Journal of New Technology and Materials

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

Ethical AI Integration in Cybersecurity Operations: A Framework for Bias Mitigation and Human Oversight in Security Decision Systems

Dr. Rebecca Collins, School of Information Security, University of Oxford, UK
Prof. David Turner, Department of Computer Science, University of Oxford, UK

Abstract

Concerns about algorithmic fairness, transparency, and supervision are at the forefront of new AI ethical dilemmas, which are affecting cybersecurity in particular. Incorporating human-centered design, accountability, and fairness into security decision-making processes, this study presents a mitigated framework for ethical AI inclusion. The study lays out the primary mechanisms for oversight, including explainable AI interfaces, continuous feedback units, Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) oversight, and technical case studies and normative models. The findings highlight the most crucial areas for future study and cross-disciplinary collaboration in cybersecurity, while also demonstrating the possibilities and limitations of ethically deploying AI.

Keywords

AI ethics, cybersecurity, algorithmic bias, human oversight, explainable AI, HITL, HOTL, ethical design, bias mitigation, security decision systems

Chapter 1: Introduction

1.1 Background

The most prevalent practice in the field of cybersecurity is the implementation of AI systems to automate a range of tasks. These tasks include access control, vulnerability assessment, behavioral analytics, incident response, and threat identification. In complicated virtual environments, where human response mechanisms could be sluggish or prone to mistake, these systems try to help with fast and informed decision-making. Among the many potential applications of AI in cybersecurity is the ability to analyze massive amounts of data in real-time, detect APTs, discover new attacks, and implement countermeasures rapidly (Charmet et al., 2022). Artificial intelligence (AI) systems may automatically triage alerts, detect abnormalities, and build connections between network events using ML, DL, and NLP. Despite the great promise of these technologies for improving cybersecurity, there are serious moral questions about their integration, especially when AI acts autonomously and affects people, businesses, and essential infrastructures. Concerns about data privacy, transparency, and accountability as well as the potential for humans to lose agency in security management processes give rise to these problems.

1.2 The Ethical Problem Statement: Bias, Autonomy, and Oversight

One major ethical concern with cybersecurity systems that rely on AI is algorithmic prejudice. A discriminatory algorithm may be trained if the training data is inadequate, prejudiced, skewed, or covers historical bias; this would cause a disproportionate number of people, geolocations, or

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

actions to be marked as suspicious. The system's credibility in identifying threats could be compromised, stakeholders could come to distrust it, and false positives or negatives could occur (Akitra, 2024). For instance, a biased training set could cause a facial recognition or behavioral analytics program that enables authentication or surveillance to miss people from minority demographics. Inadequate post-deployment monitoring, lack of transparency during feature generation, and imbalances in the data are other factors that could contribute to these biases, in addition to the model's design decisions.

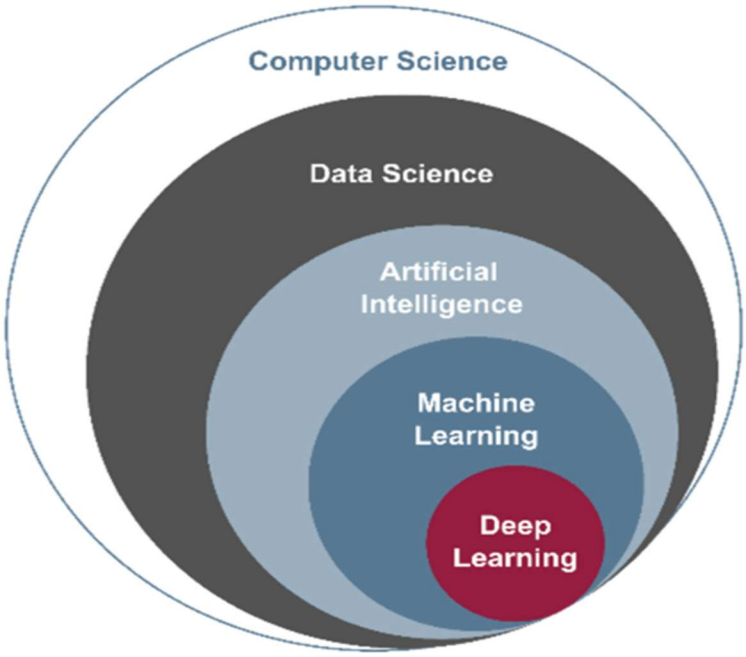

Figure 1: Classification of Artificial Intelligence

Along with bias, the unfettered independence of AI in cybersecurity is becoming an increasingly pressing issue. With the rise of cognitive decision-making agents, rule-based automated systems are blurring the line between human and machine authority. Without human oversight, self-regulating security systems have the potential to make inadequate or excessive judgments. Systemic learning (including the ability to make counterintuitive decisions) and the application of less-than-ideal human lessons resulting from simplifying assumptions allow it to accomplish this in reaction to novel or non-obvious threats. Such systems are practically unsupervised or minimally supervised unless they are correctly designed as Human-in-the-Loop (HITL) or Human-on-the-Loop (HOTL). This leads to breaches of ethical considerations, due process, and accountability, and the loss of these principles for both the affected individuals and the AI decision-makers. As a result, security teams, companies, or software developers may be open to criticism and liability because to an unclear chain of command on who is responsible for handling the fallout from AI decisions. So, to be fair, transparent, and under human control, it is essential that AI for cybersecurity be designed and implemented with ethical considerations in mind. This is not just a normative demand, but a crucial necessity.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

1.3 Research Questions and Objectives

The following chapter presents the most important questions to be used in this investigation:

- In cybersecurity systems that rely on artificial intelligence, how may bias be identified and mitigated?
- To ensure accountability and equity, what kinds of human controls will be implemented?
- How can security decision systems put the three tenets of artificial intelligence (fairness, accountability, and transparency) into practice?

The overarching goal is to present a methodical strategy for the ethical incorporation of AI into cybersecurity infrastructures, with a concentration on reducing bias, making AI explainable, maintaining control, and ensuring ongoing governance.

1.4 Significance of Ethical AI Integration in Cybersecurity Operations

Staying within the legal parameters, keeping trust, and maintaining operational efficiency all necessitate intelligent use of AI. Research has shown that firms are facing inquiries for the usage of AI systems without proper governance. The majority of these organizations lack bias testing tools, audit logs, and cross-team oversight systems (Rjoub et al., 2023; Badi, 2024). Not only can unfair treatment result from poorly managed bias or human guidance aspects of the system, but it can also generate legal liabilities, a loss of trust from stakeholders, and security vulnerabilities. Therefore, building trustworthy, long-term cybersecurity systems that ethically bolster human security defenders requires ethical AI.

2. Literature Review

2.1 Evolution of AI in Cybersecurity

Advanced threat detection, adaptive access control, behavioral analysis, vulnerability predictions, and real-time incident response are just a few ways in which artificial intelligence (AI) has revolutionized cybersecurity. These features indicate a shift towards a more adaptable approach, establishing it as an adaptive learning-based system capable of tackling the complexity and speed of modern cybercrimes (Olasehinde, 2023). Security operations centers (SOCs) may automate repetitive or monotonous processes, reduce alert fatigue, and react faster to new threats with the help of AI. In recent years, advancements in ML models and DL frameworks have made it possible to detect zero-day vulnerabilities, examine trends in attacker behavior, and make considerable predictions about the locations of threats. The extraction of threat intelligence data from unstructured sources, such as threat reports, news feeds, and dark web sources, is another application of natural language processing.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

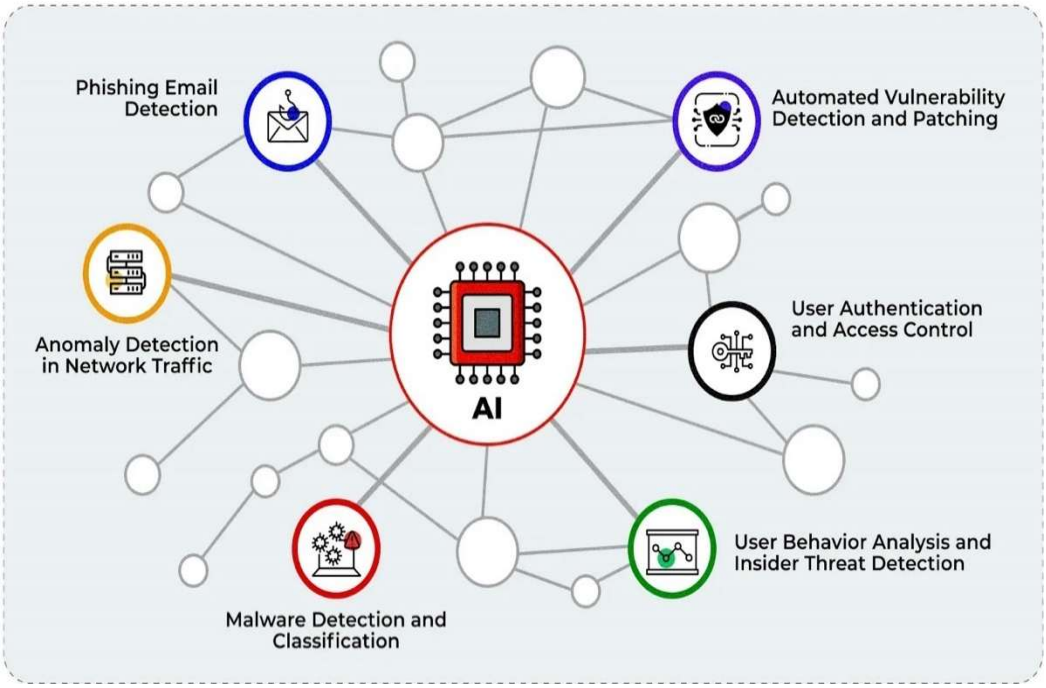

Figure 2: Uses of AI in cybersecurity

However, concerns about lack of accountability, trust, and transparency arise when AI systems and networks are increasingly incorporated into more important systems and networks, especially in cases where human engagement is minimal (Charmet et al., 2022). More research on the ethical viability of entrusting AI systems with security-related tasks is necessary for this advancement.

2.2 Algorithmic Bias in Cybersecurity Systems

2.2.1 Definition and Origins

Problems with imbalanced, incomplete, or non-representative training data can lead to algorithmic bias, which manifests as skewed, uneven, or methodically biased alternatives generated by AI programs (Buolamwini, 2018). This prejudice can impact several areas of cybersecurity, including authentication of identities, restriction of access, behavioral profiling, and threat prioritization. Possible causes of the bias include erroneous labeling of underrepresented user actions during data collection or the design of the algorithm itself, which gives some populations or behaviors more positive or negative feature values than others. When it comes to anomaly detection systems, this becomes even more important because, typically, the system's statistical baselines are utilized to define normalcy and abnormality. But theseivity might include different patterns of behavior between locations, companies, or gadgets (Olasehinde, 2023). The fact that learning models include feedback loops makes the biases much more pronounced, since human analysts may not routinely check or fix the displayed result.

2.2.2 Impacts and Ethical Concerns

The frequency of algorithmic bias in cybersecurity systems raises both practical and ethical concerns. This is illustrated by the unfair use of facial recognition and behavior-based

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

authentication, which unfairly flag minority groups as suspicious and lead to unwarranted actions like surveillance, denial of access, or even disciplinary actions (Ntoutsi et al., 2020). The analyst resources that would have been better spent on real threats are instead being squandered on these annoying false positives, which also cause distrust and problems for the user. Another way bias can lead to blind spots is by causing an organization to systematically disregard certain attack vectors. This leaves them vulnerable to hidden weaknesses. As a whole, biased AI systems are unethical because they violate the cornerstones of good governance and cybersecurity: equality, fairness, and justice. Furthermore, these biases will allow malicious actors to imitate seemingly harmless patterns of behavior, which, when identified, may expose previously unseen dangers (Sterz et al., 2024). These dangers highlight the need for proactive bias detection, comprehensive data practices, and continuous ethical auditing throughout the AI lifespan.

2.3 Human Oversight and Accountability

2.3.1 Necessity of Human-in-the-Loop (HITL)

While AI systems can analyze cyber risks at scale and with unprecedented speed, they still do not compare to human analysts when it comes to understanding context and making decisions based on ethics. According to Gourav Nagar et al. (2018), while making important security decisions, the Human-in-the-Loop (HITL) mechanism makes sure that humans check the overriding AI's output. This is particularly important when dealing with account suspension, threat attribution, or breach notifications including AI. AI alone would not be enough to address these issues because the context is often misunderstood or the purpose is overreaching. In addition to fixing AI mistakes, human oversight implies automated systems are legally and ethically responsible. As people learn to recognize the patterns that AI has identified, HITL improves learning synergy.

On the other hand, retraining AI models is based on the results of human reviews. Over the course of this co-learning cycle, both performance and dependability are enhanced. Without HITL, autonomous agents are more likely to fail operationally, ignore ethical considerations, and put themselves at risk of legal action for their wrongdoing.

2.3.2 Defining Effective Oversight

Modern multidisciplinary studies have advanced beyond simple notions of supervision to propose a formal framework that may serve as the foundation for what efficient human control could be. According to Sterz et al. (2024), in order for AI to make ethically sound decisions, four things must be in place: (1) the capacity to influence AI outputs; (2) access to data pertaining to the reasoning and underlying reasoning system; (3) the authority to act on that knowledge; and (4) the obligation to act morally.Particularly for sectors that contribute to national defense, financial, or medical infrastructure—all of which are business-and life-critical—these guidelines lay out a road map for integrating systems of meaningful supervision into AI-powered cybersecurity infrastructures. In addition to a framework for passive observation, campaign-tailored oversight should include the ability to intervene, postpone, or override automatic choices. This also calls for an explainable and transparent system, so people can understand and question AI actions without feeling left in the dark.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

2.3.3 Human-AI Teaming in Cybersecurity

The term "human-AI teaming" refers to a method of coordinated cooperation between people and AI systems that plays to each group's strengths. In cybersecurity, this means that humans will still be needed for high-context, low-quantity jobs like strategy creation or incident escalation, but artificial intelligence will be used for processing massive amounts of data with little context, like log analysis or malware classification (Sarker et al., 2023). The partnership must create AI user interfaces that can be understood and evaluated by humans. When people work together, they are able to reduce mental strain, make more informed decisions at scale, and keep security operations consistent in their ethical practices. In addition, it provides a mechanism for calibrating trust, which helps users understand when and how to trust AI tracking findings and how these results are influenced by human feedback and correction. Studies have shown that this integration can improve threat detection accuracy and decrease alert fatigue, especially when combined with a continuous learning framework.

2.4 Existing Ethical AI Frameworks

2.4.1 Established Principles and Adaptations

To help in the appropriate creation and application of AI systems, various models of ethics have been created. According to Floridi et al. (2018), the European Commission is now working on regulations for trustworthy AI that prioritize human autonomy, fairness, and the avoidance of adverse outcomes. The guidelines have been adjusted to align with the principles of cybersecurity, which place an emphasis on obtaining user permission, detecting threats without bias, and making security decisions that can be explained. Equally important to Belmont's tenets—respect for humans, beneficence, fairness, lawfulness, and community interest—is the Menlo Report's application of these values to information and communication technology research. In order to address actual security concerns, these philosophical frameworks offer fundamental precepts that must be transformed into an operational framework. Important concerns remain about the practical application of these general principles in contexts where decisions need to be taken in a matter of seconds, like SOCs.

2.4.2 Framework Limitations

Even though ethics cover a lot of ground philosophically, several ethical frameworks have taken heat for being too vague or not applicable. As pointed out by McNamara et al. (2022), due to the absence of enforcement and implementation resources, it is frequently poorly accepted by both the developer and practitioner groups. Cybersecurity ethical standards sometimes fail to address complex socio-technical factors because they are either too broad in scope or too narrow in their focus on compliance. When things become tough, the security team may be more concerned with getting the job done quickly than being fair and transparent, and the programmers may not have received enough ethical training. There is also a lack of a comprehensive framework to deal with the specific problems that arise from adaptive and adversarial AI, which change over time in reaction to new threats and how attackers act.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

Figure 3: Risks and Challenges of AI in Cybersecurity

2.4.3 Cybersecurity-Specific Frameworks

The National Institute of Standards and Technology (NIST), Microsoft, and open-source projects like AI Fairness 360 and FairML have all produced industry-specific standards to fill these gaps. The efforts propose ethical audits that can be implemented in deployment settings, documentation plans to comprehend the full scope of dataset delivery, and modular tool sets to detect bias. Fairness testing (of justice across identity and access management systems), adversarial simulation testing (to test system robustness), and drift testing (monitoring of drift in intrusion detection systems, or IDS) are some other cybersecurity applications of these technologies. But, most of these tools are underutilized because they are too complicated or do not work with the way cybersecurity is currently done (Akitra, 2024). A shift in mindset, better training for developers, and support from key stakeholders within security governance frameworks are all necessary for their wider adoption.

2.5 Emerging Research: Adaptive Human-AI Integration

An adaptive human-AI integration project has recently surfaced, with plans to use trust calibration models that can adapt AI systems' degrees of autonomy to different situations, levels of danger, and operators' levels of experience. As an example, Security Operations Centers (SOCs) might implement a tiered autonomy system that allows AI to independently make low-risk judgments while humans are re-evaluated ambiguous or high-impact events (Mohsin et al., 2024). Interoperability, interpretability, and organizational feedback loops are key components of these frameworks, which aim to continuously improve interactions between humans and AI. Finding a middle ground between innovation and protections for ethics, transparency, and human control, Kulothungan (2024) argues for standardized worldwide norms to govern cybersecurity-related

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

high-risk AI systems. Ethical AI governance, which finds a middle ground between automation and principled supervision, will eventually take shape thanks to adaptive approaches like these, which are still in their infancy.

3 Conceptual Foundations

3.1 Ethical Principles in AI

Fairness, Accountability, and Transparency (FAT) are three of the most important concepts in artificial intelligence (AI) ethics. These three principles are strongly supported by most AI ethics frameworks and play a crucial role in directing the creation of ethical cybersecurity systems that utilize AI. One aspect of AI is its fairness, which means that it does not discriminate against any user or object and does not produce biased results due to biased training data or design decisions (Floridi et al., as applied to cybersecurity). Equal opportunity measures, balancing error rates across demographic groups, or mandating procedural fairness—that is, giving people impacted by a judgment a chance to appeal—are all ways to ensure fairness. The ability to track, assign, and enforce culpability for AI system outcomes is what we mean when we talk about accountability. According to Mokander and Floridi (2021) and Turner et al. (2019), organizations that use AI should make sure that everyone knows their part, from developers to operators to oversight teams, and set up systems to track and fix mistakes. Transparency in decision logic is also necessary for algorithmic accountability, so that regulators and the general public can understand and challenge judgments if they so desire. The capacity to make the reasoning behind AI decisions clear and understandable to all involved is what we mean when we talk about transparency, also known as explicability. As part of this process, it is necessary to disclose model quirks, decision threshold mistakes, and offer an explanation that is understandable by humans (Barocas & Selbst; Diakopoulos & Koliska literature). Analysts and end-users can have faith in AI-identified events and comprehend where they came from when cybersecurity is transparent.

3.2 Algorithmic Bias and Its Impact on Security Operations

3.2.1 Examples of Bias in Cybersecurity Domains

Bias in algorithms may take a unique digital form in security capabilities:

- Additionally, intrusion detection systems (IDSs) have the potential to falsely detect legitimate but unusual activity by an underrepresented group, leading to a loss of confidence in alarms or even an attack.
- As demonstrated in the Buolamwini Gender Shades study, wherein facial recognition fails to discern darker-skinned individuals, access control and authentication systems such as biometrics, behavioral analytics, or facial recognition can mistakenly identify users based on specific demographics (Wikipedia). Pleasant Buolamwini.
- Systems that track users' actions throughout time, like surveillance cameras, could unjustly highlight minority-specific situations or behaviors that do not fit the majority profile.

This bias undermines faith in AI systems, causes unneeded disruptions, and leads to security implementations that have not been proven.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

3.2.2 Root Causes of Micro Bias

These microbiases in cybersecurity systems that heavily rely on Artificial Intelligence (AI) typically emerge from subtle but deeply embedded systemic inequities in the building of batches and other models that make it through top-to-bottom checks. First, there is a data imbalance; in other words, there are too many or too few instances of certain user actions, network operations, or dangers in the training data. This leads to a skewed understanding of the model, which in turn causes systematic mistakes that have an outsized impact on the profile of minority behaviors or on events with low likelihood but high significance (Barocas, Hardt, & Narayanan, 2023). In addition, feature selection and representation are significant sources. As sensitive features can be anything other than gender or race, proxying them can enable the algorithm to unintentionally incorporate biases into its decision-making processes (Mehrabi et al., 2021). Decisions about the structure of the model are also necessary; this is especially true with black-box deep learning training algorithms, which introduce and amplify biases by focusing on common patterns instead of ethically important variations (Raji et al., 2020). Security Some feedback loops can even make bias worse. If an intrusion detection system raises an alarm because it has mistakenly reported a certain sort of user activity too often, retraining the system using that flagged data might reinforce the initial mistake and create bias patterns (O'Neil, 2016). Algorithmic bias is a socio-technical problem with the conception, training, and application of models, as shown by these underlying reasons when taken collectively.

3.3.1 Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) Models

Integrating Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) supervision models has become an essential measure to reduce dangers connected with autonomous cybersecurity judgments. In order to prevent errors like false positives or high-impact measures like automated system lockout, the HITL model ensures that humans are involved at critical system decision points. According to Amershi et al. (2019), the strategy's foundation is the accountability and retention principle, which allows human experts to assess and adjust AI outputs in response to specific circumstances and ethical considerations. The HOTL paradigm, on the other hand, lets the system run autonomously. Once a judgment has been made, human operators step in to oversee the process and deal with any irregularities or ethical violations that may have occurred. Due to the impracticality of requiring human intervention at each step, HOTL is well-suited to systems operating at high speeds, such as real-time intrusion detection (Gunning & Aha, 2019).

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

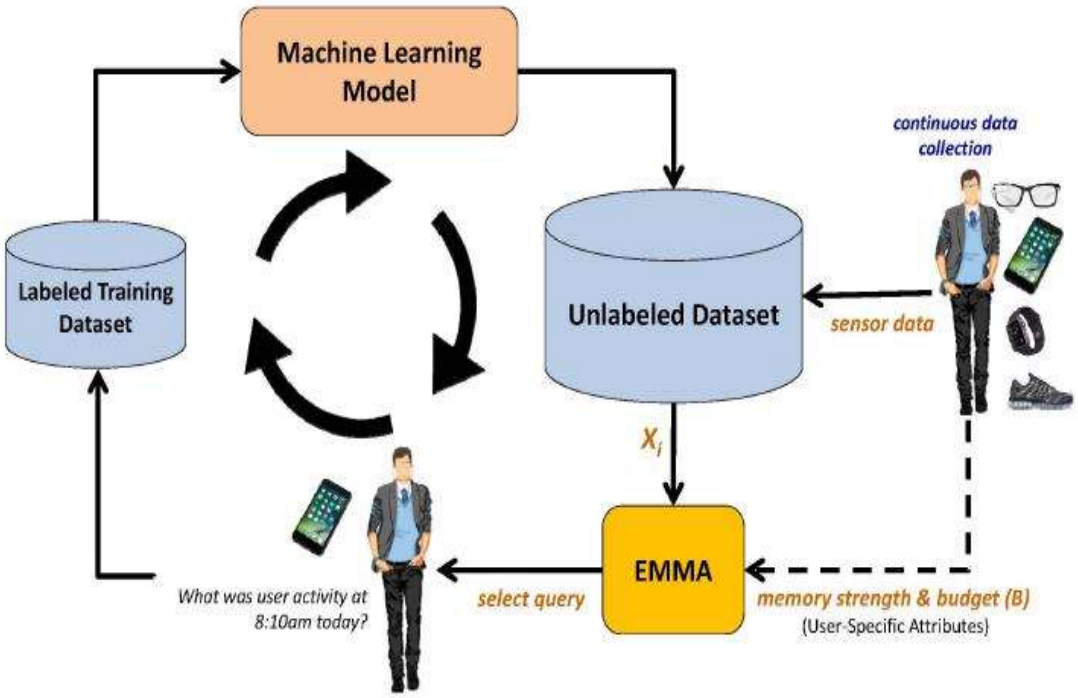

Figure 4: Human-in-the-Loop Learning

However, because HOTL permits examining, resolving, and offering comments on policies, it does not do away with human accountability. In order to avoid operator fatigue or automation bias, it is important to carefully define how the HITL and HOTL systems will interact with cybersecurity operations. This includes including on-screen user interfaces, instructional tools, and alert priority. In the end, these models help make AI-enhanced security environments more transparent, less likely to wander ethically, and better able to respect human agency.

3.3.2 Cognitive Limits and Ethical Responsibilities

Proper design and attendance are essential for good monitoring. According to Sterz et al. (2024) and Donald Farmer (2024), human operators need to have the following abilities: the ability to stop or prohibit actions, the ability to understand and interpret AI judgments, the ability to process and act upon alerts, and the moral will to set ethical constraints. The term "automation bias" describes the tendency for operators to put too much faith in AI's suggestions, particularly when they are tired from being on high alert. Effective cybersecurity management calls for clearly defined responsibilities, thorough training, a user-friendly interface, and a methodical strategy that avoids mental fatigue while promoting moral decision-making. Organizations should set up accountability systems to decide who is responsible for reviewing what, when, and how, and to record and audit the outcomes.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

4. Proposed Framework for Ethical AI Integration in Cybersecurity

4.1 Design Considerations

When planning for AI's widespread use in the future, a well-rounded strategy for cybersecurity must take ethical design concepts like transparency, justice, and responsibility into account from the ground up. Good performance and built-in protections against ethical violations are essential for AI-based security technologies. Important components include taking into account data discrepancies and the reality that underrepresented groups or behaviors might not be prominently displayed in automated decision-making using a bias-aware design approach (Binns, 2018; Cowls & Floridi, 2019). Data rebalancing before processing, in-processing limitations like adversarial debiasing, and post-processing calibration to change model outputs are all ways that system designers can speculate on algorithmic fairness (Mehrabi et al., 2021). Modular ethical architectures that enable selective monitoring and policy enforcement layers to suit diverse settings should also be an element of cybersecurity technologies. To reduce the likelihood of catastrophic failures at a single location and increase transparency in decision-making, these instruments might make use of secure enclaves or decentralized data flows (Brundage et al., 2018). Further, in order to implement decision-making systems that are in line with human values, ethical AI security design should center on the system's predictions rather than their explanation or implementation.

4.2 Human Oversight Layer

A key component of AI-based security operatives' dedication to upholding ethical standards and corporate regulations is the presence of human oversight. For AI decisions to be monitored, assessed, and even intervened with when necessary, a certain degree of next-level supervision is required. This would entail establishing explicitly stated checks where the human operator will validate critical decisions, such as denying access to a user or escalating the reaction to a danger. Such benchmarks can be established in accordance with the model's confidence intervals or a sensitivity limit (Amershi et al., 2019). Regular checks on AI behavior in diverse scenarios should be conducted by oversight committees comprised of technical and ethical specialists to strengthen the level of oversight. Decisions will be examined from an ethical and cybersecurity perspective with this system in place. The ability for analysts or frontline operators to dispute or postpone AI conclusions in the case of suspected anomalies necessitates escalation protocols. This layer also places a premium on the XAI interface. A human reviewer can understand the reasoning behind AI actions through one of the interfaces, which boosts confidence and reduces cognitive burden. Put methods like SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to use to create comprehensible visual representations of model behavior that aid in management and responsibility (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016).

4.3 Continuous Monitoring and Feedback

In order for cybersecurity to be helpful and adapt to changing threats and social ideals, it needs to constantly assess its own performance and receive feedback. Instead of a static ethical AI, cybersecurity should be constantly learning and improving. Integrating auditing procedures into operational framework can help detect patterns that point to ethical lapses, bias buildup, or

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

systemic failure (Raji et al., 2020). Automated anomaly detection, cross-demographic comparisons, or a real-time testing method implemented after the fact can all be used by such mechanisms to find discrepancies. Moreover, systems must to be built such that they can learn by ethical analysis in addition to performance indicators. With the help of a framework, adaptive learning can be guided to improve technology by retraining models with highlighted occurrences, user complaints, or oversight annotations. Some models are starting to integrate a similar component to reinforce learning (RLHF) in an effort to better reflect the preferences of complicated human choices (Christiano et al., 2017). Finally, cyber AI with continuous auditing and feedback can assist firms in meeting current governance needs in the ever-changing threat landscape, aligning with ethical aims, and sustaining public trust.

5. Case Studies and Practical Applications

5.1 IBM Watson for Cyber Security: Human-AI Collaboration in Threat Intelligence

By combining AI with human analysts, Security Operations Centers (SOCs) can improve their threat detection and investigation capabilities, as shown by IBM's Watson in Cybersecurity application. Watson discovers clues that people would miss by analyzing unstructured threat intelligence, which includes things like research papers, security feeds, and blogs. After that, the analysts would verify the reported dangers and guide Watson's development to increase investigation time to over 60% and reduce false positives to almost 30% compared to the rule-based method. Combining intelligent automation with human accountability and decision-making, this hybrid system ensures high speed and performance while yet allowing for human control, which is in line with FAT ethics (Eastgate Software, 2024).

5.2 Darktrace and Autonomous Response at Boardriders: Behavioral Modeling with Oversight

An international retail company called Boardriders has decided to adopt autonomous response technology developed by Darktrace. This technology can detect patterns of behavior that are typical of a certain user or device. Darktrace will take action autonomously to contain risks if it detects anomalies, such as irregular data access or lateral movement. However, human analysts could still see all warnings, verify all activities, and, in an emergency, override the autonomous responder's judgments. It exemplifies the Human-on-the-Loop (HOTL) approach to human-AI collaboration, which tries to find a middle ground between complete autonomy and active supervision. Strong ransomware threat containment and rapid reaction capabilities were among the business outcomes, and human agency was not lost in the process (Eastgate Software, 2024).

5.3 IBM QRadar Advisor with Watson: User-Centered XAI Interface

The international retailer Boardriders has opted to deploy Darktrace's autonomous response technology, which can detect patterns of expected behavior from a certain user and device. Unusual data access or lateral movement are examples of anomalies. Darktrace will automatically take action to contain existing dangers if it discovers them. But human analysts were still visible; they could see all warnings, verify all actions, and, in an emergency, override the autonomous

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

responder's judgments. This is an example of the Human-on-the-Loop (HOTL) approach to human-AI collaboration, which tries to find a middle ground between complete autonomy and constant supervision. With no loss of human agency, the business outcomes included fast reaction features and solid ransomware threat containment (Eastgate Software, 2024).

5.4 Human-in-the-Loop in Advanced Incident Response Systems

Although cybersecurity case studies are few, other industries, such as healthcare and hiring, have recently implemented explanation-guided forms of human oversight. In these systems, deployed models keep an eye out for discriminatory predictions in real-time and, if they are ignored, provide humans with counterfactual explanations so that they can intervene and override those predictions. The method has been used in various areas, but it provides a strong foundation that cybersecurity researchers can use, especially to expose when a system makes a biased or unclear judgment (Mamman et al., 2024).

5.5 Synthesis: Comparative Analysis

Case Study	AI Role & Technique	Human Oversight	Model Ethical Alignment
IBM Watson Cyber Security	NLP-based threat correlation	HITL – human validation	Reduces bias, maintains accountability
Darktrace @ Boardriders	Behavioral anomaly detection & response	HOTL – human review/override	Human safety net, transparency in alerts
QRadar Advisor redesigns	the XAI interface for analyst usability.	Embedded interpretability	enhances transparency and trust in AI output
Explanation-guided oversight model	Counterfactual fairness monitoring	HITL – review with real-time override	Mitigates bias, supports fairness in operations

5.6 Insights and Implications

- In the realm of artificial intelligence (AI), human-in-the-loop (HITL) models play a crucial role in verifying AI decisions that could have fatal consequences, such as threat attributions or account isolation.
- When it comes to routine anomaly detection, human-on-the-loop (HOTL) models excel. However, with review interfaces or escalation processes, the human-in-the-loop is not left behind.
- Design To ensure ethical transparency and minimal automation bias, explainable AI (XAI) improves analysts' comprehension, trust, and intervention capabilities.
- In order to guarantee long-term justice and responsibility, ethical feedback loops, user-centered design, and iterative bias detection are crucial.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

6. Discussion

There has been a break in efficiency in threat detection, analysis, and responses due to the adoption of AI in cybersecurity. Additionally, new complicated ethical challenges have emerged about automation, explainability, accountability, and ethical flexibility. When it comes to cybersecurity, unlike in the real world, it is not enough to have highly developed algorithms; a prudent mix of machine autonomy and human decision-making is also necessary. While AI is capable of rapidly processing massive amounts of data in search of irregularities, human oversight is crucial, particularly when decisions may have far-reaching ethical or legal consequences. To assist reduce the dangers of being too reliant on automation, there exist basic regulatory models like Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL). Concurrently, explainability is now a must-have for AI-based security solutions, since it aids in audits and boosts user confidence in AI systems. But there are nuanced design concerns, such as keeping sensitive logic out of the wrong hands and making models human-interpretable while preserving certain operational logic. Modular transparency and context-based counterfactuals are two examples of such efforts that might give a partial solution. Nevertheless, operational security requirements may necessitate sacrificing completely transparent operations. Adding to the complexity of the problem are questions of legal and organizational responsibility. Traditional notions of liability are not well-suited to systems in which autonomous AI has the power to cause harm or fail to avoid it. Firms are being pushed to create clear lines of accountability and integrate documentation, monitoring processes, and avenues for complaints into their AI operations due to the rise of inconsistent legal frameworks like the European AI Act.

However, effective rules will not be able to handle the variations in ethical references and hazards between industries, locales, and usage unless they are adaptable enough. An ethical framework that can adapt to changes in technology and society is necessary, as detection models developed in corporate settings may not work well or may generate bias in other states or environments. To address both the technical and social implications of AI systems fairly, an interdisciplinary strategy is necessary for ethical AI in cybersecurity. This approach should bring together experts from computer science, law, ethics, sociology, and organizational behavior. In order to incorporate the input of various stakeholders during the design and deployment phases, methods such as ethics councils, co-development, and participatory workshops can be utilized. Finally, in order to responsibly approach AI-augmented cybersecurity, it is important to prioritize human factors over expertise. This includes maintaining high technical standards, being clear about the law, and maintaining ethical stability. Innovations should be made with the ultimate goal of safeguarding both the system and society in mind.

7. Conclusion and Future Research Directions

This article has investigated the moral considerations of using AI in cybersecurity, specifically looking at how AI may help eliminate prejudice, increase transparency, and give humans more say over automated decision-making processes. The offered approach integrates ethically significant

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

features at the design level through human oversight and continuous monitoring, based on the principles of transparency, fairness, and responsibility. This study highlights the significance of ethical responsibility and accurate algorithms, especially in contexts with high stakes like intrusion detection and access control, where model bias could lead to a recurrence of systemic inequities and weaknesses (Mittelstadt et al., 2016; Binns, 2018).

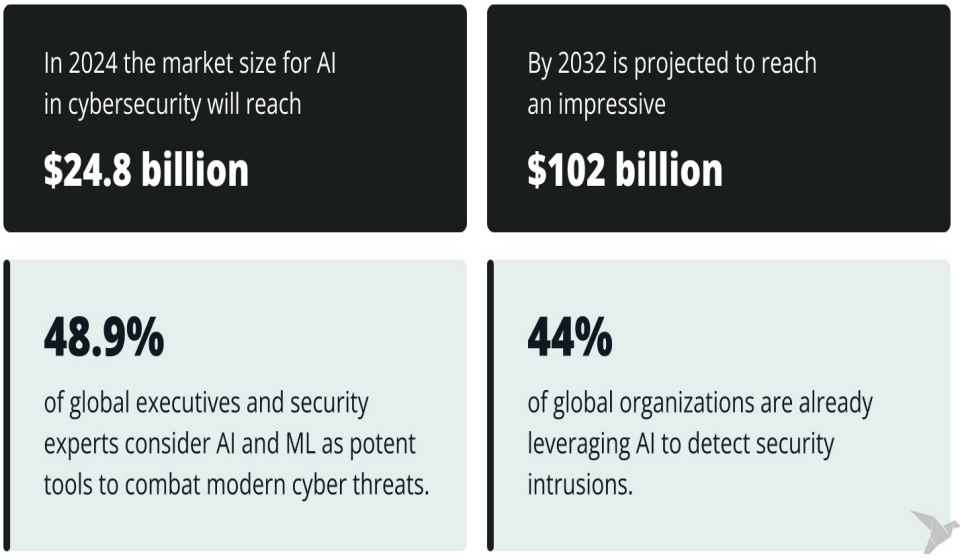

Figure 5: Market projection for AI as of 2024(according to Victoria Shutenko, August 2024)

However, despite its usefulness, the model contains serious flaws. One of these drawbacks is the difficulty in applying it to different corporate environments because of cultural differences in operational management styles, security architecture, and regulatory conformance. In addition, mental exhaustion, corporate inaction, or a lack of adequately qualified staff can prevent the human-in-the-loop structure and human-on-the-loop monitoring from being implemented in practice (Leslie, 2019; Ryan, 2021).

Some technical hurdles that limit the immediate flexibility of adopting ethical AI include explainability trade-offs in deep learning models and the rising diversity of risks from ever-changing adversarial attacks (Morley et al., 2020). The suggested methodology needs more empirical testing in various cybersecurity ecosystems, such as those associated with vital infrastructure, industry, and government. To strengthen the credibility and adaptability of ethical AI tools, there is a growing need for multidisciplinary studies that integrate AI technical design with sociology, public policy, behavioral ethics, and public policy (Cath, 2018; Wachter et al., 2017). Furthermore, in order to ensure operational efficiency, it is crucial to create automated bias auditing techniques that are both clear and easily integrated into real-time security operations. Developing an ethical AI for cybersecurity is a hard endeavor, involving both technological and social aspects. Nevertheless, this paper presents a framework that can be used as a starting point for creating systems that are effective, fair, and accountable.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

Reference

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., . . . Anderljung, M. (2020). Toward trustworthy AI Development: Mechanisms for supporting verifiable claims. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2004.07213

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Leslie, D. (2019). 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. https://doi.org/10.5281/zenodo.3240529

Ryan, M. (2018). Ethics of Public Use of AI and Big Data. *ORBIT Journal*. 2. 10.29297/orbit.v2i1.101.

Liz Rogers, *IBM Security (*2019). Bringing the Security Analyst into the Loop: From Human-Computer Interaction to Human-Computer Collaboration. EPIC Proceedings pp 341–361, ISSN 1559-8918, https://www.epicpeople.org/bringing-security-analyst-into-loop-human-computer-interaction-collaboration/

Mamman, H., Basri, S., Balogun, A., Imam, A. A., Kumar, G., & Capretz, L. F. (2024). Unbiasing on the fly: Explanation-Guided human oversight of machine learning system decisions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.17906

Eastgate Software (September 13, 2024). AI in Cybersecurity: Key Case Studies and Breakthroughs. https://medium.com/%40eastgate/ai-in-cybersecurity-key-case-studies-and-breakthroughs-39bc72ce54ea

Brundage, Miles & Avin, Shahar & Clark, J. & Toner, H. & Eckersley, P. & Garfinkel, B. & Dafoe, A. & Scharre, P. & Zeitzoff, T. & Filar, B. & Roff, H. & Allen, G. & Steinhardt, J. & Flynn, C. & O Heigeartaigh, Sean & Beard, Sj & Belfield, Haydn & Farquhar, Sebastian & Amodei, Dario. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. 10.48550/arXiv.1802.07228.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences*. Advances in Neural Information Processing Systems, 30.

Cowls, J., & Floridi, L. (2019). *A unified framework of five principles for AI in society*. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.
https://doi.org/10.48550/arXiv.1702.08608

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD International

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778

Amershi Saleema, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, & Eric Horvitz. (2019). Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19' 19). Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. https://doi.org/10.1145/3290605.3300233

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning*. http://fairmlbook.org/

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) program. *AI Magazine*, 40(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. https://doi.org/10.1145/3457607

Cathy O'Neil. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024, June). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2495-2507).

Donald Farmer (December 27, 2024). TreeHive Strategy. Human oversight enables automated data governance. https://www.techtarget.com/searchdatamanagement/opinion/Human-oversight-enables-automated-data-governance

Binns Reuben (2018). Fairness in Machine Learning: Lessons from Political Philosophy. https://proceedings.mlr.press/v81/binns18a.html

Crawford, K. (2022). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Perspectives on Science and Christian Faith. 74. 61–62. 10.56315/PSCF3-22Crawford.

Deeks, A., The Judicial Demand for Explainable Artificial Intelligence (August 1, 2019). 119 Colum. L. Rev. __ (2019 Forthcoming), Virginia Public Law and Legal Theory Research Paper No. 2019-51, Available at SSRN: https://ssrn.com/abstract=3440723

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716679679 (Original work published 2016)

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods, and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. '., . . . Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Inioluwa Deborah Raji & Joy Buolamwini. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES' 19). Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

Sandra Wachter, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, Volume 7, Issue 2, May 2017, Pages 76–99, https://doi.org/10.1093/idpl/ipx005

Weller, A. (2019). Transparency: Motivations and Challenges. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, KR. (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science(), vol 11700. Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_2

Singhal A, Neveditsin N, Tanveer H, Mago V. Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review. JMIR Med Inform. 2024 April 3;12:e50048. doi: 10.2196/50048. PMID: 38568737; PMCID: PMC11024755.

Mokander, J., Morley, J., Taddeo, M. & Floridi, L. (2021). Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. 10.48550/arXiv.2110.10980.

Turner Nicol Lee, Paul Resnick, and Genie Barton (May 22, 2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Wikipedia. Algorithmic accountability. https://en.wikipedia.org/wiki/Algorithmic_accountability?

Wikipedia. Joy Buolamwini. https://en.wikipedia.org/wiki/Joy_Buolamwini

https://redresscompliance.com/ethical-issues-ai-cybersecurity/

David Caswell, Sabthagiri Saravanan Chandramohan, Deborshi Dutt, Chris Knackstedt, Vikram Reddy Kunchala, David Mapgaonkar, Mike Morris, Abdul Rahman, Kate Fusillo Schmidt, Niels van de Vorle (2024). The CISO's Guide to Generative AI. https://www.deloitte.com/

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

Charmet, F., Tanuwidjaja, H.C., Ayoubi, S. *et al.* Explainable artificial intelligence for cybersecurity: a literature survey. *Ann. Telecommun.* 77, 789–812 (2022). https://doi.org/10.1007/s12243-022-00926-7

Akitra (September 16, 2024) Cybersecurity: Balancing Security Needs with Algorithmic Bias and Transparency. https://medium.com

Bruschi, D., Diomede, N. A framework for assessing AI ethics with cybersecurity applications. *AI Ethics* 3, 65–72 (2023). https://doi.org/10.1007/s43681-022-00162-8

Rjoub, G., Bentahar, J., Wahab, O. A., Mizouni, R., Song, A., Cohen, R., Otrok, H., & Mourad, A. (2023). A survey on Explainable Artificial intelligence for Cybersecurity. *IEEE Transactions on Network and Service Management*, *20*(4), 5115–5140. https://doi.org/10.1109/tnsm.2023.3282740

Badi, Sadi. (2024). Ethical Implications of Integrating AI in Cybersecurity Systems: A Comprehensive Examination. *International Journal of Applied Mathematics and Computer Science*. 56–63.

Roman Panarin, Mekan Bairyev (May 2023) The Role of Artificial Intelligence in Cybersecurity. https://maddevs.io/blog/artificial-intelligence-in-cybersecurity/

Victoria Shutenko (08 August 2024) AI in Cybersecurity: Exploring the Top 6 Use Cases. https://www.techmagic.co/blog/ai-in-cybersecurity

Embedded Machine Intelligence Lab (Feb 20, 2024) Human-in-the-Loop Learning. https://ghasemzadeh.com/project/human-in-the-loop-learning/

Liz Ticong (April 29, 2024) AI in Cybersecurity: The Comprehensive Guide to Modern Security. https://www.datamation.com/security/ai-in-cybersecurity/

Prof. Norbert Pohlmann (October 2024) ARTIFICIAL INTELLIGENCE AND IT SECURITY - MORE SECURITY, MORE THREATS. https://www.dotmagazine.online/issues/digital-security-trust-consumer-protection/artificial-intelligence-it-security

Muniyandi, V. (2022). Harnessing Roslyn for advanced code analysis and optimization in cloud-based .NET applications on Microsoft Azure. International Journal of Communication Networks and Security, 14(4), 979-990.

Muniyandi, V. (2021). Extending Roslyn for custom code analysis and refactoring in large enterprise applications. International Journal of Science and Technology Research Archive, 3, 271-283.

Muniyandi, V. (2022). Harnessing Roslyn for advanced code analysis and optimization in cloud-based .NET applications on Microsoft Azure. International Journal of Communication Networks and Security, 14(4), 979-990.

Muniyandi, V. (2021). Extending Roslyn for custom code analysis and refactoring in large enterprise applications. International Journal of Science and Technology Research Archive, 3, 271-283.

Muniyandi, V. (2024). Design and Deployment of a Generative AI Copilot for Veterinary Practice Management Using Azure OpenAI and RAG Architecture. Available at SSRN 5342838.

ISSN:2170161X | E-ISSN: 2588-2082 | Vol.15 Issue1 (June 2025)

Muniyandi, V. (2024). AI-Powered Document Processing with Azure Form Recognizer and Cognitive Search. Journal of Computational Analysis and Applications, 33(5).

Chellu, R. (2021). Secure Containerized Microservices Using PKI-Based Mutual TLS in Google Kubernetes Engine.

Chellu, R. (2022). Spectral Analysis of Cryptographic Hash Functions Using Fourier Techniques. Journal of Computational Analysis and Applications, 30(2).

Chellu, R. AI-Powered Intelligent Disaster Recovery and File Transfer Optimization for IBM Sterling and Connect: Direct in Cloud-Native Environments.

Chellu, R. (2024). Intelligent Data Movement: Leveraging AI to Optimize Managed File Transfer Performance Across Modern Enterprise Networks.

Chellu, R. Adaptive Quantum-Safe PKI Solutions for Nano-IoT Security Leveraging Cognitive Computing.