


# Chapter 11

## Explainable and Transparent AI Architectures

**Deepak Gupta**

 <http://orcid.org/0000-0003-3929-1362>

*Institute of Technology and  
Management, Gwalior, India*


**Abduraimova Nigora**

*Termez University of Economics and  
Service, Termez, Uzbekistan*

**Gulkhayo Gulkhayo**

*Mamun University, Khiva, Uzbekistan*

**Ergashev Nuriddin**

 <http://orcid.org/0000-0002-8274-6193>

*Karshi State Technical University,  
Karshi, Uzbekistan*

**Shokhzod Karimov**

*Tashkent State University of Economics,  
Uzbekistan*

**Mamatkhujayev Otabek**

*Alfraganus University, Tashkent,  
Uzbekistan*

**Seitnazarov Kuanishbay**

*Nukus State Pedagogical Institute,  
Uzbekistan*

### ABSTRACT

*Explainability and transparency have emerged as foundational pillars in the secure deployment of artificial intelligence (AI) systems, especially large language models (LLMs). This chapter examines the evolving landscape of explainable AI (XAI) architectures through the lens of cybersecurity, adversarial robustness, and regulatory compliance. The authors survey core XAI methodologies—including LIME, SHAP, mechanistic interpretability, attention attribution, and causal tracing—evaluating their effectiveness against adversarial threats such as jailbreaking, prompt injection, data poisoning, and hallucination exploitation. The dual nature of XAI is critically*

DOI: 10.4018/979-8-3373-8252-4.ch011

*examined: while transparency mechanisms bolster defense and trust, they simultaneously introduce novel attack surfaces that adversaries can exploit to subvert explanation systems.*

## 1. INTRODUCTION

The rapid integration of artificial intelligence into critical systems—spanning healthcare diagnostics, financial decision-making, legal reasoning, and national security—has elevated the stakes of AI transparency to an unprecedented degree. As large language models (LLMs) such as GPT-4, Claude, and open-source equivalents become embedded in operational pipelines, the question of whether we can trust, audit, and understand their decisions is no longer merely academic. It is a question of societal and systemic risk that demands rigorous, interdisciplinary inquiry at the intersection of machine learning, information security, and governance.

Explainable artificial intelligence (XAI) refers to a broad family of techniques and methodologies aimed at rendering AI model decisions interpretable to human stakeholders (Danilevsky et al., 2021; Hosain et al., 2024). In the context of LLMs, explainability operates at multiple levels: from local explanations of individual predictions to global characterizations of model behavior, and from post-hoc rationalization to intrinsic interpretability embedded within model architectures (Singh et al., 2024). The distinction between these levels is not merely taxonomic—it carries profound implications for the types of adversarial threats that can be detected, the forms of regulatory compliance that can be demonstrated, and the degree to which human oversight can be meaningfully exercised over high-stakes AI decisions.

A defining tension runs through the contemporary XAI landscape: the very mechanisms designed to make AI systems more understandable can themselves become vectors of attack. Slack et al. (2020) demonstrated that LIME and SHAP—two of the most widely deployed post-hoc explanation methods—can be systematically fooled by adversaries who construct inputs that appear benign to the explanation method while enabling the underlying model to behave discriminatorily or maliciously. This “scaffolding attack” paradigm has catalyzed a new subfield examining adversarial robustness not merely of AI predictions, but of AI explanations themselves (Baniecki & Biecek, 2024; Vadillo et al., 2025). The implications are far-reaching: if auditors and regulators rely on explanation systems that can be strategically manipulated, then compliance regimes built on those systems may provide a false sense of security.

Simultaneously, the emergence of LLM-specific attack vectors—including jailbreaking (Chao et al., 2023; Shen et al., 2024), prompt injection (Perez & Ribeiro, 2022; Greshake et al., 2023), training data poisoning (Cinà et al., 2023), and hallucination exploitation (Huang et al., 2024)—demands that XAI methodologies

26 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/chapter/explainable-and-transparent-ai-architectures/408788](http://www.igi-global.com/chapter/explainable-and-transparent-ai-architectures/408788)

## Related Content

---

### Data Sensitivity and Privacy in Connected Health Ecosystems

Olatunde Ayomide Olasehan (2026). *Threat Intelligence and Cloud Trust Models for Healthcare Security* (pp. 73-118).

[www.irma-international.org/chapter/data-sensitivity-and-privacy-in-connected-health-ecosystems/392350](http://www.irma-international.org/chapter/data-sensitivity-and-privacy-in-connected-health-ecosystems/392350)

### Convergence Analysis for Identification of Multivariable Delayed Systems

Yamna Ghouland Naoufel Zitouni (2023). *Applications of Encryption and Watermarking for Information Security* (pp. 151-162).

[www.irma-international.org/chapter/convergence-analysis-for-identification-of-multivariable-delayed-systems/320950](http://www.irma-international.org/chapter/convergence-analysis-for-identification-of-multivariable-delayed-systems/320950)

### The Insider Threat Landscape and the FinTech Sector: Attacks, Defenses, and Emerging Challenges

Zainab Abaid, Ahsan Saadatand Baria Mubashar Mirza (2023). *Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications* (pp. 65-90).

[www.irma-international.org/chapter/the-insider-threat-landscape-and-the-fintech-sector/314075](http://www.irma-international.org/chapter/the-insider-threat-landscape-and-the-fintech-sector/314075)

### An Iterative CrowWhale-Based Optimization Model for Energy-Aware Multicast Routing in IoT

Dipali K. Shende, Yogesh S. Angaland S.C. Patil. (2022). *International Journal of Information Security and Privacy* (pp. 1-24).

[www.irma-international.org/article/an-iterative-crowwhale-based-optimization-model-for-energy-aware-multicast-routing-in-iot/300317](http://www.irma-international.org/article/an-iterative-crowwhale-based-optimization-model-for-energy-aware-multicast-routing-in-iot/300317)

### Neural Network-Based Approach for Detection and Mitigation of DDoS Attacks in SDN Environments

Oussama Hannacheand Mohamed Chaouki Batouche (2020). *International Journal of Information Security and Privacy* (pp. 50-71).

[www.irma-international.org/article/neural-network-based-approach-for-detection-and-mitigation-of-ddos-attacks-in-sdn-environments/256568](http://www.irma-international.org/article/neural-network-based-approach-for-detection-and-mitigation-of-ddos-attacks-in-sdn-environments/256568)