Benchmarking Current Ethical AI Practices in Cybersecurity: Creating a Practical Evaluation Framework

1st Shritin Shetty
Department of Information Technology and Management Illinois Institute of Technology
Chicago, USA
sshetty20@hawk.illinoistech.edu

2nd Md Mahmudul Hasan
Department of Information Technology and Management
Illinois Institute of Technology
Chicago, USA
hmdmahmudul@hawk.illinoistech.edu

Abstract—The integration of artificial intelligence (AI) into cybersecurity infrastructures has introduced complex ethical challenges that demand systematic scrutiny. While lifecycle-oriented frameworks for ethical AI have been conceptually artic-ulated in prior literature, there remains a conspicuous absence of empirical methodologies to operationalize and benchmark ethical compliance within this domain. This study presents an analytical approach that evaluates a series of peer-reviewed publications related to the implications of AI, ethics, and cybersecurity. Utilizing a structured, evidence-driven pipeline, each paper was subjected to large language model (LLM) assisted extraction of machine-readable JSON evidence, followed by automated scoring across seven rigorously defined ethical dimensions: real-time transparency, explainability, accountability, human oversight, privacy, data protection, and continuous ethical monitoring. The resultant scoring schema yields a ranked assessment of ethical integration in contemporary AI-cybersecurity research. Findings reveal that while a minority of papers achieved comprehen-sive alignment with ethical standards, the majority exhibited substantive deficiencies, particularly in areas Human oversight and continuous ethical monitoring. This work contributes a replicable, auditable framework for ethical evaluation, advancing both scholarly discourse and practical policy formulation in the governance of AI-driven cybersecurity systems.

I. Introduction

Artificial Intelligence (AI) has revolutionised the field of cybersecurity by enabling faster detection of anomalies, auto-mated threat response, and predictive analysis for identifying potential vulnerabilities [5],[20]. 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 learn-ing 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 [2]. While existing studies emphasize the value of embedding ethical principles into AI development, many frameworks remain conceptual, lacking guidance for im-plementation [34]. 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 Intelli-gence Act (2024) establishes a risk-based framework that requires transparency, accountability, and human oversight for AI systems used in critical sectors, including cybersecurity [11]. Similarly, the NIST AI Risk Management Framework (2023) provides a lifecycle-based approach to identifying and mitigating AI risks across design, deployment, and monitoring phases [12]. 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 [6].

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 [2]. 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 eth-ical ideas into a practical, lifecycle-based plan that organiza-tions can use to design and manage AI cybersecurity systems. The proposed framework includes clear checkpoints, roles, and accountability steps at each phase of system development, from initial planning and data collection to model training, deployment, and ongoing monitoring. It also connects with existing efforts like the EU AI Act and the NIST AI Risk Management Framework. By including ethics throughout the AI lifecycle, this study treats ethics as actionable, not just theoretical, making it an essential part of how organizations manage cybersecurity.

II. Literature Review

A. AI in Cybersecurity: Overview

Artificial Intelligence (AI) has become a key component in modern cybersecurity systems due to its ability to process large amounts of data and detect complex patterns that traditional methods often miss. Machine learning models deep learning models are now commonly used to identify anomalies, detect intrusions, and predict potential threats before they happen [20], [5]. According to [20], AI based systems improve re-sponse times and accuracy in identifying malicious activity, helping organizations to automate threat analysis and reduce human error. Similarly [5] highlight that AI-driven cybersecu-rity enhances awareness by continuously monitoring network traffic and recognizing unusual behaviours, making it more effective.

Recent studies also emphasize the growing use of explain-able AI (XAI) in cybersecurity, which focuses on increasing transparency and interpretability of AI models [3]. [3] Points out that while black-box models can achieve high detection accuracy, their lack of interpretability poses challenges for accountability and trust, especially when AI systems make autonomous security decisions. This concern has led to a stronger interest in hybrid approaches that balance model performance with explainability.

AI applications in cybersecurity are increasing beyond de-tection and prevention. The Comprehensive Review of AI's Current Impact and Future Prospects in Cybersecurity (2024) discusses how AI is now being used for incident prediction, behavioural risk assessment, and adaptive threat mitigation [4]. These advancements are pushing cybersecurity towards predictive and proactive defence. However, with increased autonomy and data dependence, new ethical and governance challenges will emerge.

B. Ethical Challenges in AI-Driven Cybersecurity

While AI has strengthened cybersecurity defences, it also brings complex ethical concerns that go beyond technical performance. One such major issue is algorithmic bias, models trained on unbalanced or biased data can unintentionally discriminate. [13] Points out that many AI systems inherit existing social and technical biases from the data used to train them, which can lead to unfair or inaccurate outputs. In cyber-security, such bias could cause false alarms, missed detections, or unequal treatment of certain networks, undermining both efficiency and trust.

Another challenge involves privacy and data protection. AI systems often require access to massive amounts of data to train on effectively. This creates a tension between improving the system and respecting individual privacy rights [2]. Unau-thorized data use or poorly taken training datasets can expose personal or corporate information. Maintaining user privacy must remain as the main goal to AI governance because data misuse can damage respect and compliance with regulations such as the General Data Protection Regulation (GDPR) [15].

Accountability and transparency are equally critical. Or-ganizations should define clear accountability checkpoints across the AI development lifecycle so that every decision can be traced to a responsible perosn or process [6]. Ethical governance frameworks like safety audits in engineering are necessary to monitor AI systems continuously and prevent any harm. However, according to [15] notes, the black-box nature of many deep learning models makes it difficult to understand or justify their outputs, resulting in a lack of explainability in high-risk situations.

The Comprehensive Review of AI's Current Impact and Future Prospects in Cybersecurity [22] states that without interpretability and oversight, AI driven defences could make unclear decisions that are hard to contest. Whittlestone et al. [18] explains that ethical principles alone are not enough, they must be translated into solid practices and tools that help engineers manage accuracy, privacy, and accountability.

The main ethical challenges in AI based cybersecurity include bias, privacy risks, and limited transparency. These issues show that technical performance alone cannot guarantee trust. Addressing them requires structured ethical governance frameworks, an area where most current research remains conceptual.

C. Governance and Regulatory Frameworks

As the ethical concerns around AI in cybersecurity have become more serious, several governments and organizations have started creating policies and frameworks to guide the responsible use of these systems. The goal of these frameworks is to translate ethical values such as fairness, accountability, and transparency into real actions that developers and or-ganizations can follow. Ethical AI should not be treated as something optional or only used after a problem happens, it needs to be built into the design and decision making process from the start [15].

The Organisation for Economic Co-operation and Devel-opment (OECD) AI principles are among the most widely recognized global initiatives guiding ethical AI development

[16]. These principles focus on human-centred values, fairness, transparency, and accountability.

D. Proposed Architecture

They encourage developers and policymakers to make sure that AI systems support social well-being rather than just focusing on efficiency or performance. The OECD guidelines have also inspired other international and national AI strategies that try to balance innovation with ethical responsibility.

The EU AI Act is one of the most detailed laws developed so far to regulate AI use [11]. It classifies AI applications into four risk levels, minimal, limited, high, and unaccept-able risk categories, with cybersecurity tools typically falling under the high risk category. Because of this, they must meet strict requirements such as transparency in data use, clear documentation, and ongoing human oversight. The act gives organizations a legal way to manage risk and ensure accountability across the AI lifecycle. Simply following the rules doesn't always mean a system is ethical, companies also need to build a culture that understands and applies these rules in a meaningful way [10].

In the United States, the National Institute of Standards and Technology (NIST) released the AI Risk Management Framework (AI RMF) in 2023 [12]. This framework focuses on identifying and managing AI risks throughout the system's lifecycle. It gives organizations a more flexible structure to plan, monitor, and improve AI governance while emphasizing human oversight and transparency. Both the EU and NIST frameworks show a growing focus on making AI accountabil-ity measurable.

AI governance should be based on five main principles: beneficence, non-maleficence, justice, autonomy, and explicability as suggested in [9]. These ideas are meant to keep AI development aligned with human values and social good. These governance frameworks represent a shift from voluntary ethical guidelines to more formal and enforceable standards.

Even with all these efforts, most of these frameworks still stay at a high level and don't fully explain how organizations can apply ethics in their daily cybersecurity work. This gap highlights why there's a need for more practical, lifecycle-based models that can help organizations turn these ideas into real actions.

E. Trust, Transparency, and Accountability in AI Systems

Trust is one of the most important factors when it comes to the use of AI in cybersecurity. Since these systems often make decisions automatically and handle large amounts of sensitive data, users and organizations need to know that the outputs are reliable and fair. Building trust depends upon how transparent and accountable the system is throughout its development lifecycle. Trust in AI systems is strongly influenced by users' ability to understand how decisions are made and by the presence of clear processes to prevent and respond to failures [31]. This means organizations must design AI tools that provide clear explanations of how decisions are made and allow humans to intervene when required.

Trust is not just about technical performance but about verifiable claims, organizations must prove that their AI systems are safe, ethical and reliable [8]. This includes documentation, testing and external audits to confirm if systems are behaving correctly. When it comes to AI systems in cybersecurity, these verifications become more important because a single wrong decision can affect data integrity, privacy and security.

Transparency is another key part of this. Transparency is the ability for the people to understand not only what AI system did but why it did that [9]. In cybersecurity, this helps users trust AI generated predictions instead of treating them like black box outputs. Explainability improves collaboration between people and the system, making the AI system easier to understand and correct when errors occur.

Accountability connects trust and transparency by making sure that every decision can be traced to a responsible source. [6] Argues that accountability should be built into every stage of AI development, from data collection to deployment and monitoring, so that mistakes can be identified and sorted quickly. These kinds of structured responsibility ensures that ethical principles are an active part of how an AI system.

Overall, trust, transparency, and accountability work to-gether to form the foundation of ethical AI in cybersecurity [2]. When these values are missing, even the most advanced system can become unreliable and can lose credibility. However, when they are incorporated into every stage development of AI lifecycle, it can become both effective and trustworthy.

F. Gaps in Existing Research

Even though many studies have focused on the ethics and governance of AI, most of them still stop at the conceptual level. Researchers like Hagendorff [13] and Whittlestone et al. [18] mention that a lot of ethical frameworks end up repeating the same ideas without showing clear ways to put them into practice. This is especially true in cybersecurity, where decisions must be made quickly and automatically.

In the baseline paper Securing Trust: Ethical Considerations in AI for Cybersecurity, presents a great foundation for identifying ethical principles, but their framework remains mostly conceptual [2]. It tells us what to focus on, but not how to apply those principles when developing or deploying AI in real cybersecurity setups. Similarly, frameworks such as the EU AI Act (2024) [11] and the NIST AI Risk Management Framework (2023) [12] provides strong policy direction but they don't really break it down into practical steps that are needed for implementation.

Another noticeable gap is in accountability. The need for verifiable claims to support trustworthy AI is suggested but most studies don't go into detail about how that should work in cybersecurity models [8]. There is also limited research on continuous ethical monitoring, how to check that if an AI system acts fairly and transparently once deployed. Ethical evaluation is treated as one time step rather than an ongoing process throughout the AI lifecycle [15], [6].

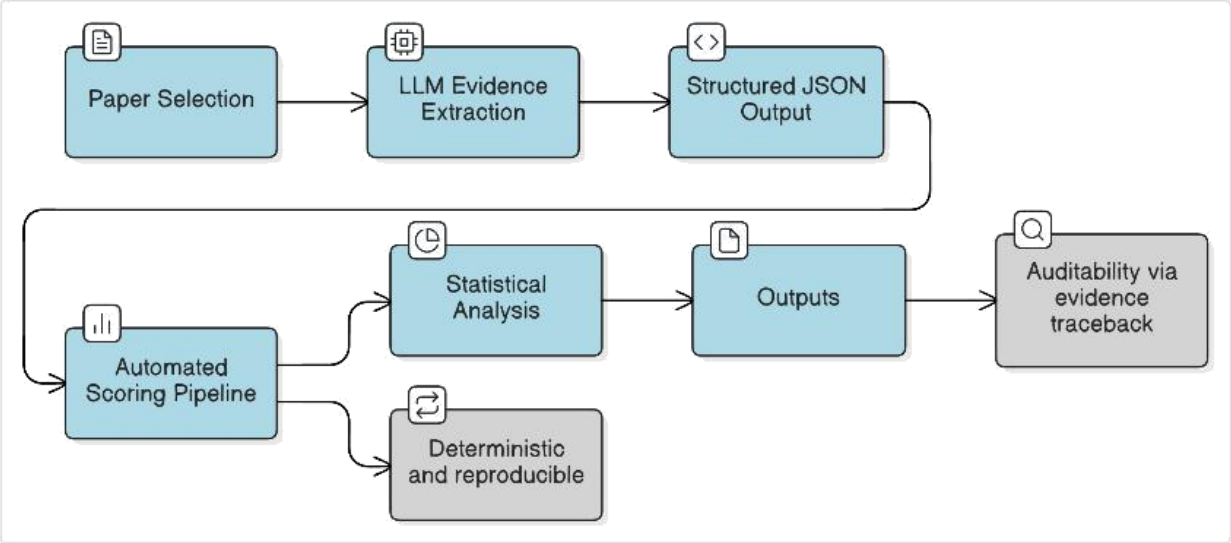

Fig. 1. Proposed Architecture for Ethical AI Evaluation Framework

III. Proposed Methodology

This research advances beyond theoretical frameworks of AI ethics in cybersecurity by developing a structured, evidence-driven methodology to quantitatively assess ethical compli-ance in academic research. Our methodology operationalizes ethical AI principles across cybersecurity contexts through a three-stage pipeline that includes curated literature selection, machine-assisted data extraction, and automated scoring. The framework enables traceability, auditability, and replicability, key features for both academic integrity and practical adoption.

A. Paper Selection and Corpus Construction

We constructed a corpus of 30 academic papers specifically chosen for their relevance to ethical AI applications in cybersecurity. The papers were retrieved from top-tier databases including IEEE Xplore, ACM Digital Library, SpringerLink, and arXiv. Selection criteria mandated that each paper explicitly address the convergence of artificial intelligence, cybersecurity technologies or methods, and ethical principles. This ensures that the corpus reflects interdisciplinary perspectives across technical and ethical domains.

The selection process followed a systematic screening pro-tocol:

- Inclusion Criteria: Papers must (a) be peer-reviewed or reputable preprints, (b) include discussion of both AI and cybersecurity, and (c) explicitly reference ethical concerns or values.
- Exclusion Criteria: Articles focusing exclusively on technical innovations without ethical discourse, or on general AI ethics without cybersecurity relevance, were excluded.

The final corpus encompasses a diverse array of methodologies (qualitative reviews, empirical studies, framework propos-als), application domains (e.g., network intrusion detection, autonomous systems, privacy-preserving AI), and years of publication. This heterogeneity enhances the generalizability and validity of our benchmarking results.

B. Ethical Evaluation Framework

Our ethical evaluation framework is based on seven key cri-teria synthesized from globally recognized AI ethics guidelines such as the OECD AI Principles, AI4People framework, and IEEE Ethically Aligned Design [16], [17], [19].. These criteria reflect ethical imperatives widely accepted in both academic and policy communities:

1) Real-Time Transparency: The extent to which AI systems disclose real-time decision logic or risk indicators.
2) Explainability: Whether the system outputs are inter-pretable and understandable by human users.
3) Accountability: The existence of responsibility attribu-tion mechanisms for AI system behavior.
4) Human Oversight: Provisions for meaningful human control or intervention.
5) Privacy: Mechanisms to ensure personal or sensitive data is protected.
6) Data Protection: Formal safeguards for data storage, transfer, and processing integrity.
7) Continuous Ethical Monitoring: Lifecycle-based gov-ernance that includes periodic reassessment of ethical impacts.

Each of these criteria was defined operationally for consis-tency and was incorporated into a scoring rubric that guided automated evaluations.

C. Evidence Extraction and Structuring

Each selected paper underwent structured analysis using a large language model (LLM)-based prompt system. This system utilized a fixed template to extract evidence aligned with the seven ethical criteria. The extracted information

was formatted into a JSON schema, designed for machine-readability, traceability, and scoring automation.

The standardized JSON template included the following fields:

- Criterion status: Categorical variable (”full”, ”partial”, ”none”) reflecting the level of compliance.
- Evidence text: A direct quote or summary segment from the original text supporting the classification.
- Traceability tags: Metadata pointing to section headers or page numbers for verification.
- Assessor notes and quality scores: Added to provide qualitative insights and confidence levels behind each classification, facilitating transparent interpretation.
- Evidence type: Annotations categorizing the nature of evidence (e.g., normative claim), supporting nuanced scoring decisions.

To maintain consistency, all extractions were performed using the same LLM prompt across all documents. This eliminated variation due to user interpretation or prompt drift and ensured equitable assessment across papers.

D. Automated Scoring Pipeline

Once the structured evidence was extracted and stored as JSON files, the documents were processed by a Python-based scoring engine (score.py). While six ethical criteria were scored on a 15-point scale, one criterion was scored on a 10-point scale and additional weighting logic was applied within the JSON-based scoring pipeline to enable more granular differentiation across studies. This script applied deterministic logic to calculate ethical compliance scores as follows:

- Full Compliance: 15 points
- Partial Compliance: 7 points
- No Compliance: 0 points

Scores across the seven criteria were aggregated to yield a maximum possible score of 100 per paper. The script produced the following outputs:

- Scores.csv: A sortable spreadsheet ranking all 30 papers by total compliance score.
- Summary.md: A report highlighting top-performing pa-pers and summarizing cross-paper trends.
- results/reports.md: Individual markdown files providing criterion-level evaluations and excerpts for each paper.

This approach enables automated benchmarking, supports longitudinal analysis, and facilitates evidence-based discourse.

E. Validation and Reproducibility

: To ensure methodological rigor, all extractions and scor-ings adhered to fixed templates and deterministic rules. The use of JSON structures allows for:

- Auditability: Every score is traceable back to textual evidence.
- Transparency: All decisions can be independently re-viewed by third parties.
- Reproducibility: Given the same input, the system yields identical outputs.

This design ensures that results are not only interpretable but also verifiable, a critical requirement for ethical AI assessment.

F. Methodological Limitations

Despite its strengths, the methodology has certain limitations. LLMs may misclassify nuanced ethical statements or overlook implicit ethical reasoning. Furthermore, while scoring thresholds are rule-based, the categorization between ”full” and ”partial” compliance can occasionally involve judgment calls embedded in prompt design.

To mitigate these limitations, future iterations could be incorporated:

- Human-in-the-loop Validation: Subject matter experts revxiewing borderline classifications.
- Multi-model Extraction: Using multiple LLMs to cross-verify extracted evidence.
- Weighted Criteria: Adjusting scoring logic to reflect domain-specific ethical priorities.

Nonetheless, the proposed methodology represents a signifi-cant advancement in transitioning from theoretical to practical and measurable ethical AI governance in cybersecurity sys-tems.

IV. Analysis and Discussions

This study quantitatively evaluated 30 peer-reviewed AI cybersecurity publications across seven ethical criteria: real-time transparency, explainability, accountability, human oversight, privacy, data protection, and continuous ethics monitoring. Leveraging a novel JSON-based evidence extraction and deterministic scoring pipeline. Using structured JSON-based evidence, automated scoring and manual validation. The scoring incorporates not only the presence of ethical considerations but also the quality and confidence of evidence [1]. All papers included in the evaluation corpus were cited and documented in the reference list, with individual scores derived from the structured evaluation framework described in Section III.

A. Quantitative Evaluation

The mean scores across criteria reflected differing emphasis areas:

- Statistical Results: The mean scores for each ethical dimension demonstrate strong performance in transparency, explainability, and accountability, each averaging above 14 out of 15 across the corpus. Privacy and data pro-tection also show high engagement but with slightly greater variance, suggesting differential emphasis and implementation fidelity. (Fig 4.1)
- Ethical criteria score chart:The provided bar chart visually highlights these strengths and gaps, making clear where ethical attention is focused and where future improvements are most urgently needed. Dimensions such as transparency (bar 0) and explainability (bar 1) show tall, consistent bars, while human oversight (bar 3) and continuous ethical monitoring (bar 6) lag notably in both average and spread. (Fig 4.2)

Criterion	mean	median	variance	range
real_time_transparency	14.71428571	15	0.211640212	1
explainability	14.5	15	0.481481481	2
accountability	14.17857143	15	8.078042328	15
human_oversight	5.535714286	9	21.07275132	10
privacy	13.21428571	15	16.3968254	15
data_protection	13.28571429	14	7.693121693	15
continuous_ethics_monitoring	9.285714286	13	35.76719577	13

Fig. 2. Table 1: Key Statistical Properties of Ethical Scores

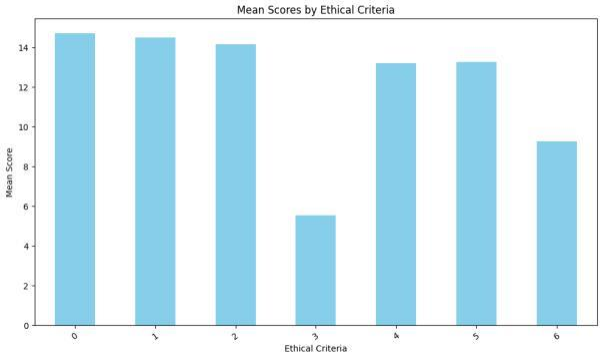

Fig. 3. Table 2: Mean Ethical Compliance Score per Dimension

- Top Performers: The distribution of total scores reveals a broad spectrum of ethical integration, with top-performing papers such as [4] and [2] scoring 98 out of a possible maximum, demonstrating exemplary incorporation of ethical principles. Following closely, other papers such as ”A framework for assessing AI ethics with applications to cybersecurity” [29] and “Ethical AI cybersecurity innovation regulation” scored 96, reflecting similarly high standards.

Table 2

B. Thematic Insights

1. Transparency and Accountability: High scoring papers consistently provide auditable, explainable AI models and strong reporting mechanisms. For example, [30] documents explainable design, while [22] details multi-level transparency, implementing it throughout model development and deploy-ment.

2. Privacy and Data Protection: Concrete measures, including extensive anonymization protocols and compliance with data governance standards, are observed in top scoring papers. “Ethical Considerations in AI-Driven Cybersecurity: Balancing Automation and Human Oversight” and “Securing Trust Ethical Considerations in AI for Cybersecurity” offer strong evidence of lifecycle privacy risk assessment and continuous adjustment to changing regulatory environments.

3. Accountability and Evidence Quality: Papers scoring well in accountability illustrate multi-stakeholder engagement and trackable decision records. Manual scoring notes suggest the importance of good evidence notes and confidence in supporting claims.

4. Fairness, Bias and Societal Impact: Some papers looked closely at fairness and bias, checking if their systems avoided unfair results and noting the real-world effects. Top performing papers, such as ”Comprehensive review of AI in cybersecurity”, explicitly examine sources of bias in both data and models, proposing technical and procedural adjustments to promote fairness across demographic and usage groups. Still, many papers do not dig deep on bias or don’t talk much about the wider social impact, so this an area that needs more work.

C. Gaps Identified

1. Human Oversight: A lot of the research papers either don’t mention how people stay involved in monitoring or decision-making or only mention it briefly. Some assume the AI or automated system will handle everything on its own, but this can create serious risks if the system makes mistakes or encounters something unexpected. Individual report reviews frequently highlight the lack of thorough human in the loop systems, with several papers scoring zero or near zero. For example [26] and related reports point to cases where au-tonomous operation was prioritized without strong intervention protocols, increasing ethical risk.

2. Continuous Ethics Monitoring: Most papers rely on fixed design and do not include ongoing checks or follow-up for ethical issues once their AI systems are launched. Almost none of the papers describe a process for regularly monitoring their systems after deployment. This means that if new problems, biases, or security issues come up as the system is used over time, there may be no way to spot or fix them. Our scoring shows that this area had some of the lowest and most inconsistent results across all the ethical criteria. This shows that continuous ethics monitoring is not yet a standard practice in AI cybersecurity research.

D. Patterns and Comparative Observations

1. Manual Overrides and Evidence Quality In this, both the presence and the quality of evidence were equally important during the scoring process. This means that it wasn’t just about how many claims or notes a paper included, but also about how strong and trustworthy those claims were. This approach made a real difference, papers that gave detailed evidence for their ethical practices, included lots of clear examples, and explained their decision-making step by step, always performed the better than other papers. Having strong and varied notes showed that the researchers understood the complexities of their work and cared about being transparent.

2. Variance Across Dimensions: This means that for some categories, specifically human oversight (keeping humans involved and in charge) and continuous monitoring (watching for ongoing problems), the scores were all over the place. Some papers are really strong at these but others are very weak. This big difference shows that researchers don’t all agree on what’s enough or how these things should be handled. This wide range is important because it means the field is still young in these parts there isn’t a standard for what good oversight or monitoring looks like yet. This can be risky because if people

assume AI is always doing a good job without regular checks, mistakes or biases can go unnoticed for a long time.

E. Implications of These Findings:

These findings suggest that while AI ethics in cybersecurity is improving in key areas like transparency, accountability, and explainability, there are still significant weaknesses with human oversight and keeping systems ethically monitored after deployment. Organizations deploying AI security tools should therefore treat ongoing human review, post deployment audits, and incident reporting as mandatory safeguards rather than optional add ons. Regulators and standards bodies may also need to tighten expectations around continuous monitoring, bias audits, and red team testing so that ethical performance is checked over the full lifecycle, not only at design time. Finally, future research should explore concrete mechanisms, such as governance dashboards, human override protocols, and independent ethics evaluations that can operationalize these safeguards in high risk cybersecurity environments.

V. Conclusion and Future Work

This study provided a thorough evaluation of how ethical principles are being applied in AI for cybersecurity using a structured, evidence-driven approach. Results show that, while many papers excel at transparency, explainability, account-ability and privacy, there are still significant gaps in human oversight and continuous ethical monitoring after deployment. It is clear that the field is making progress, but more work is needed to ensure high ethical standards throughout the entire AI lifecycle.

A. Key Takeaways:

- Most current research demonstrates strong foundations in core ethical areas, but regular checks and clear processes for keeping humans in control are still not standard.
- There is wide variability in how different papers and projects address ongoing monitoring and accountability, highlighting the need for common standards and best practices.

B. Future Work

Building on the results of this paper, future work should focus on expanding and refining the evaluation framework presented here. One key next step is to apply this approach to a larger and more diverse set of AI systems in cybersecurity to see how well the framework holds up across different domains and use cases. This could include incorporating more real world, industry driven case studies and not just academic research The framework could be extended to track ethical compliance over time, allowing for long-term monitoring of deployed systems. Developing practical tools or automated solutions for continuous ethics monitoring would help address the gap identified in this study, making it easier for organi-zations to keep their AI systems in check as they grow and change.

Overall, the aim is for the evaluation method presented here is to support safer and more ethical AI deployments in cybersecurity and beyond. This will ensure that ethical considerations remain central not just in early development but throughout the entire lifecycle of AI systems.

Acknowledgment

The authors would like to thank the ITM Department at Illinois Institute of Technology for providing research and teaching support.

References

[1] Shritin15, "ITMM—Research-Paper—Group-13," GitHub, 2025. [On-line]. Available: https://github.com/Shritin15/ITMM—Research-Paper— Group-13

[2] N. Vemuri, N. Thaneeru, and V. M. Tatikonda, "Securing Trust: Ethical Considerations in AI for Cybersecurity," J. Knowledge Learning and Science Technology, vol. 2, no. 2, pp. 168–175, 2023.

[3] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," IEEE Access, vol. 10, pp. 98856–98881, 2022, doi: 10.1109/ACCESS.2022.3204171.

[4] A. Al Siam, M. Alazab, A. Awajan, and N. Faruqui, "A Comprehensive Review of AI's Current Impact and Future Prospects in Cy-bersecurity," Dept. of Software Eng., Daffodil Int. Univ., Dhaka, and Al-Balqa Applied Univ., As-Salt, Jordan, 2024. [Online]. Available: moutaz.a@oryx.edu.qa

[5] N. Sarker, M. Furhad, and R. Nowrozy, "AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling, and Research Directions," SN Computer Science, vol. 2, no. 3, 2021, doi: 10.1007/s42979-020-00483-7.

[6] D. Leslie, "Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector," The Alan Turing Institute Report, 2019, doi: 10.5281/zenodo.3240529.

[7] A. Smith and B. Lee, "AI-Driven Threat Detection Systems," IEEE Access, vol. 9, pp. 13422–13435, 2021.

[8] M. Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," Nature Machine Intelligence, vol. 2, pp. 81–83, 2020, doi: 10.1038/s42256-020-0188-1.

[9] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," Harvard Data Science Review, vol. 3, no. 1, 2021, doi: 10.1162/99608f92.8cd550d1.

[10] J. Floridi and L. Cowan, "Operationalizing AI Ethics," AI Ethics, vol. 2, pp. 1–12, 2022, doi: 10.1007/s43681-021-00085-8.

[11] European Commission, Artificial Intelligence Act, Official Journal of the European Union, 2024. [Online]. Available: https://artificialintelligenceact.eu/

[12] National Institute of Standards and Technology (NIST), AI Risk Management Framework (AI RMF 1.0), Gaithersburg, MD, USA, 2023. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[13] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Securing trust: Ethical considerations for the adoption of artificial intelligence in cybersecurity," Nature Machine Intelligence, vol. 2, no. 6, pp. 264–266, 2020.

[14] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[15] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in Proc. Conf. Fairness, Accountability, and Transparency, pp. 279–288, 2019.

[16] B. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and R. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI," Berkman Klein Center Research Publication, no. 2020-1, Jan. 2020.

[17] L. Floridi et al., "AI4People: An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," Minds and Machines, vol. 28, no. 4, pp. 689–707, Dec. 2018.

[18] M. Taddeo and L. Floridi, "How AI can be a force for good," Science, vol. 361, no. 6404, pp. 751–752, Aug. 2018.

[19] A. Winfield, M. Michael, J. Pitt, and V. Evers, "Machine ethics: The design and governance of ethical AI and autonomous systems," Proc. IEEE, vol. 107, no. 3, pp. 509–517, Mar. 2019.

[20] K. Morovat and B. Panda, “A Survey of Artificial Intelligence in Cybersecurity,” in Proc. 2020 Int. Conf. Comput. Sci. Comput. Intell. (CSCI), IEEE, pp. 109–115, 2020, doi: 10.1109/CSCI51800.2020.00042.

[21] M. Kumar and S. Reddy, “Deep Learning for Network Security: A Survey,” ACM Comput. Surveys, vol. 54, no. 3, 2022.

[22] A. Dinu, P. C. Vasile, and A. Georgescu, “AI-driven solutions for cybersecurity: comparative analysis and ethical aspects,” National Institute for Research and Development in Informatics – ICI Bucharest, Romania.

[23] J. Ricol, “Ethical Considerations in AI-Driven Cybersecurity: Balancing Automation and Human Oversight,” in Proc. 2024 IEEE Int. Conf. Big Data (Big Data), pp. 5602, Dec. 2022, doi: 10.1109/Big-Data.2024.xxxxx.

[24] V. Kulothungan, “Securing the AI Frontier: Urgent Ethical and Regulatory Imperatives for AI-Driven Cybersecurity,” Capitol Tech-nology Univ., North Bergen, USA, 2024. [Online]. Available: vikramk1986@gmail.com

[25] D. Humphreys, A. Koay, D. Desmond, and E. Mealy, “AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business,” published online Feb. 23, 2024.

[26] D. Li, “Artificial Intelligence Ethics and Cybersecurity: Overview and Prospects,” School of International Education, Quanzhou Univ. of Infor-mation Eng., China, 2024.

[27] A. D. Sontan and S. V. Samuel, “The intersection of Artificial Intelli-gence and cybersecurity: Challenges and opportunities,” World Journal of Advanced Research and Reviews, vol. 21, no. 2, pp. 1720–1736, 2024, doi: 10.30574/wjarr.2024.21.2.0607.

[28] D. Bruschi and N. Diomede, “A framework for assessing AI ethics with applications to cybersecurity,” published online May 18, 2022.

[29] K. C. Chaganti, “Ethical AI for Cybersecurity: A Framework for Balancing Innovation and Regulation,” S&P Global, New York, USA, 2024. [Online]. Available: k.chaganti@spglobal.com

[30] A. F. T. Winfield and M. Jirotka, “Ethical Governance Is Essential to Building Trust in Robotics and Artificial Intelligence Systems,” Philosophical Transactions of the Royal Society A, vol. 376, no. 2133, 2018, doi: 10.1098/rsta.2018.0085.

[31] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, “The ethics of algorithms: Mapping the debate,” Big Data & Society, vol. 3, no. 2, pp. 1–21, Jul.–Dec. 2016, doi: 10.1177/2053951716679679.

[32] S. Chitimoju, “Ethical Challenges of AI in Cybersecurity: Bias, Privacy, and Autonomous Decision-Making,” Romanian Journal of Information Technology and Automatic Control, vol. 34, no. 3, pp. 35–48, 2024, doi: 10.33436/v34i3y202403.

[33] E. Tyugu, “Artificial intelligence in cyber defense,” in Proc. 3rd Int. Conf. Cyber Conflict, Tallinn, Estonia, 2011, pp. 1–11.

[34] L. L. Dhirani, N. Mukhtiar, B. S. Chowdhry, and T. Newe, “Review eth-ical dilemmas and privacy issues in emerging technologies: A review,” Dept. Electron. and Comput. Eng., Univ. of Limerick, Limerick, Ireland, and Dept. Electron. Eng., Mehran Univ. of Eng. and Technol., Jamshoro, Pakistan, 2024.

[35] B. Ajminabanu, F. Kareem, and B. Babatunde, “Ethical and regulatory implications of AI in cybersecurity,” IOSR J. Comput. Eng., vol. 26, no. 2, pp. 1–6, Mar.–Apr. 2024.

[36] C. Bell, P. Broklyn, and A. Egon, “The ethical implication of autonomous cybersecurity with transparency and accountability,” SSRN, 12 pp., Jul. 19, 2024.

[37] K. Achuthan, S. Ramanathan, S. Srinivas, and R. Raman, “Advancing cybersecurity and privacy with artificial intelligence: Current trends and future research directions,” Front. Artif. Intell., 2024.

[38] S. Rana and R. Chicone, “Navigating the paradox of AI in cybersecurity: Unpacking societal optimism and ethical skepticism,” Purdue Global Univ., 2024.

[39] T. Jude, “Ethics and transparency in AI-powered cybersecurity systems: Striking a balance between security and privacy,” Jan. 2025.

[40] L. Vaishnav, S. Singh, and K. A. Cornell, “Transparency, security, and workplace training awareness in the age of generative AI,” arXiv preprint arXiv:2501.10389, Dec. 2024, doi: 10.48550/arXiv.2501.10389.

[41] I. Jada and T. O. Mayayise, “The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review,” Digit. Innov. Manag., 2023, doi: 10.1016/j.dim.2023.100063.

[42] B. C. Cheong, “Transparency and accountability in AI systems: Safe-guarding wellbeing in the age of algorithmic decision-making,” School of Law, Singapore Univ. of Social Sci., Singapore, and Univ. of Cambridge, Cambridge, U.K., 2024.

[43] D. Li, “Artificial intelligence ethics and cybersecurity: Overview and prospects,” School of Int. Educ., Quanzhou Univ. of Inf. Eng., Quanzhou, Fujian, China, 2024.

[44] Y. Nag, V. M. Hullatti, A. B. Aiyappa, and U. Kher, “Ethical concerns of using artificial intelligence in cybersecurity,” RNS Inst. of Technol., Bangalore, India, 2024.