Trustworthy AI Systems Through Lenses of Post-Quantum Security and Privacy-Enhancing Techniques

Saleh Darzi, **Kiarash Sedghighadikolaei**, **Saif Eddine Nouma, Attila A. Yavuz**
{salehdarzi, kiarashs, saifeddinenouma, attilaayavuz}@usf.edu
Bellini College of AI, Cybersecurity, and Computing
Applied Cryptography Research Laboratory, 3720 Spectrum Blvd, Tampa, FL, 33612

Abstract

The rapid integration of Artificial Intelligence (AI) across sectors such as healthcare and finance has amplified concerns around privacy, integrity, and long-term security, particularly in outsourced AI-as-a-Service (AIaaS) environments. As quantum computing threatens classical cryptographic protections, ensuring *Post-Quantum (PQ)-Secure Trustworthy AI* has become imperative. This study explores a holistic framework for securing AI systems across training and deployment phases by leveraging PQ-resilient cryptographic and privacy-enhancing technologies (PETs), including Homomorphic Encryption (HE), Functional Encryption (FE), Multi-Party Computation (MPC), Differential Privacy (DP), Trusted Execution Environments (TEEs), and Zero-Knowledge Proofs (ZKPs). Each technique offers complementary strengths, ranging from secure computation and privacy-preserving inference to verifiable model integrity, yet varies in practicality, scalability, and PQ-readiness. We evaluate their utility, integration challenges, and trade-offs, advocating for hybrid designs (e.g., FL+MPC, HE+FL) to optimize security and performance. The study concludes that building PQ-secure, privacy-centric AI requires strategic selection and deployment of tailored tools across the AI pipeline to balance robustness with real-world feasibility.

Keywords: Trustworthy AI, Post Quantum Security, Machine Learning, Security and Privacy

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) enable machines to perform tasks autonomously, often surpassing human efficiency. This capability has driven their adoption across critical sectors such as healthcare, finance, and transportation. With the rise of AI-as-a-Service (AIaaS) platforms like AWS and tools like ChatGPT, AI has become widely accessible, particularly for startups and small enterprises. However, the rapid deployment of AI also introduces serious concerns around privacy, integrity, and long-term security. These risks underscore the need for ***Trustworthy AI***, systems that remain secure, reliable, and verifiable even in adversarial settings [1]. This is particularly vital in high-stakes domains like healthcare, where AI applications like Radiomics support tasks like tumor detection from medical imaging. In such contexts, breaches in data confidentiality or inference integrity can be life-threatening. Moreover, outsourcing AI services to third-party clouds exposes training datasets and models to additional vulnerabilities, necessitating the adoption of Privacy-Enhancing Technologies (PETs) throughout the AI lifecycle. Trustworthy AI hinges on two pillars: (1) **Training Security**, which ensures model integrity, protects user data (e.g., medical records, financial data), and defends against model theft or poisoning; and (2) **Deployment Security**, which safeguards against adversarial queries, model inversion, and extraction attacks that can leak or replicate sensitive models [4,7].

Post-Quantum Security for Trustworthy AI

The advent of quantum computing threatens to break classical cryptographic schemes, endangering the long-term confidentiality and robustness of AI systems [2]. For instance, ECC-based encryption used in AI-driven healthcare can be broken, risking patient privacy and diagnostic reliability. To mitigate these threats, a global transition to post-quantum (PQ) cryptographic solutions is underway, led by efforts such as the NIST standardization initiative [5]. However, integrating PQ security into AI remains an open challenge, which this study aims to address.

We focus on the concept of **PQ-Secure Trustworthy AI**, a framework for securing AI across cloud, distributed, and AIaaS settings where long-term integrity and confidentiality are paramount. Our study investigates PQ-secure techniques spanning algorithmic, statistical, and hardware-based domains, analyzing their applicability to both AI training and deployment. We identify critical gaps in current practices, examine integration and performance trade-offs, and outline directions for advancing PQ-resilient AI infrastructures. A qualitative comparison of these methods is also provided to guide future research and development toward scalable, secure, and verifiable AI systems.

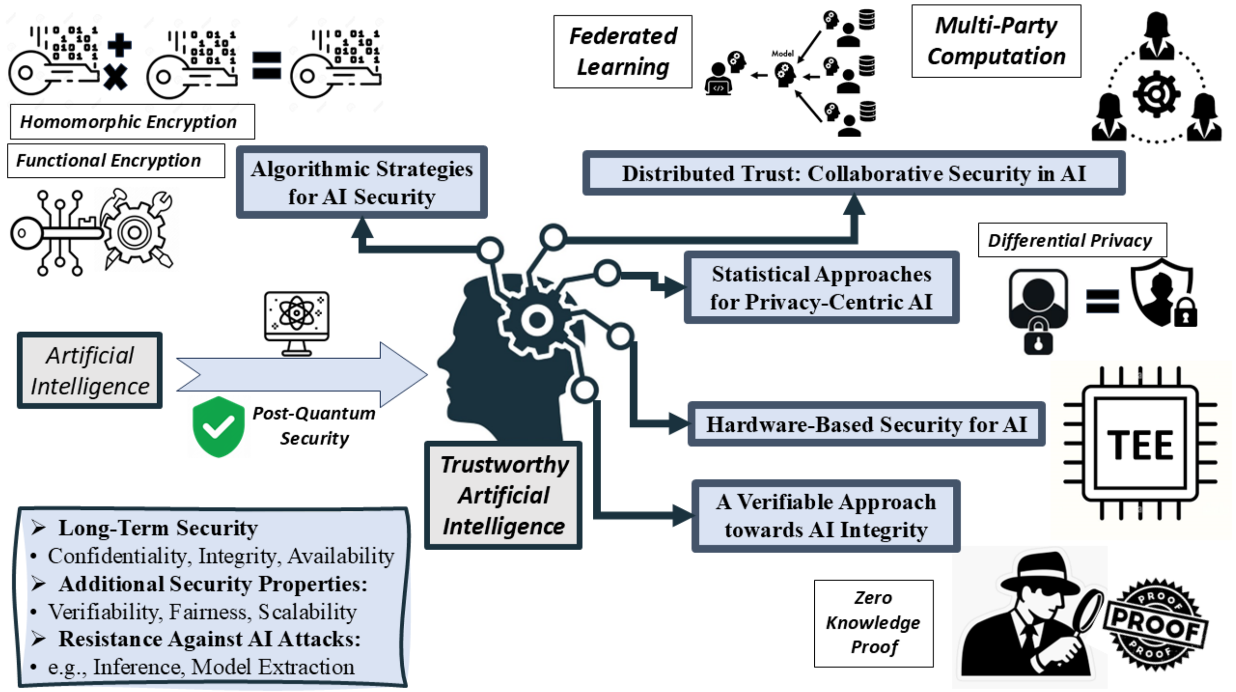
Homomorphic Encryption
Functional Encryption
Algorithmic Strategies for AI Security
Federated Learning
Model
Multi-Party Computation
Distributed Trust: Collaborative Security in AI
Differential Privacy
Statistical Approaches for Privacy-Centric AI
Artificial Intelligence
Post-Quantum Security
Hardware-Based Security for AI
TEE
Trustworthy Artificial Intelligence
A Verifiable Approach towards AI Integrity
Zero Knowledge Proof
PROOF
➢ Long-Term Security
• Confidentiality, Integrity, Availability
➢ Additional Security Properties:
• Verifiability, Fairness, Scalability
➢ Resistance Against AI Attacks:
• e.g., Inference, Model Extraction

Figure 1. A high-level taxonomy of PQ-secure techniques for Trustworthy AI

PQ-Secure AI Training & Deployment

Algorithmic Approaches for AI Security: One set of approaches to secure AI relies on encryption-based techniques with provable security, notably Homomorphic Encryption (HE) and Functional Encryption (FE). Homomorphic Encryption, and specifically Fully Homomorphic Encryption (FHE), enables computations on encrypted data without the need for decryption, ensuring end-to-end confidentiality of both user data and models. While FHE-based training remains computationally intensive and less accurate, it is widely used for privacy-preserving outsourcing and secure aggregation, especially in federated learning scenarios. FHE is most beneficial during the deployment phase, particularly for AIaaS and cloud environments, offering robust protection in untrusted settings [10]. Most practical FHE schemes are built on lattice-based problems, which provide PQ security. Continuous improvements, such as batching, parallelization, and hardware acceleration, are enhancing FHE's practicality. In addition to HE, FE [6] allows for the decryption of specific function outputs, enabling selective data disclosure in settings such as cloud-based healthcare. While FE ensures privacy through access control, it operates under an honest-but-curious model and may risk leakage of information through plaintext outputs. Additionally, FE schemes face PQ-related efficiency trade-offs, limiting their scalability for large datasets.

Distributed Trust: Collaborative Security in AI: Distributed trust in AI security is achieved through collaborative architectures like Federated Learning (FL) and cryptographic techniques such as Multi-Party Computation (MPC) [11]. FL enables decentralized model training by allowing participants to retain their raw data locally, enhancing privacy, fault tolerance, and regulatory compliance (e.g., HIPAA). While FL relies on centralized aggregation, PQ security can be achieved by integrating secure aggregation methods such as HE, MPC, TEE, or DP. MPC complements FL by enabling multiple parties to jointly compute functions over their private data without revealing it, using primitives like HE, garbled circuits, oblivious transfer, and secret sharing. During training, MPC supports secure aggregation across institutions and outsourcing in untrusted environments. In deployment, MPC enables privacy-preserving inference without trusted intermediaries, safeguarding sensitive queries across sectors like banking and healthcare. PQ-secure MPC schemes, particularly those utilizing NIST-standardized primitives and TEE-based secure function evaluations, provide scalable and resilient security frameworks suitable for collaborative AI in the PQ era.

Statistical Approaches for Privacy-Centric AI: Differential Privacy (DP) is a statistical technique that protects individual data privacy by introducing calibrated noise, making it difficult to distinguish between similar datasets. Though not based on cryptographic hardness assumptions, DP is inherently resistant to quantum attacks and is often combined with PQ-secure methods like HE, FE, and MPC to enhance system-level privacy. In AI training, DP is applied either centrally, by adding noise to objective functions or aggregated updates, or locally, where users inject noise directly into their data or labels to prevent leakage [9]. During deployment, DP secures inference results and user queries in real-time systems (e.g., recommendation engines, fraud detection), making it a practical solution for companies aiming to protect user data and resist inference attacks. An emerging extension, Quantum-DP (QDP), introduces quantum noise and leverages quantum computation to offer verifiable privacy guarantees, even with current NISQ devices. Beyond privacy, DP supports model fairness, reduces overfitting, and improves stability, reinforcing its role as a foundational tool in privacy-centric and PQ resilient AI.

Hardware-Based Security for AI: Trusted Execution Environments (TEEs) offer hardware-based security for AI by enabling confidential computation within isolated, tamper-resistant memory regions, protecting both data and model operations in untrusted centralized or distributed settings. Compared to cryptographic methods like HE, FE, and MPC, TEEs reduce computational overhead while providing stronger security guarantees. During training, TEEs ensure data and model confidentiality by isolating computation, enabling secure collaboration in sensitive domains such as pharmaceuticals or autonomous vehicles without exposing proprietary information. In deployment, TEEs preserve model privacy and prevent theft while delivering inference results securely, applicable in sectors like healthcare and finance. Although TEEs themselves are not inherently PQ secure, combining them with NIST-standardized PQC schemes during AI training and inference enables PQ security guarantees [12], making TEEs a practical and scalable solution for secure AI in the PQ era.

A Verifiable Approach Towards AI Integrity: Ensuring AI integrity is vital in outsourced or collaborative environments where malicious behavior by users or model owners can compromise outcomes, especially in sensitive domains like healthcare and finance. Verifiability is essential to guarantee correct model training and inference, prevent data poisoning, and ensure regulatory compliance. Zero-Knowledge Proofs (ZKPs) offer a compelling solution by enabling privacy-preserving verification of computation correctness and output legitimacy [8]. While traditionally built on ECC for efficiency, PQ-secure ZKPs based on symmetric, lattice, or multivariate cryptography extend this assurance into the PQ era. However, practical challenges, such as large proof sizes, high computational costs, and limited scalability, must be addressed for widespread deployment in AI systems.

Takeaways and Future Directions

As AI adoption accelerates across sectors such as healthcare and finance, driven by AIaaS and increasing data complexity, ensuring privacy and PQ security in outsourced training and inference has become increasingly essential. Quantum computing introduces new threats, requiring the evolution of PETs to remain effective. While a unified solution for securing the full AI pipeline is appealing, it often proves impractical due to computational overhead. For instance, HE is best suited for inference, and ZKPs struggle with floating-point operations due to their integer-based structure and high storage demands. Instead, selecting and applying the appropriate technique at each stage is crucial. Hybrid designs (e.g., MPC+FL or HE+FL) offer promising trade-offs, enabling secure, efficient AI workflows by combining the strengths of multiple PETs. These strategies enable flexibility and scalability while meeting diverse security and performance requirements. A detailed breakdown of these takeaways and future directions is provided in the Learning Objective section.

AI Systems with PQ-security and Privacy Guarantee: The growing adoption of AI across industries such as finance and healthcare, fueled by AIaaS, coupled with the increasing volume of data and the complexity of AI models, has driven a shift toward outsourced computation for model training and inference. In these outsourced models, ensuring the privacy and security of both our data and users' data is crucial. Furthermore, the rise of quantum computing presents new security challenges, prompting us to adapt and strengthen our PETs to withstand its increased computational power.

Not One Approach Fits All: Choosing the Best for Each Stage in Securing AI Pipelines: Using a single approach to secure the entire AI pipeline, from training to inference, may seem ideal in theory, but the computational overhead often hinders practical adoption for meeting privacy and PQ-security needs. For example, HE is more suited for inference tasks due to its high computational costs, while floating-point models face difficulties with ZKPs, which typically work over integers, resulting in significant proof generation overhead, including the need for large storage. Therefore, it's essential to carefully choose and apply each security tool at the right stage of the pipeline, considering its overhead and effectiveness for industry requirements.

Hybrid Approaches for Optimizing Security and Performance: Hybrid approaches, such as MPC+FL or HE+FL, have proven effective in optimizing both security and performance, often outperforming single-approach systems like HE-only. For example, HE, while not ideal for the entire training cycle, can be combined with FL for secure execution. FL ensures data privacy by storing data locally, while HE facilitates secure central aggregation of model updates, enhancing both privacy and efficiency. In cases where centralized computation is a vulnerability, MPC can be used to aggregate FL model updates securely. Overall, hybrid PETs offer the flexibility to balance security and performance for specific application needs.

Final Notes**:** PQ-security and privacy are two critical requirements for building trustworthy AI systems. As these requirements become increasingly complex with the evolution of technology, it becomes essential to adopt strategies that balance security, performance, and practicality. This ensures that AI systems can be deployed safely and efficiently in real-world applications. Careful planning, thorough evaluation, and gradual implementation are vital to addressing these challenges, enabling a smooth transition toward more secure and reliable AI systems.

References

1. Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, and Jiquan Pei. 2023. Trustworthy AI: From principles to practices. Comput. Surveys 55, 9 (2023), 1–46.

2. Saleh Darzi, Attila A. Yavuz, Rouzbeh Behnia. Post-Quantum Security for Trustworthy Artificial Intelligence: An Emerging Frontier. Authorea Preprints, (2024).

3. Kiarash Sedghighadikolaei and Attila A Yavuz. 2024. Privacy-preserving and trustworthy deep learning for medical imaging. *arXiv preprint arXiv:2407.00538* (2024).

4. Muhammad Tayyab, Mohsen Marjani, NZ Jhanjhi, and Ibrahim Abaker Targio Hashem. 2023. A comprehensive review on deep learning algorithms: Security and privacy issues. Computers & Security 131 (2023), 103297.

5. Saleh Darzi, Kasra Ahmadi, Saeed Aghapour, Attila Altay Yavuz, and Mehran Mozaffari Kermani. 2023. Envisioning the future of cyber security in post-quantum era: A survey on pq standardization, applications, challenges, and opportunities. arXiv preprint arXiv:2310.12037 (2023).

6. Dan Boneh, Amit Sahai, and Brent Waters. 2011. Functional encryption: Definitions and challenges. In Theory of Cryptography: 8th Theory of Cryptography Conference, TCC 2011, Proceedings 8. Springer, 253–273.

7. Kha Dinh Duy, Taehyun Noh, Siwon Huh, and Hojoon Lee. 2021. Confidential machine learning computation in untrusted environments: A systems security perspective. IEEE Access 9 (2021), 168656–168677.

8. ZKProof Standards. Retrieved Nov. 2024 from https://zkproof.org/ 2024.

9. Ahmed El Ouadrhiri and Ahmed Abdelhadi. 2022. Differential privacy for deep and federated learning: A survey. IEEE Access 10 (2022), 22359–22380.

10. Runhua Xu, James BD Joshi, and Chao Li. 2019. Cryptonn: Training neural networks over encrypted data. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). IEEE, 1199–1209.

11. Kiarash Sedghighadikolaei and Attila Altay Yavuz. 2023. A comprehensive survey of threshold signatures: NIST standards, post-quantum cryptography, exotic techniques, and real-world applications. arXiv preprint arXiv:2311.05514 (2023).

12. Yavuz, Attila A., Saif E. Nouma, Thang Hoang, Duncan Earl, and Scott Packard. "Distributed cyber-infrastructures and artificial intelligence in hybrid post-quantum era." In 2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), pp. 29-38. IEEE, 2022.

Presenter Bio: Attila Altay Yavuz is an Associate Professor at the Bellini College of Artificial Intelligence, Cybersecurity, and Computing at the University of South Florida (USF), where he also directs the Applied Cryptography Research Laboratory. Previously, he was an Assistant Professor at Oregon State University (2014–2018) and USF (2018–2021), following his role as a research scientist at the Robert Bosch Research and Technology Center North America (2011–2014). He holds a Ph.D. in Computer Science from North Carolina State University (2011) and an M.S. from Bogazici University (2006). Dr. Yavuz's broad research interests center on designing, analyzing, and deploying cryptographic techniques to strengthen the security of computer systems and next-generation networks. His work has been recognized with numerous honors, including the NSF CAREER Award, multiple research awards from Bosch (five) and Cisco (four), three USF Excellence in Research Awards, several major federal grants, and numerous best paper awards. His research leadership extends to editorial board service (e.g., IEEE TDSC) and organizing roles in major conferences (e.g., ACM CCS). His work encompasses 115 peer-reviewed publications in top-tier venues (e.g., Usenix, NDSS, CCS, IEEE TIFS), patents, and technology transfers to industry partners, particularly in searchable encryption and intra-vehicular network security, impacting tens of millions of users worldwide. He is a Senior Member of the IEEE, the National Academy of Inventors, and ACM.

Speaking Experiences: Associate Professor Dr. Attila A. Yavuz has delivered numerous presentations and talks at flagship venues (e.g., PETs, CCS), the National Science Foundation (NSF) as an invited Principal Investigator, and prestigious talks for institutions such as SVCC, companies (e.g., Robert Bosch), and several universities (e.g., Oregon State University, University of Pittsburgh, USF) in various capacities. Among many, some selected speaking experiences:

* National Science Foundation, US-Taiwan Research Initiative, "Unleashing and Advancing Secure Computation for NextGen Networks in the Post-Quantum and Artificial Intelligence Era" (2025)

* Invited Research Seminar, TSF Technology Seminars, "Envisioning the Future of NextGen Networked Systems and Digital Signatures" (2024)

* Distinguished Research Award from Silicon Valley Cybersecurity Institute, "Energy-Aware Digital Signatures for Embedded Medical Devices" (2020).