



## Chapter 2

# Explainable AI for Cybersecurity

**Siva Raja Sindiramutty**  
Taylor's University, Malaysia


**Abdalla Hassan Gharib**  
Zanzibar University, Tanzania

**Chong Eng Tan**  
 <https://orcid.org/0000-0002-3990-3501>  
Universiti Malaysia Sarawak, Malaysia

**Amaranadha Reddy Manchuri**  
 <https://orcid.org/0000-0002-3873-0469>  
Kyungpook National University, South Korea

**Sei Ping Lau**  
Universiti Malaysia Sarawak, Malaysia

**Navid Ali Khan**  
Taylor's University, Malaysia

**Rajan Thangaveloo**  
 <https://orcid.org/0000-0002-0402-6019>  
University Malaysia Sarawak, Malaysia

**Wee jing Tee**  
Taylor's University, Malaysia

**Lalitha Muniandy**  
Tunku Abdul Rahman University of Management and Technology, Malaysia

### ABSTRACT

*In recent years, the utilization of AI in the field of cybersecurity has become more widespread. Black-box AI models pose a significant challenge in terms of interpretability and transparency, which is one of the major drawbacks of AI-based systems. This chapter explores explainable AI (XAI) techniques as a solution to these challenges and discusses their application in cybersecurity. The chapter begins with an explanation of AI in cybersecurity, including the types of AI commonly utilized, such as DL, ML, and NLP, and their applications in cybersecurity, such as intrusion detection, malware analysis, and vulnerability assessment. The chapter then highlights the challenges with black-box AI, including difficulty identifying and resolving errors, the lack of transparency, and the inability to understand the decision-making process. The chapter then delves into XAI techniques for cybersecurity solutions, including interpretable machine-learning models, rule-based systems, and model explanation techniques.*

DOI: 10.4018/978-1-6684-6361-1.ch002

## INTRODUCTION

The emerging field of XAI aims to create ML and DL algorithms that are transparent and explainable. In recent years, the fast growth of AI and its associated technologies have led to significant improvements in various fields, including finance, healthcare, and transportation. Despite the efforts to make AI systems more transparent and explainable, the increasing complexity and sophistication of these systems have made them more challenging to comprehend and interpret. The lack of accountability and transparency in AI systems, especially in crucial applications where their decisions can lead to significant outcomes, has raised concerns. XAI strives to tackle these concerns by creating AI systems that can offer explicit and comprehensible justifications for their decisions and actions. By doing so, XAI can improve the trustworthiness, accountability, and reliability of AI systems, making them more accessible to users and stakeholders. XAI includes various approaches such as rule-based systems, model-based systems, and post-hoc explanations. Rule-based systems depend on a predefined set of rules or heuristics to arrive at decisions. These rules are explicitly defined and can be easily understood by humans. Model-based systems, on the other hand, use statistical models to make decisions, and the explanations provided are based on the model's parameters and assumptions. Post-hoc explanations, on the other hand, entail analyzing an AI model's decision-making process after the event to recognize the elements that influenced the outcome (Adadi, & Berrada 2018; Gilpin et al., 2018; Sujatha et al., 2022).

Research in XAI has gained significant momentum in recent years, with numerous studies and initiatives exploring various aspects of the field. Doshi-Velez and Kim (2017) introduced a classification system for XAI, which groups explanations into various categories, such as algorithmic transparency, interpretable models, and interactive explanations. Further research has concentrated on devising particular methods for XAI, such as local interpretable model-agnostic explanations (LIME) proposed by Ribeiro, Singh, and Guestrin (2016), and counterfactual explanations (Wachter, Mittelstadt, & Russell, 2018). XAI has several potential applications in various domains, including healthcare, finance, and defence. In healthcare, XAI can help medical professionals make better decisions by providing clear and understandable explanations for diagnoses and treatment plans. In finance, XAI can help investors and regulators understand the factors that influence financial decisions and mitigate the risks associated with automated trading systems. In defence, XAI can improve the transparency and accountability of autonomous weapons systems, ensuring that they are used ethically and in compliance with international laws and regulations (Guidotti et al., 2018). Despite the significant progress in XAI research, several challenges still need to be addressed. One major challenge is the trade-off between transparency and accuracy. In some cases, the most accurate AI systems are also the least interpretable, making it difficult to understand how they arrived at their decisions. Another challenge is the need for more standardized evaluation metrics and benchmarks for XAI systems. Currently, there is no consensus on the best evaluation metrics for XAI, making it challenging to compare different systems and techniques. To sum up, XAI is a developing area of research that seeks to produce transparent and interpretable AI systems. XAI can have considerable potential in numerous fields, such as healthcare, finance, and defence. However, several challenges need to be addressed, including the trade-off between transparency and accuracy and the need for more standardized evaluation metrics and benchmarks.

65 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-ai-for-cybersecurity/336871](http://www.igi-global.com/chapter/explainable-ai-for-cybersecurity/336871)

## Related Content

---

### Information Technology Service Management and Opportunities for Information Systems Curricula

Sue Conger (2009). *International Journal of Information Systems in the Service Sector* (pp. 58-68).

[www.irma-international.org/article/information-technology-service-management-opportunities/2528](http://www.irma-international.org/article/information-technology-service-management-opportunities/2528)

### Inclusive Democracy in the Digital Age: The Case of Brasil Participativo Using Decidim Platform

Claucia Piccoli Faganello and Edimara Mezzomo Luciano (2024). *Emerging Developments and Technologies in Digital Government* (pp. 333-354).

[www.irma-international.org/chapter/inclusive-democracy-in-the-digital-age/344623](http://www.irma-international.org/chapter/inclusive-democracy-in-the-digital-age/344623)

### Enhanced Trust Path between Two Entities in Cloud Computing Environment

Usha Divakarla and K. Chandrasekaran (2016). *International Journal of Cloud Applications and Computing* (pp. 15-31).

[www.irma-international.org/article/enhanced-trust-path-between-two-entities-in-cloud-computing-environment/159835](http://www.irma-international.org/article/enhanced-trust-path-between-two-entities-in-cloud-computing-environment/159835)

### Service-Oriented Enterprise Engineering: A Modeling Discipline Based on the Viable Systems Approach (vSa) for Strategic Sourcing Decision-Making

Laleh Rafati and Geert Poels (2018). *International Journal of Information Systems in the Service Sector* (pp. 20-40).

[www.irma-international.org/article/service-oriented-enterprise-engineering/206895](http://www.irma-international.org/article/service-oriented-enterprise-engineering/206895)

### Determinants of Goal-Directed Mobile Ticketing Service Adoption Among Internet Users: The Case of Taiwan

Shen-Yao Wang and Ting Lie (2012). *Innovative Mobile Platform Developments for Electronic Services Design and Delivery* (pp. 21-36).

[www.irma-international.org/chapter/determinants-goal-directed-mobile-ticketing/65939](http://www.irma-international.org/chapter/determinants-goal-directed-mobile-ticketing/65939)