# Chapter 4
# Ethical AI in Cybersecurity: Foundations, Challenges, and Evaluation Perspectives

**Shritin Shetty**
http://orcid.org/0009-0008-7181-1405
*Illinois Institute of Technology, USA*

**Md Mahmudul Hasan**
*Illinois Institute of Technology, Ukraine*

## ABSTRACT

*The integration of artificial intelligence into cybersecurity systems introduces significant ethical challenges that require systematic and measurable evaluation. Although prior studies propose lifecycle-based ethical AI frameworks, practical methodologies for benchmarking ethical compliance remain limited. This study presents an evidence-driven analytical framework for evaluating ethical integration in AI-driven cybersecurity research. A corpus of peer-reviewed publications was analyzed using a structured pipeline that employs large language models (LLMs) to extract machine-readable JSON evidence, followed by automated scoring across seven ethical dimensions: transparency, explainability, accountability, human oversight, privacy, data protection, and continuous ethical monitoring. The results reveal substantial variability in ethical compliance, with notable deficiencies in human oversight and post-deployment monitoring. The proposed framework is replicable, auditable, and supports evidence-based ethical governance of AI-enabled cybersecurity platforms.*

DOI: 10.4018/979-8-3373-8252-4.ch004

Copyright © 2026, IGI Global Scientific Publishing. Copying or distributing in print or electronic forms without written permission of IGI Global Scientific Publishing is prohibited. Use of this chapter to train generative artificial intelligence (AI) technologies is expressly prohibited. The publisher reserves all rights to license its use for generative AI training and machine learning model development.

## I. INTRODUCTION

Artificial Intelligence has revolutionised the field of cybersecurity by enabling faster detection of anomalies, automated threat response, and predictive analysis for identifying potential vulnerabilities (Sarker et al., 2021),(Morovat & Panda, 2020). As organizations rely more on AI-driven systems to protect critical digital infrastructure, the ethical and governance challenges associated with the system have become a major concern. The use of autonomous algorithms in cybersecurity introduces potential risks related to fairness, transparency, accountability, and privacy, issues that directly affect trust and reliability in the organizations that depend on them. The growing use of AI systems in cybersecurity poses complex ethical dilemmas. Machine learning models often train on sensitive data and make high stakes decisions without human supervision. A biased or inaccurate detection model could lead to privacy violations or unnecessary restrictions (Vemuri et al., 2023). While existing studies emphasize the value of embedding ethical principles into AI development, many frameworks remain conceptual, lacking guidance for implementation (Dhirani et al., 2024). This gap between ethical goals and practical application creates uncertainty for organizations trying to keep up with changing regulatory standards. Recent global policy efforts highlight the need for ethical governance in AI. The European Union's Artificial Intelligence Act (2024) establishes a risk-based framework that requires transparency, accountability, and human oversight for AI systems used in critical sectors, including cybersecurity (European Commission, 2024). Similarly, the National Institute of Standards and Technology AI Risk Management Framework (2023) provides a lifecycle-based approach to identifying and mitigating AI risks across design, deployment, and monitoring phases (National Institute of Standards and Technology, 2023). The effectiveness of these frameworks really depends on how well organizations can apply their principles into action in technical settings. Researchers highlight that ethical AI should be built on key principles such as beneficence, non-maleficence, justice, autonomy, and explicability, which help ensure that AI systems act in ways that benefit people and society. In a similar way, explains that ethics should be part of the entire AI lifecycle, from collecting data to continuous monitoring and should include clear checkpoints for accountability and human oversight (Leslie, 2019). The baseline paper, Securing Trust: Ethical Considerations in AI for Cybersecurity, identifies transparency, accountability, fairness, and privacy as the main ethical pillars for building trustworthy AI systems (Vemuri et al., 2023). While their framework provides a strong starting point, it mainly stays conceptual and doesn't show how organizations can apply these principles in real cybersecurity systems. Building on that work, this research aims to turn those ethical ideas into a practical, lifecycle-based plan that organizations can use to design and manage AI cybersecurity systems.

16 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/ethical-ai-in-cybersecurity/408781

## Related Content

Combining Elliptic Curve Cryptography and Blockchain Technology to Secure Data Storage in Cloud Environments

Faiza Benmenzerand Rachid Beghdad (2022). *International Journal of Information Security and Privacy (pp. 1-20).*

www.irma-international.org/article/combining-elliptic-curve-cryptography-and-blockchain-technology-to-secure-data-storage-in-cloud-environments/307072

K-Means Cluster-Based Interference Alignment With Adam Optimizer in Convolutional Neural Networks

Tirupathaiah Kanaparthi, Ramesh S.and Ravi Sekhar Yarrabothu (2022). *International Journal of Information Security and Privacy (pp. 1-18).*

www.irma-international.org/article/k-means-cluster-based-interference-alignment-with-adam-optimizer-in-convolutional-neural-networks/308307

Trust and Reputation in Secured AIoT: A Communication and Social Science Perspective on Authentication Ecosystems

Astri Dwi Andriani (2026). *Advanced Approaches for Trust and Identity Management in AIoT Environments (pp. 347-374).*

www.irma-international.org/chapter/trust-and-reputation-in-secured-aiot/411116

Detecting the Risk of Online Harms on People With Social Orientation Impairments: The Role of Automated Affective Content Screening of Neuro-Response Plasticity

Jonathan Bishopand Darren Bellenger (2021). *Handbook of Research on Cyber Crime and Information Privacy (pp. 739-753).*

www.irma-international.org/chapter/detecting-the-risk-of-online-harms-on-people-with-social-orientation-impairments/261753

Child Security in Cyberspace Through Moral Cognition

Satya Prakash, Abhishek Vaish, Natalie Coul, SaravanaKumar G, T.N. Srinidhiand Jayaprasad Botsa (2013). *International Journal of Information Security and Privacy (pp. 16-29).*

www.irma-international.org/article/child-security-cyberspace-through-moral/78527