MRIs). The system is integrated into a distributed hospital network and receives data from multiple connected clinics. An external attacker attempts to exploit a network vulnerability to launch a MITM attack during the data collection phase. By intercepting medical images and patient metadata, the attacker violates the principles of privacy and security, as the intercepted data contain sensitive personal health information that may be exposed or altered without patient consent. This undermines confidentiality-a core component of data privacy-and compromises data integrity. Both are fundamental to ensuring secure and trustworthy AI systems, particularly in healthcare contexts where the consequences of misuse or misdiagnosis can be severe. At the same time, the attacker launches a data poisoning attack during the training phase by injecting manipulated images to confuse the model's ability to correctly identify tumor patterns. The system is supported by an intelligent monitoring framework that analyzes risks and proposes specific countermeasures. In this case, the MITM attack was unsuccessful, as post-quantum encryption had been implemented. Even with access to a quantum computer, the attacker was unable to decrypt communications protected by PQC. As a result, the framework did not flag an ethical violation of the principles of privacy and security. The data poisoning attack was also unsuccessful, as the system had been trained using quantum adversarial training, which increased its resilience to manipulated inputs. The monitoring module confirmed that the model maintained its reliability and did not produce clinically significant alterations in diagnostic outcomes. Therefore, the risks to privacy, security, and reliability were effectively mitigated, thanks to robust, by-design defensive mechanisms integrated into the system. The principle of non-maleficence was preserved, as the model did not

Table 4 Comparison between classical and hybrid AI systems: process structure

Aspect	Classical AI system	Hybrid AI system
Main phases	Data collection, preprocessing, training, validation, inference, maintenance	Preprocessing (classical), quantum processing (training, validation, testing), and classical post-processing via measurement
Processing	Traditional computation using classical processors (CPU, GPU)	Hybrid computation using classical processors (CPU/GPU) together with quantum processing units (QPUs), often accessed through cloud platforms

generate incorrect or harmful diagnoses. In the opposite case, where such measures were not present, the framework would have flagged the ethical violations and recommended the implementation of appropriate design-stage defenses to minimize harm. It would be promising to introduce a phase for measuring both technical and ethical risk, which could be explored as technical future work.

4 Discussion

In this section, we discuss the results by comparing the two frameworks—classical and hybrid classical-quantum - to understand how the ethical security challenges evolve across these different AI systems. Specifically, we analyze how the integration of quantum components reshapes the landscape of technical and ethical risks, offering insights into where traditional approaches remain applicable and where novel solutions are required. Building on the methods presented in Sect. 2, where we transition from security to security ethics, we now examine the results detailed in Sect. 3, focusing on the system phases, the technical security measures, the associated risks and attacks, and finally the corresponding ethical principles. First of all, the phases of the system are the same but the typology is clearly different (Table 4). In the hybrid system we have the data processing phase (training, validation and testing) which is quantum (Fig. 3). This recalls the need to convert classical data into quantum data through different techniques such as basis, amplitude and phase coding, and then be processed in the quantum circuit and be converted back into classical data with the measurement process. Moreover, computation in hybrid systems is not limited to classical processors such as CPUs and GPUs, but may also involve QPUs, which are the hardware responsible for executing quantum algorithms on qubits. Currently, QPUs are primarily accessible through cloud platforms offered by companies such as IBM, Google, and IonQ, and are used in combination with classical resources within the hybrid system. For example, in quantum machine learning models, the QPU repeatedly executes a parameterized quantum circuit, while a classical CPU receives the measurement results and updates the circuit's parameters using an optimization algorithm, such as gradient descent or COBYLA. Tools like Qiskit allow developers to design and run quantum circuits either on simulators (which use classical processors) or on real QPUs via cloud access. Within hybrid ML workflows, QPUs, when employed, can handle the quantum component of computation, working alongside classical processors in tasks such as data preprocessing, inference, and variational circuit optimization. When it comes to technical security measures, encryption, model and data protection, attack detection and robustness to adversarial training are different between the two systems (Table 5). In both systems we can find AES-256 symmetric encryption to protect the stored data with a long key that guarantees security in case of quantum attacks. For data in transit with TLS, both systems could use QKD or PQC algorithms for the exchange of secure keys to be used to encrypt data with a secure algorithm such as AES-256.

Table 5 Comparison between classical and hybrid AI systems: technical security measures

Category	Classical AI system	Hybrid AI system
Cryptography	AES-256 and TLS used for data in transit; adoption of post-quantum cryptography (PQC) anticipated	Quantum cryptography integrated into both data transmission and computation. Adoption of PQC in classical computation
Model protection	Obfuscation techniques and classical watermarking applied during training and deployment	Quantum watermarking used to protect intellectual property during quantum model execution and inference
Data protection	Differential privacy and anonymization used to preserve confidentiality	Differential privacy enhanced with quantum mechanisms to ensure secure processing, e.g., in federated learning
Attack detection	Conventional logging and anomaly detection systems for monitoring and response	Quantum-secured logging supported by quantum digital signatures to ensure tamper-proof event records
Robustness to adversaries	Adversarial training based on classical vulnerabilities and known threat models	Quantum adversarial training designed to mitigate manipulation of quantum circuits and inputs

As an alternative to TLS, HE could be used, although it is not an efficient solution for classical models due to the computational weight, it is rather a secure and sustainable solution for the hybrid model with quantum processing. This is useful for secure cloud computing, such as using Qiskit, for example. The other differences in technical security measures are watermarking to protect the model's intellectual

property, differential privacy to protect data confidentiality, logging to protect the system from suspicious or malicious activity, and adversarial training to protect the system from adversarial attacks. These differences find the use of quantum mechanics compared to classical mechanics for protection. Let's come to the differences in technical risks and attacks (Table 6). The most frequent attacks and the greatest risks occur in inadequate protection of data (as in the case of interception) and model protection (as in the case of theft or alteration of the model). One case, in particular, is that of adversarial attacks that can target input data. Again, as for the measurements, the difference between the two systems lies in the different foundations, i.e., classical and quantum.

For example, attacks on the hybrid AI system will be due to flaws in the quantum environment such as model alteration, model theft or introduction of malicious inputs specific to quantum models or to those classical components that are vulnerable to quantum attacks.

All listed technical measures will have a positive impact on security. These technical measures are not only defensive measures—they are deeply intertwined with fundamental ethical principles that shape responsible AI governance. Specifically, they operationalize values such as privacy, security, explainability, non-maleficence, reliability, intellectual property, and fairness, each of which takes on more specific meanings depending on whether we consider classical or hybrid classical-quantum AI systems (Table 7).

Privacy, traditionally focused on protecting personal data, must evolve in hybrid AI contexts to address the computational capabilities of quantum systems. This includes adopting PQC and QFHE, which enable encrypted computations in quantum deep learning without compromising sensitive information. From the ethical standpoint, this reinforces the right of individuals to control their data and avoid exposure to inference or reconstruction attacks—especially critical in sectors like healthcare and finance where data misuse can have severe human consequences. However, this comes at a very high computational and technical cost. With current technologies, QFHE remains poorly scalable and difficult to adopt in real-time or resource-constrained environments. Even simple arithmetic operations on encrypted data can require orders of magnitude more computation than traditional processing. Latency can increase dramatically—from milliseconds to minutes or even hours, depending on the complexity of the task and the available hardware. QFHE also requires specialized cryptographic infrastructure and advanced quantum-safe hardware, which are challenging to deploy at scale. These trade-offs between ethical compliance and system performance must be considered, and future research might apply the framework to understand how to overcome them according to specific context sensitive domains and scenarios [98, 99].

Thus, QFHE holds theoretical promise for enabling privacy-preserving quantum machine learning, but at present it incurs significant computational overhead.

Security, more traditionally understood as the protection of infrastructure—including networks, systems, and devices—must also adapt. While classical AI systems rely on tools like firewalls and intrusion detection systems (IDS), hybrid systems demand the integration of quantum-specific techniques such as QKD and quantum-secured communication channels. From the ethical standpoint, this broadens the duty of non-maleficence: organizations must proactively secure infrastructures not just against known threats but against emerging quantum-enabled attacks, ensuring that system failures or breaches do not cause harm.

Explainability—critical for ensuring that AI decisions can be understood, audited, and justified—also faces distinct challenges. In classical AI, methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are widely used to clarify model reasoning. However, the probabilistic nature of measurements in quantum computing complicates interpretability, requiring the development of ad hoc quantum explainability tools. From the ethical standpoint, this reinforces principles of explainability and fairness: users and stakeholders have the right to understand how decisions are made, especially when these decisions affect rights, opportunities, or access to resources.

Non-maleficence and reliability similarly require rethinking. While classical systems address these principles through techniques that detect and mitigate adversarial

Table 6 Comparison between classical and hybrid AI systems: technical risks and attacks

Aspect	Classical AI system	Hybrid AI system
Data interception	Data theft through classical techniques such as man-in-the-middle (MITM) attacks and phishing	Quantum-enabled attacks (e.g., using Shor's algorithm) capable of breaking classical encryption and exposing secure keys
Model theft	Unauthorized model duplication through reverse engineering or extraction techniques	Model theft during quantum execution or through attacks targeting quantum watermarking mechanisms
Model alteration	Tampering with training data or model parameters to degrade performance or introduce bias	Alterations within the quantum processing environment, exploiting vulnerabilities in qubit manipulation or circuit design
Adversarial attacks	Input perturbations designed to mislead classical models and cause incorrect outputs	Malicious quantum-specific inputs targeting variational circuits or quantum data encodings to manipulate predictions

Table 7 Comparison between classical and hybrid AI systems: ethical implications

Ethical principle	Classical AI system	Hybrid AI system
Privacy	Data protection through encryption and anonymization	Enhanced confidentiality through post-quantum and quantum cryptography to mitigate risks from quantum attacks
Security	Protection based on traditional technologies such as firewalls and intrusion detection systems	Security enhanced by integrating quantum techniques to protect data and models during quantum processing
Explainability	Model behavior explained using Explainable AI (XAI) tools like SHAP or LIME	Need to develop new XAI techniques for quantum decisions, which are often less interpretable due to quantum complexity
Non-maleficence	Mitigation of manipulation and bias through classical testing methods	Incorporation of quantum-specific validation to ensure robustness against malicious or biased quantum inputs
Reliability	Models hardened against adversarial attacks in classical environments	Quantum adversarial training to ensure decision accuracy under quantum-specific attack conditions
Intellectual property	Protection of models through classical watermarking and obfuscation	Advanced IP protection using quantum watermarking, even during quantum computation and inference
Fairness	Bias mitigation through data- and model-level balancing and fairness auditing	Additional testing required to prevent new forms of discrimination introduced by quantum data encoding or processing

attacks, hybrid systems face quantum-specific adversarial risks—for instance, exploiting quantum noise or entanglement properties to disrupt model outcomes. Here, the ethical duty extends beyond avoiding intentional harm to proactively anticipating how new attack surfaces could be misused, ensuring models remain robust and reliable.

Intellectual property and fairness take on a sharper edge in hybrid AI. Advanced quantum watermarking techniques are needed to protect proprietary quantum models from theft or unauthorized replication, safeguarding not only commercial value but also the ethical principle of justice—recognizing and compensating the rightful contributions of creators. At the same time, the quantum system's tendency to magnify existing biases raises serious fairness concerns. During training and validation, poisoning attacks can exploit this by amplifying errors on underrepresented groups or by pushing the hybrid model toward biased outcomes through sensitive quantum features. To address these risks, it is essential to implement fairness checks and bias mitigation strategies across both classical and quantum components of the hybrid AI system. In summary, the ethical dimensions embedded in security frameworks are not abstract ideals; they are concrete operational imperatives. As we transition from classical to hybrid AI systems, these principles must be deliberately rethought, expanded, and reinforced to ensure that ethical commitments remain robust and meaningful in the face of new technical challenges.

Among the key technical measures supporting this ethical security transition, encryption is the most important aspect in the transition from the classic system to the hybrid one. It does not change in the collection and preparation phase if we think of cryptographic techniques resistant to quantum attacks such as AES 256, QKD and PQC. A separate discussion concerns FHE and QFHE, which are used to protect data processed in deep learning systems. QFHE leverages the quantum computing component of the AI model to gain the computational advantage necessary for performing calculations more efficiently while keeping the data encrypted. This, in fact, is not only what differentiates the hybrid system from the classical but also the real advantage of using quantum in a hybrid system. This homomorphic encryption is made necessary when the data will be processed not locally but on cloud platforms, even in federated learning situations. Homomorphic encryption will thus allow data protection when users invited to participate in federated learning have to participate without seeing the data in plain text.

Another important consideration is that quantum processing often does not require encryption, as the principles of quantum mechanics-such as the disturbance caused by measurement-naturally protect quantum information. However, despite this intrinsic protection, certain types of physical-layer quantum attacks may still allow adversaries to extract partial information from a quantum system, potentially enabling reconstruction of the original data. These attacks do not target the software layer, but rather exploit physical aspects of quantum hardware or implementation weaknesses. For instance, in a quantum Trojan-horse attack, an adversary may inject additional signals or exploit imperfections in quantum devices to access sensitive information

without being detected. Such vulnerabilities highlight the importance of not only theoretical security, but also the robustness of real-world quantum hardware.

This highlights the relevance of a scenario that assumes no quantum interference or hardware-based vulnerabilities-such as quantum noise, decoherence, or photon loss. Precisely for this reason, the proposed framework integrates QFHE, enabling encrypted data processing under all conditions, whether locally or on external platforms.

4.1 Use case scenario in a healthcare context

Data Collection and Preparation Phase. In the initial phase, the system collects large volumes of highly sensitive information: genomic sequences, pathology reports, treatment histories, and demographic data. This stage represents a prime target for malicious actors. Attacks such as brute force against weak credentials or malware injected into data streams by third-party providers could compromise the datasets on which the system is built. To mitigate these risks, the framework adopts a layered security approach, combining QKD, PQC, and advanced symmetric encryption. Transferred data are protected using AES-256, a high-performance symmetric encryption standard still considered secure even in post-quantum scenarios, provided that the encryption keys are well protected. In trusted environments with direct connections between nodes (such as between a hospital and a data center), QKD is used to generate and distribute AES keys in a physically secure manner. Thanks to quantum mechanics, any attempt to intercept the key produces detectable anomalies, offering theoretically unconditional protection. However, since QKD requires dedicated optical channels and is not applicable in all contexts (e.g., over the Internet, with remote vendors or mobile devices), the framework also integrates post-quantum cryptographic algorithms, such as Kyber for secure key exchange and Dilithium for digital signatures. In these scenarios, PQC ensures resistance against future quantum attacks and preserves the integrity and authenticity of the data, even in distributed and less controlled environments. Once the data are securely received-protected by AES-256 encryption and secure key exchange through QKD or PQC-the system applies a series of internal measures to ensure data quality and protection. Data validation and sanitization techniques verify that the information is not corrupted, incomplete, or potentially malicious. Then, differential privacy is used to inject statistical noise, ensuring that patient identities cannot be reconstructed, even from aggregated datasets-thus guaranteeing equal protection for all individuals, including those from vulnerable or minority groups. Finally, strict access control policies, such as multi-factor authentication, ensure that only authorized personnel can access or process the data. These measures complete the system's security perimeter and represent a concrete implementation of the ethical principles of the framework: privacy, security, and fairness, achieved through the combined use of advanced encryption and differential anonymization, as well as reliability and responsibility through control, access, and traceability mechanisms.

Quantum Processing Phase. During the processing phase, the hybrid system leverages quantum algorithms for highly complex tasks, such as quantum clustering for tumor subtype identification or personalized optimization of chemotherapy regimens. While these capabilities offer enormous potential in terms of accuracy and speed, they also introduce new vulnerabilities-particularly concerning the confidentiality and integrity of data during quantum computation. To address these risks, the framework adopts QFHE, which allows computations to be performed directly on encrypted data. This means that even those with access to the quantum infrastructure cannot view or manipulate the plaintext data, ensuring constant protection during the most sensitive computational stages. In addition to encryption, a key component is the structural and logical separation between classical and quantum modules, known as quantum-classical isolation. This architectural measure prevents any interference between the two computational environments, ensuring that a compromise of the classical system-such as an attack on a hospital server or orchestration layer-does not propagate to the quantum module (e.g., by tampering with input data or execution commands). Likewise, it protects the classical system from anomalies or malicious behaviors originating from the quantum level. This compartmentalization not only strengthens the technical resilience of the architecture but also enables more precise and traceable control over data flows between the two environments. In AI systems applied to healthcare, where each computation can have a direct clinical impact, these are not merely technical features-they are essential elements of an ethical approach committed to privacy, security, and responsibility toward patients and practitioners.

Post-Processing and Clinical Interface Phase. Once quantum processing is complete, the results-such as predicted responses to oncological treatments or patient risk

stratification-are returned to the clinical team, which may include oncologists, radiologists, hospital pharmacists, and therapeutic planning coordinators. At this stage, the AI system's output comes into direct contact with real clinical decisions, influencing drug allocation, initiation of experimental protocols, or adjustment of treatment strategies. However, this critical step introduces new vulnerabilities: results may be intercepted during transmission, tampered with in transit across hospital IT systems, or even altered through user interface manipulation (e.g., a compromised clinical software that changes the output displayed to the physician). Such errors could lead to incorrect or inappropriate treatment choices with direct consequences for the patient's health. To prevent these risks, the framework implements output encryption throughout the transmission phase, protecting privacy and blocking unauthorized access. At the same time, cryptographic integrity checks ensure that the data received matches exactly the data generated, ruling out tampering during transit. Beyond technical protection, a key element is the integration of XAI tools directly into clinical interfaces. These tools allow medical professionals to visualize the reasoning behind the algorithm's recommendations. Only by understanding the decision-making process can clinicians validate or challenge the AI outputs, detect potential errors or clinical biases, and ensure adherence to fairness principles, delivering impartial assessments even for patients from minority groups.

Continuous Monitoring and Updates. In a hospital or healthcare network, security is not a one-time installation-it must be maintained over time. The framework addresses persistent and advanced threats, such as data poisoning-where an attacker corrupts clinical datasets used to train the model, leading to harmful treatment recommendations-or quantum adversarial attacks, which exploit specific vulnerabilities in quantum architectures to produce anomalous behaviors. To counter these risks, the system includes continuous monitoring capable of detecting suspicious activities or anomalous outputs from the quantum module. This monitoring can be integrated with existing clinical audit systems (e.g., diagnosis traceability, protocol verification), enabling hospital IT teams and ethics committees to intervene promptly in case of irregularities. In parallel, the framework includes a robust software update infrastructure, featuring: post-quantum patches to defend against emerging computational threats;pre-deployment testing to ensure that new versions do not introduce vulnerabilities; regular audits certifying the operational correctness and reliability of the AI system. These measures are essential to ensure that AI continues to function securely, reliably, and in alignment with clinical and ethical standards. In doing so, the system not only protects computational integrity, but also upholds the principle of non-maleficence-avoiding harm-and the principle of ongoing responsibility in patient care.

5 Limitations and future work

Given that this is a theoretical framework oriented toward future scenarios and the integration of advanced quantum technologies, we also consider security measures that are not yet fully implementable, as they remain the subject of ongoing research. This is the case, for example, with QFHE, which-despite its strong theoretical potential for secure encrypted data processing in quantum environments-still faces significant limitations in terms of scalability and computational efficiency [100]. Nonetheless, we argue that even theoretical models are essential: without a foundational conceptual step, there can be no meaningful guidance for future technical research aimed at developing and deploying ethically secure QAI systems. In this sense, the framework provides a necessary ethical and methodological reference point for aligning research trajectories with emerging governance needs. At this stage, we limit ourselves to outlining future research directions in these areas, which may, over time, support the practical integration of such techniques into operational systems. A fundamental step will be the standardized and widespread release of post-quantum cryptographic systems, such as those currently being adopted by NIST (e.g., Kyber, Dilithium), which are already designed to withstand future quantum attacks [101]. In parallel, advances in quantum hardware may enable the operational implementation of innovative techniques such as quantum watermarking, which, although still in the prototyping phase, are already recognized as promising for applications like intellectual property protection and forensic traceability of AI models [102]. Despite this forward-looking orientation, several limitations remain. One important limitation concerns the scalability of the framework, which depends on its adaptability across domains with varying levels of technical maturity and ethical risk. Another challenge lies

in the implementation barriers-many of the techniques referenced (e.g., QFHE, quantum watermarking) are still under development and not yet mature enough for integration into real-world systems. As future work, it will be necessary to empirically test the framework and its individual components, and to develop targeted case studies to assess its applicability and implementation in specific domains. This would allow for the identification of context-sensitive trade-offs and the evaluation of which techniques and protections are ethically advisable depending on the scenario. Moreover, the framework should be translated into actionable policies-for both designers and adopters of QAI technologies-that are sensitive to the particularities of each application domain. These developments-and the progressive overcoming of the aforementioned limitations-are central to our ongoing and future technical, ethical, and policy research. However, we emphasize that this framework constitutes a necessary theoretical foundation to guide, legitimize, and coordinate such efforts in a coherent and functional manner, aligned with the evolving landscape of AI ethics and governance.

6 Practical recommendations for implementation

The following recommendations are intended to guide both organizations and policymakers in ensuring ethical security in the lifecycle of AI systems, particularly in light of post-quantum threats and evolving regulatory expectations [103].

6.1 Recommendations for organizations

- **Develop Quantitative Risk Assessment Tools** Based on the proposed framework, organizations should develop internal tools capable of measuring AI-related security and ethical risks using quantitative, reproducible metrics to support governance and decision-making.
- **Assess AI Systems Using the Proposed Framework** Organizations should map existing AI applications to the stages outlined in the proposed framework to identify both technical and ethical security gaps. Particular attention should be given to:
  - Implementing advanced encryption methods (e.g., AES-256, Kyber)
  - Integrating digital watermarking to ensure data traceability and integrity
  - XAI capabilities within decision-making interfaces
  - Applying fairness mitigation strategies such as equalized odds post-processing and reweighting
- **Train Personnel** Continuous training programs should be established to educate professionals on emerging risks, quantum-resilient security practices, and the ethical implications of AI use.
- **Form Multidisciplinary Security and Ethics Teams** Institutions are encouraged to establish dedicated teams that bring together experts in AI, cybersecurity,and ethics. These teams should oversee the design, deployment, and ongoing evaluation of AI systems.
- **Implement Continuous Monitoring and Auditing** AI systems should be subject to continuous monitoring of inferences, system logs, and software updates. Periodic audits can help detect anomalies, vulnerabilities, and compliance issues before they escalate.
- **Utilize Open-Source Toolkits for Fairness and Explainability** Where appropriate, organizations should adopt existing open-source solutions-such as IBM's AI Fairness 360 toolkit-to assess and improve the transparency and fairness of their AI systems.

6.2 Recommendations for policymakers and regulators

- **Set Minimum Standards for Post-Quantum Security** Regulatory bodies should define and enforce baseline post-quantum security requirements, including advanced encryption, robust authentication, and digital watermarking for AI systems.
- **Mandate Explainability in High-Impact Clinical AI** AI systems that directly impact decisions should be required to incorporate explainability features, ensuring that outputs can be interpreted and justified by all stakeholders.
- **Support Research on Quantum-Resilient Technologies** Public funding should prioritize exploratory

research and prototyping in areas such as QFHE and quantum-based watermarking, which offer strong potential for long-term security.

- **Promote Open and Interoperable AI Standards** Standard-setting organizations should encourage the adoption of open, transparent, and interoperable technical standards for XAI components, audit logging, and hybrid (classical-quantum) system architectures.
- **Establish Ethical and Technical Oversight Committees** Local and national governments should facilitate the creation of dedicated committees tasked with reviewing AI deployments, offering both technical scrutiny and ethical guidance in line with best practices.
- **Use the proposed framework to develop security ethics guidelines** Policymakers should use the proposed framework to develop security ethics guidelines which are currently lacking for both classical and hybrid AI systems. In the case of AI, there is still no comprehensive framework that meaningfully integrates security ethics into broader AI ethics and governance. For hybrid systems, such a framework is entirely absent.
- **Test the framework in real-world domains** Policymakers should test the framework in real-world domains-such as healthcare, finance, mobility, and public services-to develop context-specific policy guidance through the involvement of relevant stakeholders.

7 Conclusion

This paper introduces the first ethical security framework for both classical and hybrid AI, aimed at paving the way for future efforts to support stakeholders (academia, industry, and institutions) in governing the entire AI lifecycle in compliance with current regulations, from both a technical and ethical perspective.

This framework represents the first theoretical tool for ethical security in a classical–quantum hybrid system, even though, to date, no such framework exists even for purely classical AI systems. In fact, building on that initial foundation, our work extends beyond classical technologies to address hybrid systems. The framework introduces a notion of security ethics that goes far beyond the existing literature. This means starting from concrete security measures and corresponding ethical principles justifying them for AI and hybrid AI risk management to provide a comprehensive security ethics tool. Structurally, the framework is developed considering existing regulations and ethical frameworks developed for classical AI systems, which are expanded to consider security ethical challenges, and then overall reinterpreted considering also quantum AI technologies.

To translate this framework into action, we propose a set of practical recommendations tailored to both organizations and policymakers.

For organizations, we emphasize the need to develop quantitative risk assessment tools, assess existing AI deployments against the proposed framework, prioritize robust encryption, fairness and explainability, and invest in continuous monitoring, training, and interdisciplinary governance structures. Policymakers and regulators, in turn, should establish minimum post-quantum security standards, mandate explainability for high-impact clinical systems, support research and standardization efforts focused on resilient and transparent AI practices, use the proposed framework to develop security ethics guidelines, and test the framework in real-world domains.

In combination, these recommendations aim to ensure a trustworthy AI system capable of withstanding future challenges.

Appendix A Security frameworks for artificial intelligence systems

This appendix presents three supporting tables that complement the main results (Sects. 3.1, 3.2). Table 8 provides a generic security framework across the AI lifecycle (risks, mitigations, ethical principles, their justification, and regulatory coverage). Tables 9 and 10 tailor the framework to classical AI and hybrid AI, including phase-specific security measures and examples of cyber attacks.

Table 8 Security framework across the AI lifecycle

Lifecycle stage/ risk	Data collection and preparation	Model training	Validation and testing	Deployment	Inference and operations	Maintenance and updates	Decommissioning
Unauthorized access	Encryption. Privacy and Security are called into question in case of encryption techniques since they ensure the confidentiality and integrity of personal data, preventing unauthorized access. GDPR, Cybersecurity Act, NIS2	Environmental isolation Security is called into question in case of environmental isolation since it prevents external intrusions and preserves the integrity of the system. NIST RMF, Cybersecurity Act					Key revocations. Security is called into question in case of key revocation since it prevents ongoing access to systems or data after decommissioning, preserving systemic security. NIS2, ISO/IEC 27002
Privacy violation	Differential privacy. Privacy and Fairness are called into question in case of differential privacy since it masks individual data points, minimizing re-identification risks and ensuring equal protection across all demographic groups. GDPR, AI act, ISO/IEC 29134	Encryption. Privacy and Security are called into question in case of homomorphic encryption since it allows data to be processed in encrypted form, protecting sensitive information throughout the training phase. Cybersecurity Act, ISO/IEC 27002, GDPR			Output encryption. Privacy is called into question in case of output encryption since it protects the results of sensitive analyses from unauthorized exposure. GDPR, Cybersecurity Act		
Manipulation	Data validation. Reliability is called into question in case of data validation since this measure ensures that only accurate data are processed, supporting reliable AI outcomes. NIST AI RMF, ALTAI	Adversarial training. Non-maleficence and Reliability are called into question in case of adversarial training since it strengthens the model against manipulative inputs, reducing the likelihood of harmful or erroneous decisions. NIST AI Guidelines, ISO/IEC 27001		Secure protocols. Reliability is called into question in case of secure protocols since protect communication channels from interception, reinforcing model reliability in operation ISO/IEC 27002, Cybersecurity Act	Input validation. Non-maleficence and Reliability are called into question in case of input validation since filtering problematic inputs prevents harmful errors and ensures consistent model performance. NIST AI Guidelines, ALTAI		

Table 8 (continued)

Lifecycle stage/ risk	Data collection and preparation	Model training	Validation and testing	Deployment	Inference and operations	Maintenance and updates	Decommissioning
Misuse	Access control. Responsibility is called into question in case of access control since it restricts system access to authorized users, thereby enforcing responsibility and preventing potential data misuse Cybersecurity Act, NIS2, DORA				Monitoring. Responsibility is called into question in case of monitoring since it enables continuous oversight, ensuring ethical use and rapid intervention in case of misuse. DORA, ISO/IEC 27001		
Malicious data	Data sanitization. Non-maleficence is called into question in case of data sanitization since it removes malicious or corrupt data prior to processing, thus avoiding the generation of harmful or misleading outputs. GDPR, AI act, ISO/IEC 27001						
Theft		Watermarking. Intellectual Property and Justice are called into question in case of watermarking since this technique helps track ownership and supports equitable protection of proprietary innovations. AI Act, ISO/IEC 42001		Encryption. Security is called into question in case of model encryption since it prevents unauthorized interception or duplication of the deployed system, safeguarding its confidentiality and integrity. Cybersecurity Act, NIS2			

Table 8 (continued)

Lifecycle stage/ risk	Data collection and preparation	Model training	Validation and testing	Deployment	Inference and operations	Maintenance and updates	Decommissioning
Unauthorized changes		Logging/Monitoring. Transparency and Responsibility are called into question in case of logging and monitoring since these practices provide clear traceability of system behavior and enable oversight by stakeholders. ALTAI, GDPR, DORA		Access Control. Responsibility is called into question in case of access control since limits modification rights, ensuring only authorized actors influence the deployed system, fulfilling responsibility. GDPR, DORA, NIS2			
Data poisoning		Data quality control. Responsibility and Transparency are called into question in case of data quality control since verifying data accuracy directly supports honest system behavior and ethical responsibility. AI Act. ISO/IEC 25012					
Undetected bias			Bias testing. Fairness is called into question in case of bias testing since it helps to identify and mitigate systemic discrimination, ensuring equal treatment for all individuals regardless of background. AI Act, ISO/IEC 24027, ALTAI				

Table 8 (continued)

Lifecycle stage/ risk	Data collection and preparation	Model training	Validation and testing	Deployment	Inference and operations	Maintenance and updates	Decommissioning
Robustness loss			Attack simulation. Reliability is called into question in case of attack simulation since it ensures that the model remains robust and reliable even when exposed to real-world threats and adversarial conditions. NIST RMF, Cybersecurity Act				
Result corruption			Isolated testing. Security and Responsibility are called into question in case of isolated testing since this protects the system from internal threats and supports the responsible rollout of AI models. NIST RMF, ISO/IEC 27001				
Lack of clarity/ Black-box decisions			Explainable AI. Explainability and Responsibility are called into question in case of explainable AI since making decisions intelligible is a responsibility of developers and fosters trust and informed use of the technology AI Act, ALTAI, ISO/IEC 24029-1		Explainable AI. Explainability and Fairness are called into question in case of explainable AI since it clarifies model outputs allowing stakeholders to assess whether decisions are equitable and based on ethically sound logic. AI Act, ISO/IEC 24029-1		

Table 8 (continued)

Lifecycle stage/ risk	Data collection and preparation	Model training	Validation and testing	Deployment	Inference and operations	Maintenance and updates	Decommissioning
Illicit use				Persistent Watermark. Intellectual Property and Justice are called into question in case of persistent watermarking since it protects rights and deter unlicensed deployment. AI Act, ISO/IEC 42001			
Irregular behavior					Anomaly detection. Reliability is called into question in case of anomaly detection since it helps identify unusual or malicious behaviors in real-time, supporting secure and consistent AI use. Cybersecurity Act, ISO/IEC 27001		
Vulnerabilities						Security patches. Security is called into question in case of security patches since they close vulnerabilities that could be exploited, thereby maintaining the security operation of the system. Cybersecurity Act, NIST RMF	Final audit. Responsibility is called into question in case of final audit as it confirms that any remaining risks have been addressed by the appropriate parties, ensuring responsible system decommissioning. GDPR, ISO/IEC 27001, DORA

Table 8 (continued)

Lifecycle stage/ risk	Data collection and preparation	Model training	Validation and testing	Deployment	Inference and operations	Maintenance and updates	Decommissioning
Faulty updates						Pre-deploy Testing. Responsibility is called into question in case of pre-deploy testing since validates updates before deployment, maintaining responsibility for functional integrity. ISO/IEC 25010, ALTAI	
Missed threats						Regular audits. Responsibility and Reliability are called into question in case of regular audits since they identify missed threats and ensure reliable, responsible AI system operation. GDPR, DORA, ISO/IEC 27001	

Table 8 (continued)

Lifecycle stage/ risk	Data collection and preparation	Model training	Validation and testing	Deployment	Inference and operations	Maintenance and updates	Decommissioning
Confusion						Version control. Transparency is called into question in case of version control since it enables tracking of all system modifications, supporting transparency in model evolution. ISO/IEC 27001, AI Act	
Reuse							Secure destruction. Non-maleficence is called into question in case of secure destruction since it ensures data cannot be retrieved or misused after system termination. Cybersecurity Act, ISO/IEC 27040

Table 9 Security framework for classical AI

AI phase	Security measure	Risk	Cyber attack	Ethical principle	Regulations
Data collection and preparation	Encryption (TLS, AES-256)	Unauthorized data access	MITM, exfiltration	Privacy, Security	GDPR, Cybersecurity Act, NIS2
	Differential Privacy	Privacy violation	Privacy attacks	Privacy, Fairness	GDPR, AI Act, ISO/IEC 29134
	Data Validation	Data manipulation	Poisoning attacks	Reliability	NIST AI RMF, ALTAI
	Access Control	Data misuse	Account takeover (social engineering)	Responsibility	Cybersecurity Act, NIS2, DORA
	Data Sanitization	Malicious data	Payload attacks (malware)	Non-maleficence	GDPR, AI Act, ISO/IEC 27001
Model training	Env. Isolation	Unauthorized access	Account takeover (social engineering)/ noise injection	Security	NIST RMF, Cybersecurity Act
	Watermarking	Illicit use	Reverse engineering (malware)	Intellectual property, Justice	AI Act, ISO/IEC 42001
	Adversarial Training	Manipulated inputs (misleading results)	Poisoning attacks	Non-maleficence, Reliability	NIST AI Guidelines, ISO/IEC 27001
	Logging/Monitoring	Unauthorized changes	Data tampering (prompt injection)	Transparency, Responsibility	ALTAI, GDPR, DORA
	Homomorphic Encryption	Privacy leak	Privacy attacks	Privacy, Security	Cybersecurity Act, ISO/IEC 27002, GDPR
	Data Quality Control	Data manipulation	Data poisoning attacks	Responsibility, Transparency	AI Act, ISO/IEC 25012
Validation and testing	Bias Testing	Biased results	Poisoning attacks	Fairness	AI Act, ISO/IEC 24027, ALTAI
	Attack Simulation	Robustness loss	Evasion attacks	Reliability	NIST RMF, Cybersecurity Act
	Isolated Test environment	Data (or results) alteration	Account takeover (social engineering)/ noise injection	Security, Responsibility	NIST RMF, ISO/IEC 27001
	Explainable AI techniques	Black-box decisions	Poisoning attacks	Explainability, Responsibility	AI Act, ALTAI, ISO/IEC 24029-1
Deployment	Model Encryption	Model tampering	MITM attacks or exfiltration	Security and reliability	Cybersecurity Act, NIS2, ISO/IEC 27002
	Access Control	Unauthorized modifications	Account takeover (phishing)	Responsibility	GDPR, DORA, NIS2
	Persistent Watermarking	Illicit use	Reverse engineering	Intellectual property, Justice	AI Act, ISO/IEC 42001
Inference and operations	Input Validation	Manipulated results	Poisoning attacks	Non-maleficence, Reliability	NIST AI Guidelines, ALTAI
	Anomaly Detection	Irregular behavior	Prompt injection	Reliability	Cybersecurity Act, ISO/IEC 27001
	Output Encryption	Privacy breach	Privacy attacks	Privacy and Security	GDPR, Cybersecurity Act
	Monitoring	Model misuse	Prompt injection	Responsibility	DORA, ISO/IEC 27001
	Explainable AI techniques	Black-box decisions	Poisoning attacks	Explainability, Fairness	AI Act, ISO/IEC 24029-1
Maintenance and updates	Security Patches	Vulnerabilities	(1 day-n days) exploits	Security	Cybersecurity Act, NIST RMF
	Pre-deploy Testing	Faulty updates	zero-day exploits	Responsibility	ISO/IEC 25010, ALTAI
	Regular Audits	Missed threats	Undetected attacks	Responsibility, Reliability	GDPR, DORA, ISO/IEC 27001
	Version Control	Outdated version	(1 day-n days) exploits	Transparency	ISO/IEC 27001, AI Act
Decommissioning	Secure Destruction	Model misused	Malware	Non-maleficence	Cybersecurity Act, ISO/IEC 27040
	Material and access revocation	Continued access	Account takeover (malware)	Security	NIS2, ISO/IEC 27002
	Final Audit	Residual vulnerabilities	(1 day-n days) exploits	Responsibility	GDPR, ISO/IEC 27001, DORA

Table 10 Security framework for hybrid AI

AI phase	Security measure	Risk	Cyber attack	Ethical principle
Data Collection and Preparation	Quantum Key Distribution (QKD)	Key interception	MITM on QKD	Privacy, Security
	Post-quantum Cryptography (PQC)	Key compromise (unauthorized data access)	Shor's algorithm attack	Privacy, Security
	Data Validation	Data manipulation	Poisoning attacks	Reliability
	Access Control	Data misuse	Account take-over (social engineering)	Responsibility
	Data Sanitization	Malicious data	Payloads attacks (malware)	Non-maleficence
Model Training	Quantum-Classical Isolation	Cross-system interference	Noise injection	Security
	Quantum Fully Homomorphic Encryption	Privacy leak	Privacy attacks	Privacy, Security
	Watermarking	Model theft	Reverse engineering	Intellectual property, Justice
	Quantum Adversarial Training	Input manipulated	Quantum poisoning attacks	Reliability, Non-maleficence
	Quantum Log Signing/monitoring	Unauthorized changes	Data tampering	Responsibility, Transparency
	Data Quality Control	Data manipulation	Data poisoning attacks	Responsibility, Transparency
Validation and Testing	Bias Testing	Biased results	Poisoning attacks	Fairness
	Quantum Aware Attack Simulations	Robustness loss	Evasions attacks	Reliability
	Isolated Test Environments	Data (or results) alterations	Noise injection	Security, Responsibility
	Explainable AI (XAI) techniques	Black-box decisions	Poisoning attacks	Explainability, Responsibility
Deployment	PQC	Model tampering	MITM or exfiltration attacks	Security, Reliability
	Access Control	Unauthorized modifications	Account takeover (phishing)	Responsibility
	Persistent Watermarking	Illicit use	Reverse engineering	Intellectual property, Justice
Inference and Operations	Input Validation	Manipulated results	Quantum poisoning attacks	Non-maleficence, Reliability
	Anomaly Detection	Irregular behavior	Prompt injection	Reliability
	Output Encryption	Privacy breach	Privacy attacks	Privacy and Security
	Quantum Monitoring	Model misuse	Prompt injection	Responsibility
	Explainable AI (XAI) techniques	Black-box decisions	Poisoning attacks	Explainability, Fairness
Maintenance and Updates	Post Quantum Patches	Vulnerabilities	(1 day-n days) exploits	Security
	Pre-deployment Testing	Faulty updates	Zero-day exploits	Responsibility
	Regular Audits	Missed threats	Undetected attacks	Responsibility, Reliability
	Version control	Outdated version	(1 day-n days) exploits	Transparency
Decommissioning	Secure Destruction	Model reuse	Malware	Non-maleficence
	Material and Access Revocation	Continued access	Account takeover	Security
	Final Audit	Residual vulnerabilities	(1 day-n days) exploits	Responsibility

Author contributions LI implemented the system and conducted the experiments. LI, ST, and FC developed the methodology and carried out the analysis. ST and FC designed the study. LI, ST, and FC wrote the original draft of the manuscript and contributed to its review and editing. ST and FC supervised the project.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement. This work was supported by the Italian Ministry of University and Research (MUR) under the PRIN 2022 (Progetti di Ricerca di Rilevante Interesse Nazionale) Project "Trustworthy hybrid quantum-classical Artificial Intelligence for Medical Image Analysis (ThAI-MIA)" [Project code: MUR-20227HSE83] funded by the European Union-Next Generation EU.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval and consent to participate This work does not involve human participants or animals, and therefore ethical approval and informed consent were not required.

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
2. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015). https://doi.org/10.1126/science.aaa8415
3. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S.: Quantum machine learning. Nature **549**(7671), 195–202 (2017). https://doi.org/10.1038/nature23474
4. Schuld, M., Killoran, N.: Quantum machine learning in feature Hilbert spaces. Phys. Rev. Lett. **122**(4), 040504 (2019). https://doi.org/10.1103/PhysRevLett.122.040504
5. Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L., Coles, P.J.: Variational quantum algorithms. Nat. Rev. Phys. **3**(9), 625–644 (2021). https://doi.org/10.1038/s42254-021-00348-9
6. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al.: The malicious use of artificial intelligence: forecasting. Prevent. Mitigat. **20** (2018)
7. Kundu, S., Ghosh, S.: Security aspects of quantum machine learning: Opportunities, threats and defenses. In: Proceedings of the Great Lakes Symposium on VLSI 2022. GLSVLSI '22, pp. 463–468. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3526241.3530833
8. Franco, N., Sakhnenko, A., Stolpmann, L., Thuerck, D., Petsch, F., Rüll, A., Lorenz, J.M.: Predominant aspects on security for quantum machine learning: Literature review. In: 2024 IEEE International Conference on Quantum Computing and Engineering (QCE), vol. 01, pp. 1467–1477 (2024). https://doi.org/10.1109/QCE60285.2024.00173
9. High-Level Expert Group on AI: The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. Technical report, European Commission, Brussels (2020). https://doi.org/10.2759/002360
10. Organization, W.H.: Ethics and governance of artificial intelligence for health: WHO guidance. Accessed: 2025-05-16 (2021). https://www.who.int/publications/i/item/9789240029200
11. Macnish, K.N.J., Ham, J.V.D.: Ethical approaches to cybersecurity (2022) https://doi.org/10.1093/oxfordhb/9780198857815.013.28
12. Chesney, R., Citron, D.: Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. Law Rev. **107**, 1753–1819 (2019). https://doi.org/10.2139/ssrn.3213954
13. Westerlund, M.: The emergence of deepfake technology: a review. Technol. Innov. Manag. Rev. **9**(11), 40–53 (2019). https://doi.org/10.22215/timreview/1282
14. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019). https://doi.org/10.1126/science.aaw4399
15. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342
16. Vellido, A.: The importance of interpretability and visualization in machine learning for applications in medicine and healthcare. Neural Comput. Appl. **32**(24), 18069–18083 (2020). https://doi.org/10.1007/s00521-019-04051-w
17. High-level expert group on AI: Ethics guidelines for trustworthy AI. Technical report, European Commission, Brussels (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
18. IEEE global initiative on ethics of autonomous and intelligent systems: ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems (First Edition, Version 2). IEEE (2019). https://ethicsinaction.ieee.org/
19. Tabassi, E.: Artificial intelligence risk management framework (AI RMF 1.0). NIST trustworthy and responsible AI, National Institute of Standards and Technology, Gaithersburg, MD (2023). https://doi.org/10.6028/NIST.AI.100-1
20. ISO/IEC JTC 1/SC 42: ISO/IEC 23894:2023—Artificial intelligence—guidance on risk management. International Organization for Standardization (2023). https://www.iso.org/standard/77304.html
21. Friedman, B., Hendry, D.G.: Value sensitive design: shaping technology with moral imagination. MIT Press, Cambridge, MA (2019). https://doi.org/10.7551/mitpress/7585.001.0001
22. Economic Co-operation, O., Development: OECD principles on artificial intelligence. https://www.oecd.org/going-digital/ai/principles/. Accessed: 2024-05-01 (2019)
23. Formosa, P., Wilson, M., Richards, D.: A principlist framework for cybersecurity ethics. Comput. Secur. **109**, 102382 (2021). https://doi.org/10.1016/j.cose.2021.102382
24. Kozhuharova, D., Kirov, A., Al-Shargabi, Z.: Ethics in cybersecurity. What are the challenges we need to be aware of and how

to handle them? In: Kołodziej, J., Repetto, M., Duzha, A. (eds.) Cybersecurity of digital service chains: challenges, methodologies, and tools. Lecture notes in computer science, vol. 13300, pp. 202–221. Springer (2022). https://doi.org/10.1007/978-3-031-04036-8_9
25. Weber, K., Kleine, N.: Cybersecurity in health care. Ethics Cybersecur. **21**, 139–156 (2020). https://doi.org/10.1007/978-3-030-29053-5_7
26. Domingo-Ferrer, J., Blanco-Justicia, A.: Ethical value-centric cybersecurity: a methodology based on a value graph. Sci. Eng. Ethics **26**(3), 1267–1285 (2020). https://doi.org/10.1007/s11948-019-00138-8
27. Hernandez-Jaimes, M.L., Martinez-Cruz, A., Ramírez-Gutiérrez, K.A., Feregrino-Uribe, C.: Artificial intelligence for IoMT security: a review of intrusion detection systems, attacks, datasets and cloud-fog-edge architectures. Internet Things **23**, 100887 (2023). https://doi.org/10.1016/j.iot.2023.100887
28. Yeng, P.K., Yang, B., Fauzi, M.A., Nimbe, P.: A framework for exploring incentive methods towards reducing phishing susceptibility in Healthcare: based on a review and In-the-wild-field study approach. In: 2023 Intelligent Methods, Systems, and Applications (IMSA), pp. 228–234 (2023). https://doi.org/10.1109/IMSA58542.2023.10217499
29. Lorenzini, G., Shaw, D.M., Elger, B.S.: It takes a pirate to know one: ethical hackers for healthcare cybersecurity. BMC Med. Ethics **23**(1), 131 (2022). https://doi.org/10.1186/s12910-022-00872-y
30. European Union: Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing regulation (EU) No 526/2013 (Cybersecurity Act) (Text with EEA Relevance). http://data.europa.eu/eli/reg/2019/881/oj/eng
31. Create a holistic approach to data protection | IBM. https://www.ibm.com/resources/the-data-differentiator/data-protection-strategy
32. Conference of the Independent Data Protection Supervisory Authorities of the Federation and the Länder (Datenschutzkonferenz, DSK): The standard data protection model: a method for data protection advising and controlling on the basis of uniform protection goals. Technical report, Datenschutzkonferenz (DSK) (2023). https://www.datenschutz-mv.de/static/DS/Dateien/Datenschutzmodell/SDM_V3_en.pdf
33. World Health Organization (WHO). Regional Office for Europe: The protection of personal data in health information systems: principles and processes for public health. Technical Report WHO/EURO:2021-1994-41749-57154, WHO Regional Office for Europe, Copenhagen (2021). https://www.who.int/europe/publications/i/item/WHO-EURO-2021-1994-41749-57154
34. Grother, P., Ngan, M., Hanaoka, K.: Face recognition vendor test (frvt) part 3: Demographic effects. Technical Report NISTIR 8280, National Institute of Standards and Technology (2019). https://doi.org/10.6028/NIST.IR.8280
35. Grother, P., Ngan, M., Hanaoka, K.: Frvt: Demographic summaries. Technical Report NISTIR 8429, National Institute of Standards and Technology (2022). https://doi.org/10.6028/NIST.IR.8429
36. Winfield, A.F., Jirotka, M.: Ethical standards in robotics and AI. Nat. Electron. **2**(2), 46–48 (2019). https://doi.org/10.1038/s41928-019-0213-6
37. Floridi, L.: Establishing the rules for building trustworthy AI. Nat. Mach. Intell. **1**(6), 261–262 (2019). https://doi.org/10.1038/s42256-019-0055-y
38. Radhika, M., Gupta, A.: Quantum computing for healthcare: a review. Fut. Internet **15**(3), 94 (2023). https://doi.org/10.3390/fi15030094
39. Youn, J., Lee, S.-Y.: Quantum computing in medicine. J. Transl. Med. **21**(1), 662 (2023). https://doi.org/10.1186/s12967-023-04467-0
40. Orús, R., Mugel, S., Lizaso, E.: Quantum computing for finance: overview and prospects. Rev. Phys. **4**, 100028 (2019). https://doi.org/10.1016/j.revip.2019.100028
41. Krelina, M.: Quantum technology for military applications. EPJ Quantum Technol. **8**(1), 24 (2021). https://doi.org/10.1140/epjqt/s40507-021-00113-y
42. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88 (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj
43. Possati, L.: Ethics of quantum computing: an outline. Philos. Technol. **36** (2023) https://doi.org/10.1007/s13347-023-00651-6
44. Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., Hamin, M.: Adversarial machine learning: a taxonomy and terminology of attacks and mitigations. Technical Report NIST AI 100-2e2025, National Institute of Standards and Technology, Gaithersburg, MD (2025). https://doi.org/10.6028/NIST.AI.100-2e2025
45. International Organization for Standardization, International Electrotechnical Commission: Information technology–artificial intelligence–management system. Technical Report ISO/IEC 42001:2023, International Organization for Standardization, Geneva, Switzerland (2023). https://www.iso.org/standard/81230.html
46. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 277, 12 July 2024, pp. 1–157 (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
47. Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS2 Directive). https://eur-lex.europa.eu/eli/dir/2022/2555/oj. Official Journal of the European Union, L 333, 27 December 2022, pp. 80–152 (2022)
48. Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011 (Digital Operational Resilience Act). https://eur-lex.europa.eu/eli/reg/2022/2554/oj. Official Journal of the European Union, L 333, 27 December 2022, pp. 1–79 (2022)
49. Ross, R.S.: Risk management framework for information systems and organizations: a system life cycle approach for security and privacy. Technical Report 800-37 Revision 2, Gaithersburg, MD (2018). https://doi.org/10.6028/NIST.SP.800-37r2
50. International Organization for Standardization, International Electrotechnical Commission: information technology–security techniques–guidelines for privacy impact assessment. Technical Report ISO/IEC 29134:2023, International Organization for Standardization, Geneva, Switzerland (2023). https://www.iso.org/standard/86012.html
51. Information security, cybersecurity and privacy protection–information security management systems–requirements. Standard, International Organization for Standardization, Geneva, Switzerland (2022). https://www.iso.org/standard/27001
52. International Organization for Standardization, International Electrotechnical Commission: information security, cybersecurity

and privacy protection–Information security controls. Technical Report ISO/IEC 27002:2022, International Organization for Standardization, Geneva, Switzerland (2022). https://www.iso.org/standard/75652.html
53. International Organization for Standardization, International Electrotechnical Commission: software engineering–software product quality requirements and evaluation (SQuaRE)—data quality model. Technical Report ISO/IEC 25012:2008, International Organization for Standardization, Geneva, Switzerland (2008). https://www.iso.org/standard/35736.html
54. International Organization for Standardization, International Electrotechnical Commission: information technology—artificial intelligence (AI)—Bias in AI systems and AI aided decision making. Technical Report ISO/IEC TR 24027:2021, International Organization for Standardization, Geneva, Switzerland (2021). https://www.iso.org/standard/77607.html
55. International Organization for Standardization, International Electrotechnical Commission: artificial intelligence (AI)—assessment of the robustness of neural networks—Part 2: methodology for the use of formal methods. Technical Report ISO/IEC 24029-2:2023, International Organization for Standardization, Geneva, Switzerland (2023). https://www.iso.org/standard/79804.html
56. International Organization for Standardization, International Electrotechnical Commission: systems and software engineering–systems and software quality requirements and evaluation (SQuaRE)–system and software quality models. Technical Report ISO/IEC 25002:2024, International Organization for Standardization, Geneva, Switzerland (2024). https://www.iso.org/standard/78175.html
57. International Organization for Standardization, International Electrotechnical Commission: information technology–security techniques–storage security. Technical Report ISO/IEC 27040:2024, International Organization for Standardization, Geneva, Switzerland (2024). https://www.iso.org/standard/80194.html
58. Qiao, C., Li, M., Liu, Y., Tian, Z.: Transitioning from federated learning to quantum federated learning in internet of things: a comprehensive survey. Commun. Surv. Tuts. **27**(1), 509–545 (2025). https://doi.org/10.1109/COMST.2024.3399612
59. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. CCS '17, pp. 1175–1191. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3133956.3133982
60. NIST: NIST releases first 3 finalized post-quantum encryption standards. [Online]. https://www.nist.gov/news-events/news/2024/08/nist-releases-first-3-finalized-post-quantum-encryption-standards
61. Han, S., Buyukates, B., Hu, Z., Jin, H., Jin, W., Sun, L., Wang, X., Wu, W., Xie, C., Yao, Y., Zhang, K., Zhang, Q., Zhang, Y., Joe-Wong, C., Avestimehr, S., He, C.: Fedsecurity: A benchmark for attacks and defenses in federated learning and federated llms. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '24, pp. 5070–5081. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3637528.3671545
62. Joint Task Force: Security and privacy controls for information systems and organizations. NIST special publication 800-53, revision 5, National Institute of Standards and Technology (2020). https://doi.org/10.6028/NIST.SP.800-53r5
63. Phalak, K., Saki, A.A.-, Alam, M., Topaloglu, R.O., Ghosh, S.: Quantum PUF for security and trust in quantum computing. IEEE J. Emerg. Select. Topics Circuits Syst. **11**(2), 333–342 (2021). https://doi.org/10.1109/JETCAS.2021.3077024
64. Saki, A.A., Suresh, A., Topaloglu, R.O., Ghosh, S.: Split compilation for security of quantum circuits. In: 2021 IEEE/ACM International Conference on Computer Aided Design (ICCAD), pp. 1–7 (2021). https://doi.org/10.1109/ICCAD51958.2021.9643478
65. Ash-Saki, A., Alam, M., Ghosh, S.: Analysis of crosstalk in nisq devices and security implications in multi-programming regime. In: Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design. ISLPED '20, pp. 25–30. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3370748.3406570
66. Watkins, W.M., Chen, S.Y.-C., Yoo, S.: Quantum machine learning with differential privacy. Sci. Rep. **13**(1), 2453 (2023). https://doi.org/10.1038/s41598-022-24082-z
67. Hirche, C., Rouzé, C., França, D.S.: Quantum differential privacy: an information theory perspective. IEEE Trans. Inf. Theory **69**(9), 5771–5787 (2023). https://doi.org/10.1109/TIT.2023.3272904
68. Huang, J.-C., Tsai, Y.-L., Yang, C.-H.H., Su, C.-F., Yu, C.-M., Chen, P.-Y., Kuo, S.-Y.: Certified robustness of quantum classifiers against adversarial examples through quantum noise. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095030
69. Heredge, J., Kumar, N., Herman, D., Chakrabarti, S., Yalovetzky, R., Sureshbabu, S.H., Li, C., Pistoia, M.: Characterizing privacy in quantum machine learning. NPJ Quantum Inf. **11**(1), 80 (2025). https://doi.org/10.1038/s41534-025-01022-z
70. Kumar, N., Heredge, J., Li, C., Sureshbabu, S.H., Eloul, S., Pistoia, M.: Expressive quantum circuits provide inherent privacy in federated learning. APS March Meet. Abstracts **2024**, 51–54 (2024). https://doi.org/10.48550/arXiv.2309.13002
71. Franco, N., Wollschläger, T., Gao, N., Lorenz, J.M., Günnemann, S.: Quantum robustness verification: a hybrid quantum-classical neural network certification algorithm. In: 2022 IEEE International Conference on Quantum Computing and Engineering (QCE), pp. 142–153 (2022). https://doi.org/10.1109/QCE53715.2022.00033
72. Lu, S., Duan, L.-M., Deng, D.-L.: Quantum adversarial machine learning. Phys. Rev. Res. **2**(3), 033212 (2020). https://doi.org/10.1103/PhysRevResearch.2.033212
73. Gong, W., Deng, D.-L.: Universal adversarial examples and perturbations for quantum classifiers. Natl. Sci. Rev. **9**(6), 130 (2022). https://doi.org/10.1093/nsr/nwab130
74. Guan, J., Fang, W., Ying, M.: Robustness verification of quantum classifiers. In: International Conference on Computer Aided Verification, pp. 151–174 (2021). https://doi.org/10.1007/978-3-030-81685-8_7
75. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J. Comput. **26**(5), 1484–1509 (1997). https://doi.org/10.1137/S0097539795293172
76. Grover, L.K.: A fast quantum mechanical algorithm for database search. In: Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing. STOC '96, pp. 212–219. Association for Computing Machinery, New York, NY, USA (1996). https://doi.org/10.1145/237814.237866
77. Chen, L., Jordan, S.P., Liu, Y., Moody, D., Peralta, R.C., Perlner, R.A., Smith-Tone, D.C.: Report on post-quantum cryptography. NIST interagency/internal report (NISTIR) 8105, National Institute of Standards and Technology, Gaithersburg, MD (2016). https://doi.org/10.6028/NIST.IR.8105
78. European Union Agency for Cybersecurity.: Post-quantum cryptography: current state and quantum mitigation. Publications Office, LU (2021). https://doi.org/10.2824/92307

79. Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings 35th Annual Symposium on Foundations of Computer Science, pp. 124–134 (1994). https://doi.org/10.1109/SFCS.1994.365700
80. Portmann, C., Renner, R.: Security in quantum cryptography. Rev. Mod. Phys. **94**, 025008 (2022). https://doi.org/10.1103/RevModPhys.94.025008
81. Singh, M., Sood, S.K., Bhatia, M.: Post-quantum cryptography: a review on cryptographic solutions for the era of quantum computing. Arch. Comput. Methods Eng. (2025). https://doi.org/10.1007/s11831-025-10412-7
82. Vidick, T., Wehner, S.: Introduction to quantum cryptography. Cambridge University Press, Cambridge, UK (2024). https://doi.org/10.1017/9781009026208
83. Goyal, S.B., Kumar, V., Islam, S.M.N., Ghai, D. (eds.): Quantum computing, cyber security and cryptography: issues, technologies, algorithms, programming and strategies. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-4948-8
84. Vernam, G.S.: Secret signaling system. 1310719, July 1919. U.S. Patent No. 1,310,719. https://patents.google.com/patent/US1310719A
85. Shannon, C.E.: Communication theory of secrecy systems. Bell Syst. Tech. J. **28**(4), 656–715 (1949). https://doi.org/10.1002/j.1538-7305.1949.tb00928.x
86. Liao, S.-K., Cai, W.-Q., Liu, W.-Y., Zhang, L., Li, Y., Ren, J.-G., Yin, J., Shen, Q., Cao, Y., Li, Z.-P., Li, F.-Z., Chen, X.-W., Sun, L.-H., Jia, J.-J., Wu, J.-C., Jiang, X.-J., Wang, J.-F., Huang, Y.-M., Wang, Q., Zhou, Y.-L., Deng, L., Xi, T., Ma, L., Hu, T., Zhang, Q., Peng, C.-Z., Wang, J.-Y., Pan, J.-W.: Satellite-to-ground quantum key distribution. Nature **549**, 43–47 (2017). https://doi.org/10.1038/nature23655
87. Deviani, R.: The application of fully homomorphic encryption on XGBoost based multiclass classification. JIEET (J. Inf. Eng. Educ. Technol.) **7**(1), 49–58 (2023). https://doi.org/10.26740/jieet.v7n1.p49-58
88. Alagic, G., Dulek, Y., Schaffner, C., Speelman, F.: Quantum fully homomorphic encryption with verification. In: Takagi, T., Peyrin, T. (eds.) Advances in cryptology—ASIACRYPT 2017, pp. 438–467. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70694-8_16
89. Maqousi, A., Alauthman, M., Almomani, A.: Homomorphic encryption enabling computation on encrypted data for secure cloud computing. In: innovations in modern cryptography, pp. 215–240. IGI Global (2024). https://doi.org/10.4018/979-8-3693-5330-1.ch009
90. Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., Zhang, X.: Trojaning attack on neural networks. In: Proceedings 2018 Network and Distributed System Security Symposium. Internet Society, San Diego, CA (2018). https://doi.org/10.14722/ndss.2018.23291
91. Marcano, N.J.H., Moller, M., Hansen, S., Jacobsen, R.H.: On fully homomorphic encryption for privacy-preserving deep learning. In: 2019 IEEE Globecom Workshops (GC Wkshps), pp. 1–6 (2019). https://doi.org/10.1109/GCWkshps45667.2019.9024625
92. Beauchamp, T.L., Childress, J.F.: Principles of biomedical ethics, 8th edn. Oxford University Press, New York, NY (2019). https://global.oup.com/academic/product/principles-of-biomedical-ethics-9780190640873
93. World Medical Association: Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA **310**(20), 2191–2194 (2013). https://doi.org/10.1001/jama.2013.281053
94. Floridi, L., Cowls, J., Beltrametti, M., Chiarello, F.: Ai4people—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
95. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
96. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**(2), 1–21 (2016). https://doi.org/10.1177/2053951716679679
97. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**, 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5
98. Ettore Canonici, F.C.: QUBO-based SVM for credit card fraud detection on a real QPU. arXiv preprint arXiv:2409.11876 (2024)
99. Ettore Canonici, F.C.: Privacy-preserving neutral atom-based quantum classifier towards real healthcare applications. arXiv preprint arXiv:arXiv:2505.04570 (2025)
100. Dulek, Y., Schaffner, C., Speelman, F.: Quantum homomorphic encryption for polynomial-sized circuits. In: Robshaw, M., Katz, J. (eds.) Advances in cryptology—CRYPTO 2016, Part III. Lecture Notes in Computer Science, vol. 9816, pp. 3–32. Springer, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53015-3_1
101. Alagic, G., Apon, D., Cooper, D., Dang, Q., Dang, T., Kelsey, J., Lichtinger, J., Liu, Y.-K., Miller, C., Moody, D., Peralta, R., Perlner, R., Robinson, A., Smith-Tone, D.: Status report on the third round of the nist post-quantum cryptography standardization process. NIST interagency/internal report (NISTIR) 8413-upd1, National Institute of Standards and Technology (NIST), Gaithersburg, MD (2022). https://doi.org/10.6028/NIST.IR.8413-upd1
102. Limengnan Zhou, H.W.: Watermarking quantum neural networks based on sample grouped and paired training (2025). arxiv:2506.12675v1
103. European Commission: A coordinated implementation roadmap for the transition to post-quantum cryptography|shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/coordinated-implementation-roadmap-transition-post-quantum-cryptography