JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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

Dr. Nur Afiqah Binti Rahman[1], Dr. Ahmad Fadzil Bin Ismail[2]
University of Malaya, Kuala Lumpur, Malaysia

Abstract
The application of AI to cybersecurity raises new ethical questions about algorithmic fairness, transparency, and supervision, among others. 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

Automating threat detection, incident response, behavioral analytics, vulnerability assessment, and access control is the most popular use of artificial intelligence (AI) systems in the cybersecurity environment. Decisions in complicated virtual environments can be made quickly and with more information with the help of these systems, as manual response mechanisms could be delayed or prone to human mistake. Artificial intelligence (AI) opens up a world of possibilities for cybersecurity. Some of these possibilities include analyzing massive amounts of data in real-time, detecting APTs, discovering new attacks, and fast implementing countermeasures (Charmet et al., 2022). Automated alert triage, anomaly detection, and correlation creation are all possible with the help of AI-based systems that use ML, DL, and NLP technologies. Even though there is a lot of room for improvement in cybersecurity with the help of these technologies, there are serious moral questions that arise when AI acts autonomously and affects people, businesses, and essential

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

infrastructures. Data privacy, responsibility, and explainability are all related concerns, as is the possibility of losing human control over security management.

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 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). A biased training set could cause a program that uses face recognition or behavioral analytics for authentication or monitoring, for instance, 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.

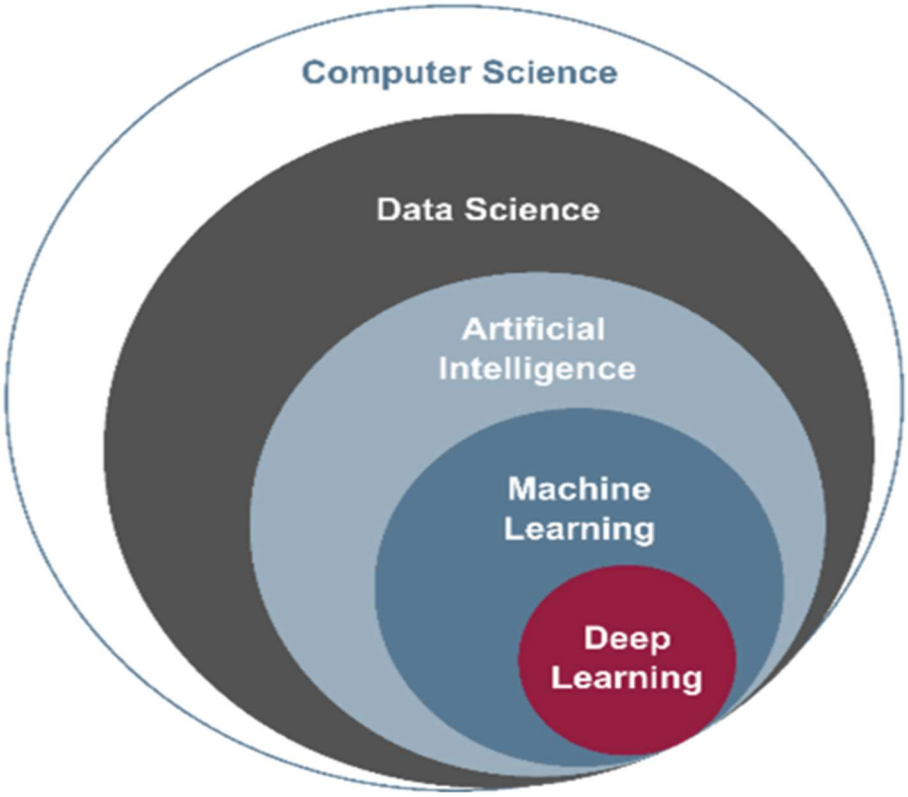

Figure 1: Classification of Artificial Intelligence

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

Along with bias, the unfettered independence of AI in cybersecurity is becoming an increasingly pressing issue. Systems that use rule-based automation technologies are moving toward cognitive decision-making agents, which means that human authority is becoming more and more blurred. When left unchecked, self-regulating security systems might make reckless or harmful decisions. 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. Both the affected individuals and those responsible for making decisions in AI risk losing sight of ethical considerations, due process, and accountability when systems are not designed as Human-in-the-Loop (HITL) or Human-on-the-Loop (HOTL). This is because these systems are effectively unmonitored or minimally supervised. There will be a lack of clear accountability for handling the fallout from AI decisions due to this oversight, which might expose security teams, businesses, or software developers to criticism and litigation. 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.

1.3 Research Questions and Objectives

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

- How can cybersecurity systems that rely on artificial intelligence detect and combat bias?
- 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 general aim is to propose a systematic approach to the ethical integration of AI into cybersecurity infrastructures, specifically focusing on bias mitigation, explainability, control, and 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. Rjoub et al. (2023) and Badi (2024) noted that enterprises are increasingly facing inquiries into the use of AI systems that lack appropriate governance. The majority of these organizations do not have bias testing tools, audit logs, or oversight mechanisms across teams. 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. Responsiblely enhancing human security defenders through the use of trustworthy and long-lasting cybersecurity technology requires ethical AI.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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). Automating mundane or repetitive processes, reducing alert fatigue, and responding faster to new assaults are all possible thanks to AI capabilities in security operations centers (SOCs). These days, sophisticated ML models and DL frameworks can detect zero-day vulnerabilities, examine patterns in attacker behavior, and, to a considerable degree, forecast the locations of threats. Threat reports, news feeds, and dark web sources are all examples of unstructured information that can be mined for threat intelligence data using natural language processing.

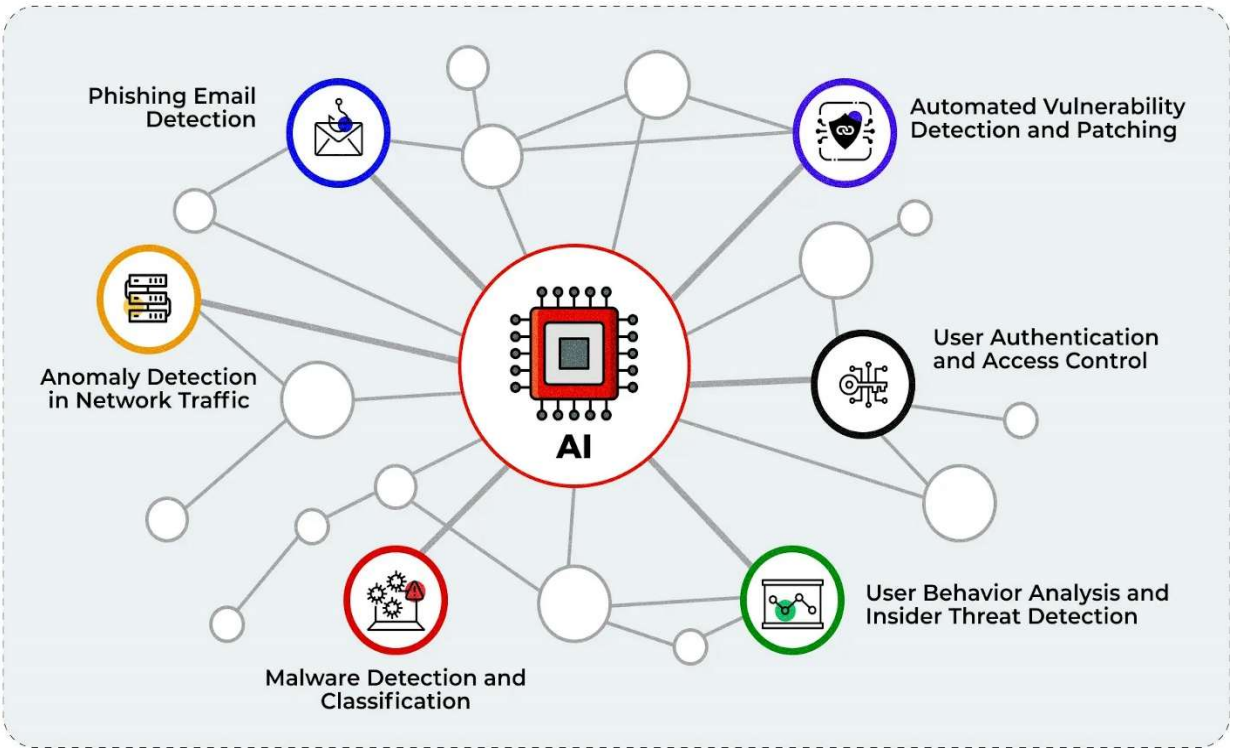

Figure 2: Uses of AI in cybersecurity

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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). The moral viability of entrusting AI systems with security-related tasks needs to be further investigated in light of this progress.

2.2 Algorithmic Bias in Cybersecurity Systems

2.2.1 Definition and Origins

When AI programs produce biased, uneven, or skewed results, it is called algorithmic bias. This happens when the training data is imbalanced, incomplete, or does not match the real world (Buolamwini, 2018). Identity identification, access control, behavioral profiling, and threat prioritization are some of the cybersecurity sectors where this bias could manifest. The data collection process might be to blame if it labels underrepresented user activities incorrectly, or the algorithm could be flawed if it favors or penalizes particular demographics or behaviors with its feature value assignments. This becomes very important when dealing with anomaly detection systems, as these systems typically use their statistical baselines to determine what is normal and what is abnormal. (Olasehinde, 2023) Theseivity, on the other hand, may include a wide range of patterns of behavior across locations, companies, and gadgets. In addition, because learning models include feedback loops, human analysts may not routinely audit or fix the displayed output, which could exacerbate the biases.

2.2.2 Impacts and Ethical Concerns

The frequency of algorithmic bias in cybersecurity systems raises both practical and ethical concerns. Such unfair practices include, for instance, the over-alerting of minority groups by behavior- or face-based authentication systems, leading to unwarranted decisions like surveillance, denial of access, or even disciplinary actions (Ntoutsi et al., 2020). These false positives are a major drain on analyst resources that may be better utilized in addressing genuine threats; they also annoy users, sow distrust, and cause difficulties. Additionally, bias can lead to blind spots, wherein certain attack routes are consistently disregarded, so exposing an organization to hidden vulnerabilities. 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). Because of these dangers, ethical auditing must be continuous throughout the AI lifespan, data procedures must be comprehensive, and bias detection must be proactive.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

2.3 Human Oversight and Accountability

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

Artificial intelligence systems can analyze cyber risks at scale and in record time, but they can 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. In cases where account suspension, threat attribution, or breach notifications are being handled by AI, this would be extremely important. AI alone would not be enough to prevent mistakes in context or overreach. In addition to fixing AI mistakes, human oversight implies automated systems are legally and ethically responsible. In addition, HITL improves learning synergy since humans can learn to recognize the patterns that AI has identified.

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. There is a high probability of operational failure, ethical negligence, and legal culpability as a result of autonomous agents' wrongdoing in the absence of HITL.

2.3.2 Defining Effective Oversight

The current state of multidisciplinary study has advanced beyond the most basic ideas of supervision and proposed a formal framework that might serve as the foundation for what good human control could be. The capacity to influence AI outputs, access to data about the reasoning and underlying reasoning system, the authority to act on that knowledge, and a moral desire to make ethical decisions are the four requirements laid out by Sterz et al. (2024).Businesses and people's lives are at stake when it comes to AI-powered cybersecurity infrastructures, and these standards lay down the groundwork for establishing effective oversight mechanisms in these systems, especially in industries like healthcare, finance, and national defense. 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.

2.3.3 Human-AI Teaming in Cybersecurity

The idea behind human-AI teaming is that humans and AI systems can work together in a structured way, maximizing each party'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). As a group, you must develop AI user interfaces that can be understood and evaluated by humans. Working as a team improves decision-

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

making accuracy, reduces cognitive overload, and keeps 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. Evidence suggests that this integration, especially when coupled with a continuous learning framework, might lessen warning fatigue and improve threat detection accuracy.

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. Human autonomy, justice, and explicability are key to the norms that the European Commission is developing for trustworthy AI (Floridi et al., 2018). User permission, non-discriminatory threat detection, and explainable security decisions are key to the cybersecurity concept, which has led to modifications to the principles. 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. The fundamental values provided by these philosophical frameworks need to be transformed into an operational framework in order to address security challenges in the real world. How these general ideas may be put into practice in contexts like SOCs, where decisions need to be taken in a matter of seconds, remains an important open subject.

2.4.2 Framework Limitations

Despite ethics' solid conceptual grounding, many have argued that existing ethical frameworks are too theoretical and fail to provide sufficient guidance. 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. Under duress, the security team may be more concerned with speed and performance than with being fair and transparent, and the programmers may be uneducated in matters of ethics. In addition, the specific difficulties posed by adaptive and adversarial AI—which change over time in reaction to new threats and attacker tactics—are not adequately addressed by the existing paradigm.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

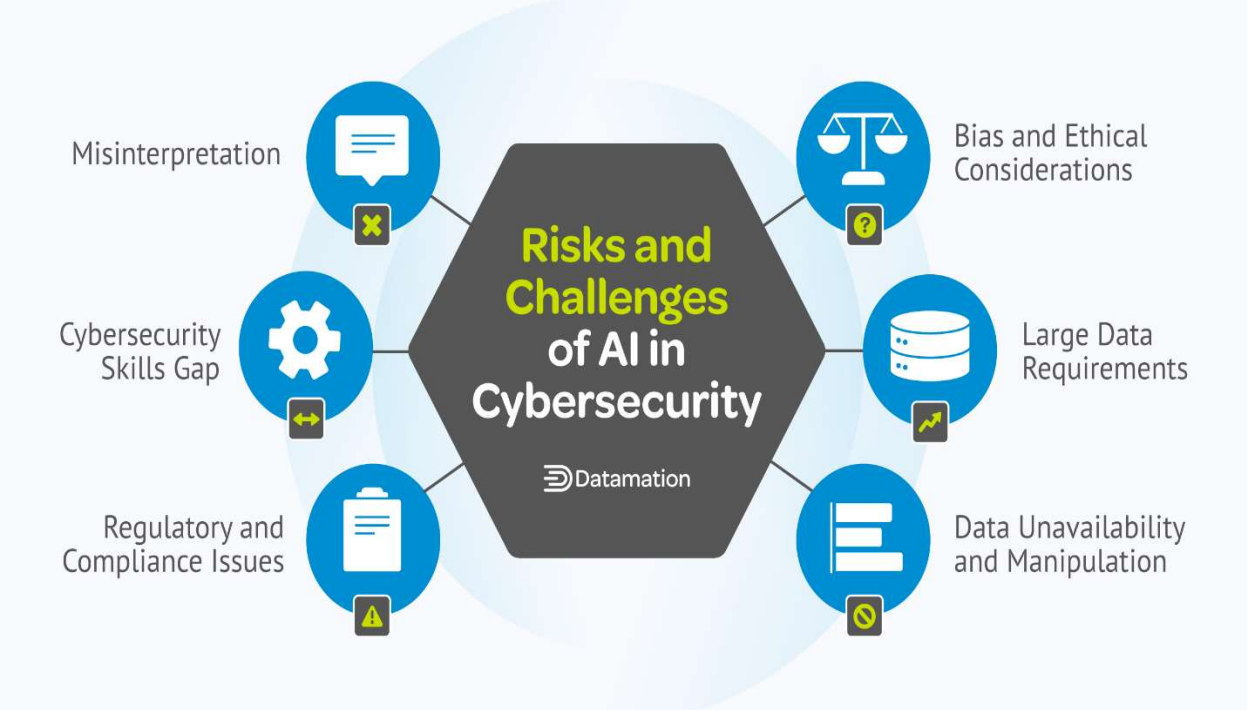

Figure 3: Risks and Challenges of AI in Cybersecurity

2.4.3 Cybersecurity-Specific Frameworks

A number of industry-specific standards have evolved to fill these gaps; they include NIST and Microsoft standards as well as open-source projects like AI Fairness 360 and FairML. 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. However, owing to their complexity or a lack of interaction with current cybersecurity processes, the majority of these tools are not deployed effectively (Akitra, 2024). Stakeholder support within security governance frameworks, developer education, and a shift in cultural attitudes are necessary for their wider use.

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

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

while humans are re-evaluated ambiguous or high-impact events (Mohsin et al., 2024). In order to improve human-AI interactions over time, these frameworks stress the need of interoperability, interpretability, and organizational feedback loops. To manage cybersecurity's high-risk AI systems, Kulothungan (2024) calls for unified worldwide rules that find a middle ground between technological advancement and protections for ethics, openness, and human oversight. While these adaptive approaches are still in their infancy, they are the wave of the future when it comes to ethical AI governance. This wave will find a happy medium between complete automation and principled oversight.

3 Conceptual Foundations

3.1 Ethical Principles in AI

When it comes to developing ethical AI-enabled cybersecurity measures, the three main concepts in AI ethics that are widely acknowledged are transparency, accountability, and fairness (FAT). 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). To be fair, something could be done by employing equal opportunity metrics, balancing error rates across demographic groups, or by mandating procedural fairness, which means that people impacted by a judgment have an option to appeal. 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. To be transparent, or explicable, means that stakeholders should be able to understand and comprehend the reasoning behind AI decisions. The literature by Barocas and Selbst as well as Diakopoulos and Koliska explains that this requires disclosing model quirks and decision threshold failures while also offering an explanation that is understandable by humans. 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:

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

- 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.
- Facial recognition, biometrics, and behavioral analytics are all examples of access control and authentication systems that have the potential to incorrectly identify users based on certain demographics. For example, in the Buolamwini Gender Shades study, it was found that facial recognition failed to identify individuals with darker skin tones (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.

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. Systematic errors that disproportionately impact the minority behavioral profile or low-probability but highly consequential events come from this distorted understanding of the model (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). In the context of black-box deep learning training systems, decisions regarding model architecture are especially important; these systems introduce and amplify biases by focusing on common patterns instead of ethically important variations (Raji et al., 2020). Safety measures Some feedback loops can even make bias worse. Retraining systems using flagged data can further entrench that original mistake, generating bias patterns. This happens, for instance, when an intrusion detection system raises an alarm because it excessively flags a certain sort of user activity (O'Neil, 2016). When considered collectively, these underlying factors show that algorithmic bias is a socio-technical problem with the conception, training, and usage of models, and not just a data problem.

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

Incorporating Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) oversight models has become an essential protection to reduce risks related to autonomous cybersecurity decisions. In

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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. On the other hand, the system is able to function autonomously thanks to the HOTL paradigm. 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. For systems operating at high speeds, like real-time intrusion detection, where it would be impractical to manually intervene at each step, HOTL is an ideal solution (Gunning & Aha, 2019).

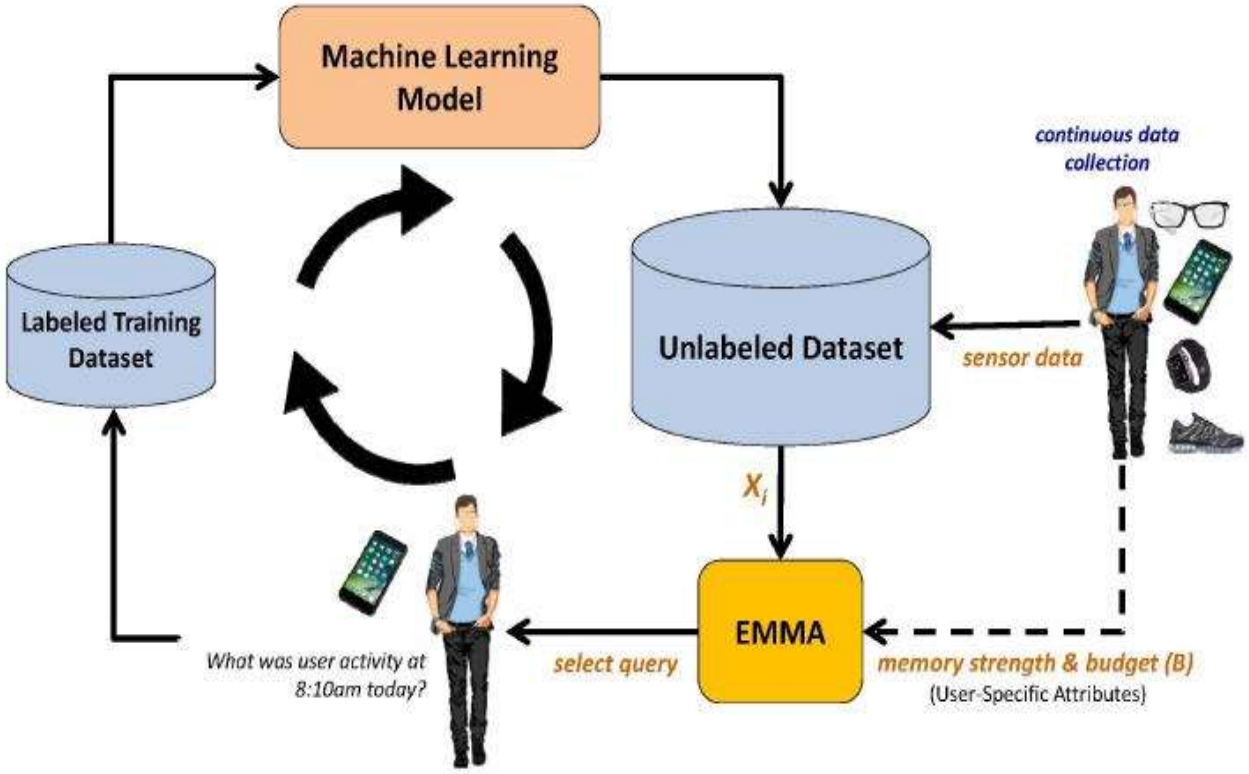

Figure 4: Human-in-the-Loop Learning

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.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

Models like this help keep human agency intact in AI-enhanced security settings while also increasing openness and reducing the likelihood of ethical drift.

3.3.2 Cognitive Limits and Ethical Responsibilities

Both presence and appropriate layout are necessary for effective monitoring. In order to put moral constraints on AI decisions, humans need causal power to stop and prohibit actions, cognitive ability to respond to alerts, and epistemic access to explanations (Sterz et al., 2024; Donald Farmer, 2024). Automation bias happens when operators put too much faith in automation, which causes them to follow AI's suggestions, particularly when they are tired from being on high alert. An easy-to-understand interface, clear responsibilities, thorough training, and a methodical strategy that promotes ethical decision-making are all necessary for good cybersecurity management. Establishing accountability mechanisms is crucial for businesses. These structures should specify who is responsible for reviewing what, when, and how findings are recorded and validated.

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. Effective AI-based security systems should be designed with built-in protections against unethical behavior. 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). Furthermore, in order to implement decision systems that are in line with human values, ethical AI security design should center on the system's predictions rather than their explanation and subsequent implementation.

4.2 Human Oversight Layer

Ethical and rule-abiding behavior on the part of AI-based security operatives can only be guaranteed with human oversight. In order to monitor, assess, and intervene (when required) with AI judgments, a certain degree of next-level supervision is required. Key choices, such as denying

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

access to a user or escalating the reaction to a threat, will be confirmed by the human operator through formally defined criteria. Both the model's confidence intervals and sensitivity limits can be used to establish such checkpoints (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. Additionally, there needs to be an escalation mechanism in place so that analysts or frontline operators can challenge or postpone AI judgments if they suspect anything out of the ordinary. At this level, the XAI interface is crucial as well. A human reviewer can understand the reasoning behind AI actions through one of the interfaces, which boosts confidence and reduces cognitive burden. Implement methods like SHAP (Shapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to produce comprehensible model behavior visualizations that aid in management and responsibility (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016).

4.3 Continuous Monitoring and Feedback

Cybersecurity should not have a static ethical AI; instead, it should constantly assess its own performance and offer input to ensure it is effective and adapts to new threats and societal norms. The operating structure should include auditing systems that can detect trends that could indicate a shift in ethics, the buildup of prejudice, or a system breakdown (Raji et al., 2020). Some examples of such methods include automated anomaly identification, comparisons across demographics, or real-time testing after implementation to spot discrepancies. The most critical thing is that systems should be built to learn from both performance metrics and ethical analysis. 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 add a similar component to better reflect the preferences of complicated human choices, and this has led to attempts to incorporate human feedback into reinforcement learning (RLHF) (Christiano et al., 2017). Finally, enterprises may better align with ethical aims, maintain public trust, and meet governance needs in the ever-changing threat landscape through the use of continuous auditing and feedback in cybersecurity AI.

5. Case Studies and Practical Applications

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

Security Operations Centers (SOCs) can benefit from combining human analysts with artificial intelligence (AI) with IBM's Watson in Cybersecurity. This allows for better threat identification and investigation. Blog posts, research papers, and security feeds are examples of unstructured

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

threat intelligence that Watson processes to identify opportunities that people would miss. By confirming the flagged risks and directing Watson's improvement, analysts may speed up the investigation process to over 60% and eliminate false positives by about 30% compared to the rules-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

The global retailer Boardriders has opted to deploy Darktrace's autonomous response technology, which can detect patterns of expected behavior from a certain user and device. Darktrace will take action autonomously to contain risks if it detects anomalies, such as irregular data access or lateral movement. Human analysts, however, were still visible; they could review warnings, verify activities, and, in an emergency, override the autonomous 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. 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

To better fit the mental model of security analysts, IBM has abandoned the traditional design of QRadar Advisor with Watson, which made use of a complicated knowledge graph representation, in favor of interfaces that are both easy to use and understand. Ethnographic user research has demonstrated that users prioritize interpretability and cognitive fit over advanced visualization. Incorporating human agency and openness into the decision-making process, the redesign enhanced confidence, usability, and adoption of AI tools (Liz Rogers, 2019).

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

Deployed models monitor discriminatory predictions (in real-time) and, when overlooked, report counterfactual explanations to humans that can intervene and overrule specific predictions. This approach has been suggested in other fields, such as healthcare and hiring, as an explanation-guided form of human oversight, although there are few examples of this in cybersecurity. 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).

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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

- Validating AI decisions that could have life-or-death consequences, such as threat attributions or account isolation, is a significant usage of human-in-the-loop (HITL) models even in AI.
- HOTL models excel at routine anomaly detection and do not abandon the human-in-the-loop component through review interfaces or escalation processes.
- Design To ensure ethical transparency and minimal automation bias, explainable AI (XAI) improves analysts' comprehension, trust, and intervention capabilities.
- For lasting justice and responsibility, ethical feedback loops, user-centered design, and incremental bias detection are crucial.

6. Discussion

Aside from the increased ethical and difficult challenges around automation, explainability, accountability, and ethical adaptation, the adoption of AI in cybersecurity has caused a discontinuity in the efficiency of threat detection, analysis, and responses. Cybersecurity, in contrast to real-world applications, necessitates a delicate equilibrium between machine autonomy and human decision-making in addition to a high level of algorithm development. 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. One

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

way to reduce the dangers of being too dependent on automation is to use basic regulatory models like Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL). At the same time, explainability is becoming a must-have for AI-based security apps, since it aids in user trust in AI systems and makes auditing easier. Subtle design considerations include keeping sensitive logic out of the wrong hands and making models understandable by people while retaining certain operational logic. Some solutions may be partially provided by such efforts as modular transparency and the use of context-based counterfactuals. Nevertheless, operational security requirements may necessitate sacrificing completely transparent operations. Traditional theories of culpability are unfit to function in systems where autonomous AI has the power to do or fail to avoid harm, adding another layer of complexity to this already intricate scenario. Companies are under increasing pressure to establish clear lines of accountability as a result of new, inconsistent regulatory frameworks like the European AI Act. They must also integrate paperwork, monitoring processes, and channels for customer complaints into their AI operations.

However, robust rules will not be effective unless they can adapt to different sectors, places, and uses in order to handle ethical references and hazards. It is necessary to have a framework of ethical devices that can adapt to changes in technology and society, as detection models developed in corporate settings may not work well or even generate bias in environments that are culturally or linguistically different. To provide a fair assessment of the social and technical impacts of AI systems, an interdisciplinary strategy integrating computer science, law, ethics, sociology, and organizational behavior is necessary for ethical AI in cybersecurity. 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

The incorporation of AI that supports bias reduction, increased transparency, and human control in automatic decision-making models has been the subject of this research paper's exploration of the ethical aspects of AI that should be implemented in cybersecurity. Human oversight and continuous monitoring ensure that the offered model has ethically significant features at the design level, based on the values of transparency, fairness, and responsibility. This study highlights the significance of ethical responsibility and accurate algorithms, especially in contexts with high

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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).

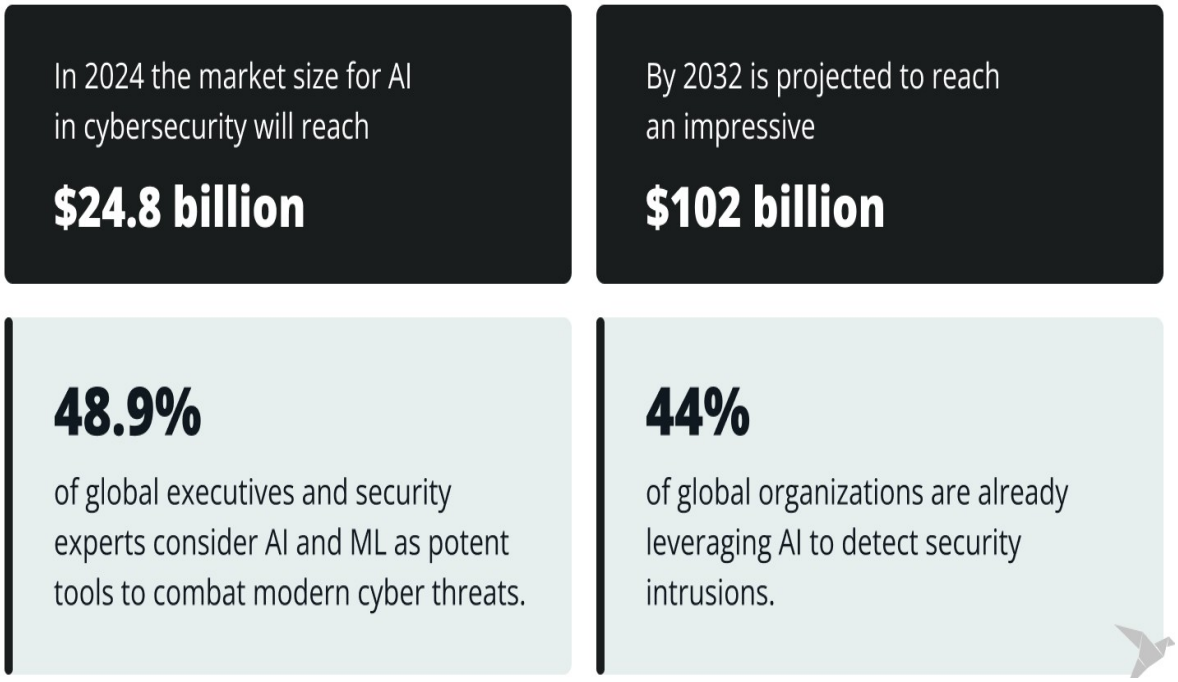

Figure 5: Market projection for AI as of 2024(according to Victoria Shutenko, August 2024)

Nevertheless, despite its merits, the paradigm is severely lacking in key areas. Because of cultural variations in security architecture, regulatory adherence, and operational management styles, it is difficult to generalize these findings to other organizational environments. Also, mental exhaustion, corporate apathy, or a lack of appropriately 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).

Further technological hurdles that limit the rapid adaptability of ethical AI adoption include explainability trade-offs in deep learning models and the growing variety of dangers presented by constantly evolving 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). Also, without sacrificing operational efficiency, real-time security operations must be accompanied by automated bias auditing procedures that are both intelligible and easy to

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

understand. 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.

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.

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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 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).

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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)

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

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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/

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/

JOURNAL OF MEDICAL AND BIOMEDICAL SCIENCE

ISSN: 2026-6294 | Volume No. 11 Issue No. 3 (2025)

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.

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.