# Bias and Fairness Issues in Artificial Intelligence-driven Cybersecurity


## ABSTRACT

**Aim:** This paper aims to examine the bias and fairness issues accorded with artificial intelligence (AI)-driven cybersecurity.
**Problem Statement:** The evolving global dependence on cybersecurity has exposed organizations, individuals, and nations to different vulnerabilities and security threats. However, merging of cyberspace with AI technologies has the potential to transform multiple domains but the implementation of AI is faced with bias problems limiting its application.
**Significance of Study:** Artificial intelligence and cybersecurity have been identified as two transformative and interconnected entities with great potential to revolutionize numerous areas of human life. However, it is imperative to critically look at the bias and fairness accorded with the implication of artificial intelligence-driven cybersecurity which are keywords limiting the usage and efficiency of the approach.
**Discussion:** The concept of artificial intelligence and cybersecurity was discussed together with their interconnectivity which enhances the application in tackling cyber threats. Various areas of artificial intelligence deployment in cyberspace were presented. The sources and solutions to bias and fairness in artificial intelligence-driven cybersecurity were also discussed.
**Conclusion:** Artificial intelligence-driven cybersecurity has found wide industrial applications in different areas. However, there is a need to critically address the issues of bias and fairness attached to it to improve its efficiency.



*Keywords: Artificial Intelligence, Cybersecurity, Bias, Fairness, Automation*



UNDER REVIEW


## 1. INTRODUCTION

Since the existence of the internet, the whole world has evolved around it enhancing the efficiency of different kinds of industrial activities. In the earliest century, studies have been reported on the exponential rapid growth of communication skills which has led to societies with attributed cyberspace for ease of information dissemination and task execution. Various multifaceted areas such as bank transfers, industrial machines control, automated houses management, stock acquisitions, control of automated aerial vehicles, and many more have found broad applications with the help of bridged communication via cyberspace [1]. The ease of operation and inestimable advantages attached to cyberspace have made society rely so much on its usage for different purposes. This over-dependence has called for reasons to protect the information transmitted via this means and ensure that its privacy is adequately maintained. A loose cyber-world can be catastrophic in nature and thus, become vulnerable to attack. Information manipulation by third parties can result in economic and social losses requiring huge amounts of money to fix the atrocities that might have happened [2].

Intruding into secretive and sensitive data/information about financial transactions such as credit card processing, e-commerce, bank stock dealings, and transactions can cause a business to be bankrupt. These are serious challenges posing an unfriendly threat to securing cyberspace information. Furthermore, the improvement of science and technology also poses new challenges to cyberspace security. Thus, cybersecurity is very imperative to keep our precious information safe from attacks and disallowing unauthorized parties from having access to it. Cybersecurity is the control of any form of damage accorded to electronic communication systems and services [3]. This could be achieved via information protection such that its confidentiality, integrity, availability, authentication, and non-repudiation are maintained. Cybersecurity was defined according to CISCO as an act of practicing numerous layers of protection across networks and systems to disallow any attacks on business operations or sensitive information. Cyber security has exhibited some challenges such as complexity in network infrastructure; increasing network capacity; network threats (such as eavesdropping and wiretapping) and breaches of security parameters such as availability, privacy, authorization, non-repudiation, and integrity. However, Artificial Intelligence (AI) has modernized the way cybersecurity is being handled by information security professionals. Newer AI-driven cybersecurity systems and tools have the potential to protect data in better ways against threats by speedily recognizing and automating processes, and behavior patterns and detecting anomalies [1].

One of the key technologies of the Fourth Industrial Revolution is Artificial intelligence, which is applicable in protecting Internet-connected systems from cyberattacks, threats, unauthorized access, or damage. To intelligently tackle associated problems of cybersecurity, artificial intelligence approaches can be excellently adopted. This involves natural language processing concepts; deep learning and machine learning methods; reasoning and knowledge representation; and the basic principle of rule-based expert systems modeling. AI is a multidisciplinary area of computer science that emphasizes establishing machines that are efficient in performing tasks that typically need human intelligence [4]. AI systems development intends to simulate human reasoning functions, such as problem-solving, learning, language understanding, and perception. The concepts of AI include a wide range of applications and technologies that reflect its various and developing nature.

AI is of two types namely; weak AI (Narrow AI) and strong AI (General AI). The former is trained and designed to perform specific tasks. It performs excellently for a particular function but has deficiencies in human intelligence and broad cognitive abilities. Examples are image recognition systems, virtual personal assistants, and speech recognition software. The latter has the potential to learn, understand, and apply knowledge over a wide range of tasks which are identical to human intelligence. Strong AI achievement is a long-term goal making it a substantial area for future research. AI systems have the learning potential from experiences and data [5]. Machine learning algorithms allow systems to increase their efficiency on a specific task as time progresses without being explicitly programmed. Problems can be solved and decisions can be made via logical reasoning. They weigh different factors, process information, and arrive at conclusions depending on learned patterns or predefined rules. AI is designed to solve complex problems by breaking them down into components which involves pattern identification, data analysis, and solutions generation [6].

AI systems based on Natural Language Processing (NLP) can interpret, understand, and generate languages that mimic human thinking. NLP allows communication between machines and humans which facilitates effective interactions between the duo via text or speech. AI systems can interpret and use sensory input to respond to the environment [3]. A subset of AI called “computer vision” allows machines to understand and analyze visual information such as videos and images. AI systems show adaptability via the adjustment of their behavior depending on new information or changing circumstances. This attribute allows

UNDER PEER REVIEW

AI to deal with dynamic and evolving situations. There is the existence of cordial relationships influencing AI and cybersecurity. AI can analyze, monitor, detect, and respond to cyber threats in real-time. AI algorithms can analyze enormous volumes of data for pattern detection signifying its potential for cyber threat detection. It can equally scan the whole network for weaknesses to control various kinds of cyberattacks. AI mainly analyzes and monitors behavior patterns. With these patterns, a baseline is created allowing AI to restrict unauthorized access to systems and also detect unusual behaviors. It can also assist in prioritizing risk, detecting malware possibility instantly and intrusions before they start [6].

AI can serve as a security automation engine when properly implemented. This gives freedom of time and employee resources via the automation of repetitive tasks [4]. AI can also minimize the occurrence of human error via the removal of humans from a process or task. Cybersecurity protection using artificial intelligence will never fully displace security professionals because there will always be a need for creative problem-solving and more complex challenges in the workplace. Before the advent of AI, signature-based detection tools and systems were usually adopted by security professionals to know potential cyber threats. These security tools compare incoming network traffic with malicious code signatures or known threat databases. Once detected, an alert is triggered by the system which alarms the security professional to take quick action to quarantine or block the threat [7]. This signature-based security methodology has been reasonably effective against known threats. However, the signature-based detection method has proven to be inadequate against unknown threats. These tools usually result in a higher frequency of false positives sending security professionals chasing unrealistic goals. However, knowledge regarding the current situation of bias and fairness issues is still scanty. This paper technically examined the concept of artificial intelligence-driven cybersecurity and the bias and fairness issues that are associated with it.

## 2. CYBERSECURITY AND ARTIFICIAL INTELLIGENCE INTERSECTION

Cyberspace and AI have arisen as two transformative and interconnected entities that have transformed numerous areas of human life. AI incorporates the advancement of intelligent machines with the potential to execute tasks that typically need human intelligence while cyberspace is the interconnected network of digital infrastructure and computer systems. AI application and development have speedily evolved over the past few years. AI involves sub-disciplines, such as natural language processing, machine learning, robotics, expert systems, and computer vision. These advancements have resulted in intelligent systems creation which can learn from patterns, analyze huge data volumes, and make decisions or predictions with insignificant human interference [8]. Currently, interconnections between AI, cyber capabilities, and autonomous systems have become subjects of discussion. However, cyberspace is a virtual environment that allows the exchange of information and global connectivity via computer networks. It has changed how individuals access information, communicate, interact, and conduct business around the globe. With internet propagation, cyberspace has improved exponentially thereby, facilitating rapid data dissemination, online communities' growth, and e-commerce platforms development.

However, the growing dependence on cyberspace has also exposed organizations, individuals, and nations to different vulnerabilities and security threats. The existence of interplay between cyberspace and AI clenches insightful relevance in shaping the future of society. Merging cyberspace with AI technologies can transform multiple domains, including education, healthcare, security, transportation, and governance [3]. AI-powered algorithms have the potential to analyze huge data amounts generated in cyberspace to enhance decision-making processes, derive valuable insights, and optimize resource utilization.

UNDER PEER REVIEW

Nonetheless, AI-driven cybersecurity systems can detect and control cyber threats protecting personal data, critical infrastructure, and national security.

Cyberspace and AI have substantial interconnectivity because AI technologies are progressively being positioned in cyberspace to boost different aspects of digital activities. AI models and algorithms are adopted to analyze vast data volumes generated in cyberspace, extract valuable insights, and detect patterns [7]. They are used in cybersecurity for cyber threat detection and prevention; arrangement of datasets for analysis and optimization of processes in automation to improve user experiences. Rapid improvements in AI are significantly impacting the cybersecurity field, especially in the military domain. Considerations are currently being given to the application of AI in the digital sphere in offensive and defensive tasks. This allows States to further fortify their networks' resilience, and robustness and tackle hostile cyber operations via the provision of strong strategic and tactical purposes. The level of human control should be context-specific. There should be variation based on the context in which autonomous cyber capabilities are applied as functions to military and humanitarian considerations [9].

The relationship existing between cyberspace and AI is mutually beneficial. AI technologies enrich and optimize several aspects of cyberspace while cyberspace offers a vast and different ecosystem for AI to adapt, learn, and perform the necessary tasks. Data training and generation; and AI-enabled cybersecurity are evidence that proves the existence of an intersection between artificial intelligence and cybersecurity. Numerous data volumes are generated by cyberspace which are used for AI model training. The data generated via social media interactions, online activities, sensors, and transactions in cyberspace act as inputs for AI algorithms [10]. Nonetheless, AI systems influence cybersecurity in cyberspace via the analysis of large volumes of data to recognize patterns as revelations for cyber threats.

## 2.1 AREAS OF ARTIFICIAL INTELLIGENCE DEPLOYMENT IN CYBERSPACE

Cyberspace creates an enabling environment for AI technology deployment as a result of its diverse and vast digital landscape. AI algorithms can be adopted to manage the huge volume of data generated in cyberspace and derive valuable perceptions from it. Some applicable areas of artificial intelligence deployment in cyberspace are discussed.

### 2.1.1 Data Analysis

AI algorithms can analyze and process large datasets in cyberspace to reveal patterns, correlations, and trends. This allows organizations to derive insights from vast amounts of information and make data-driven decisions. The total volume of data generated within cyberspace presents both opportunity and challenge [1]. AI algorithms perform excellently at analyzing and processing large datasets which enable organizations to reveal hidden patterns, trends, and correlations. In cyberspace, AI-driven data analytics authorizes decision-makers with actionable insights. Whether it's optimizing business processes, understanding user behavior, or predicting future trends, AI enriches the capability to get valuable information from the vast sea of data circulating in cyberspace [11].

### 2.1.2 Cybersecurity

Artificial intelligence is adopted to prevent cyber threats such as phishing attacks, malware, and network intrusions after their detection. Pattern analysis can be executed with it in user behavior and network traffic to recognize potential risks and anomalies. One of the principal applications of AI in cyberspace lies within the cybersecurity realm. AI acts as a powerful associate in constantly tackling cyber threats. By utilizing complex algorithms, AI can quickly

UNDER PEER REVIEW

detect and prevent a countless number of cyber threats ranging from phishing attacks and malware to network intrusions [12]. AI systems can recognize potential risks and anomalies in real-time via pattern analysis of user behavior and network traffic. This fortifies the defenses of individuals and organizations against developing cyber threats.

### 2.1.3 User Experience Enhancement

User experiences in cyberspace, such as targeted advertisements, content recommendations, and voice assistants that understand and respond to user queries are personalized using AI technologies. User experience is vital in the virtual realms of cyberspace. AI technologies play a critical role in enhancing and personalizing user interactions. Cyberspace entities can offer targeted advertisements, personalized content recommendations, and responsive voice assistants with the help of sophisticated algorithms [11]. These AI-driven enhancements contribute to the creation of a more user-friendly and engaging online environment and also cater to individual preferences. Cyberspace exists as an expansive and dynamic domain where AI technology deployment manifests across various applications. Cybersecurity defenses to unravel insights via streamlining processes through automation, data analytics, and elevating user experiences. The synergy between cyberspace and AI creates new room for innovation. As both AI and cyberspace continue to progress, their interplay is poised to shape the future landscape of digital interactions and advancements [13].

### 2.1.4 Automation and Optimization

AI-powered automation systems can restructure processes in cyberspace just to reduce manual effort and increase efficiency. This includes tasks like network traffic intelligent routing, customer support chatbots, and predictive maintenance. The transformative impact of AI encompasses process optimization and automation within cyberspace. Automation powered by AI can streamline numerous tasks which reduces manual effort and increases general efficiency. The contribution of AI-driven automation to oper¬ational excellence in cyberspace ranges from customer support chatbots which provide immediate assistance to network traffic intelligent routing for optimal task execution. Predictive maintenance is another important area in which AI algorithms are used to analyze data to forecast potential issues [14]. This minimizes disruptions and allows proactive intervention.

## 3. SOURCES AND SOLUTIONS TO BIAS AND FAIRNESS IN ARTIFICIAL INTELLIGENCE-DRIVEN CYBERSECURITY

A concern with the establishment and use of any artificial intelligence application is referred to as bias. Users of AI such as vendors and cybersecurity organizations should be more vigilant regarding its limiting bias as they integrate more AI into their defenses. Figure 1 represents the transformation of the bias in discriminatory results. Bias can be introduced into AI models by humans in several ways, but there are steps organizations can follow to moderate that. AI models trained based on unrepresentative datasets or biases may amplify and inherit existing biases causing discriminatory or unfair outcomes. Biases in training data can originate from societal prejudices, historical disparities, or sampling biases [15]. These may lead to AI systems making biased decisions or predictions that disproportionately impact certain individuals or groups. AI is inherently biased in one way or the other. AI refers to the process by which machines learn how to do certain things with the aid of data supplied to the system. To achieve this, a particular dataset is needed. However, any dataset is by definition biased because there is no such thing as a complete dataset [10].

UNDER PEER REVIEW

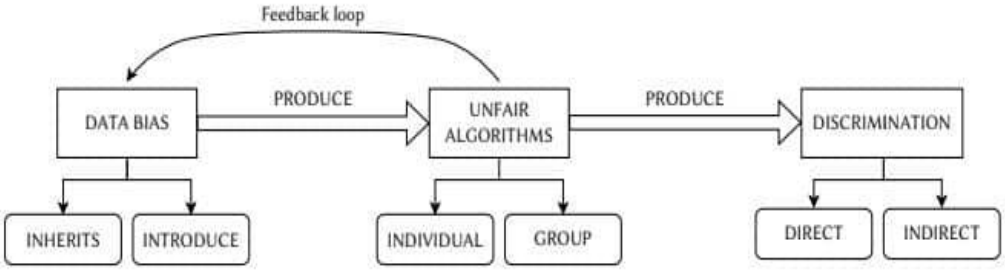


Figure 1: Transformation of the bias in discriminatory results [16]

Fairness in AI is a multifaceted and sophisticated concept that has been the subject of discussion in the field of artificial intelligence. It simply means the absence of discrimination or bias in AI systems. However, achieving fairness in AI can be tasking because careful consideration is necessary for the diverse types of bias that can come up in these systems and how they can be tackled [13]. Various types of fairness include individual fairness, group fairness, and counterfactual fairness. Individual fairness involves the treatment of similar individuals in a similar way by AI systems irrespective of their group membership. This can be achieved via distance-based or similarity-based measures to ensure that individuals having similar terms based on attributes or characteristics are treated similarly by the AI system. Group fairness refers to ensuring that different groups are treated proportionally or equally in AI systems [6]. This can be further subdivided into various types, such as demographic parity, which ensures that the negative and positive outcomes are distributed equally across different demographic groups. Counterfactual fairness is a more current concept that aims at ensuring that AI systems are fair even in hypothetical scenarios. Specifically, counterfactual fairness aims at ensuring that an AI system, regardless of its group membership, would make the same decision for an individual even if they possess different attributes [17].

Bias can infiltrate the process at different points in an AI application cycle. The main drivers of potential AI bias in cybersecurity are the data, algorithm, and cyber AI team. The data is the most crucial place to start when a discussion about AI bias arises. A machine learning algorithm will still execute a task even when the source data lacks completeness or diversity. However, its decision-making will be twisted. For example, a biased spam detection tool can generate false positives and disallow non-spam emails. Experts are now advising that training data for cybersecurity applications should be largely unhurt and unclassified [9]. Also, organizations should be careful when third-party data that may not be significant to their specific cybersecurity wants are being utilized. Bias in uploaded data sets can be measured with the help of open-source tool kits like Aequitas. Also, data bias can be reduced using bias-mitigation algorithms. The algorithm becomes a potential driver for AI bias in cybersecurity when AI models are built by data scientists and are influenced by their own unconscious experiences or ideas. It's imperative that security experts together with data scientists design algorithms in the context of the business requirements. Figure 2 presents an overview of bias impacting fairness [18].

UNDER PEER REVIEW

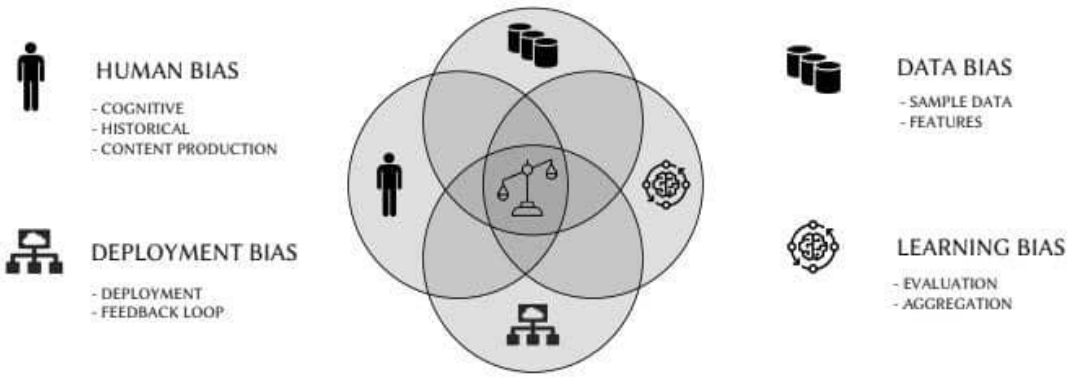


Figure 2: An overview of bias impacting fairness [16]

The formation of skewed models can be inhibited by putting necessary processes into place such as ensuring that the team has vast cybersecurity knowledge and experience to execute a third-party code review. The IBM-developed AI Fairness 360 can be a useful tool in addressing this by combining several bias-mitigating algorithms to detect problems in machine learning models. Lastly, the team engaged in the models and tools development should have an up-to-date understanding of the landscape threat, business knowledge, and strong security experience. They should also possess diverse mindsets and backgrounds [2]. A variety of backgrounds and experiences coupled with cognitive diversity are required to create better-rounded AI cybersecurity systems having the potential of understanding a wide range of threats and behavioral patterns. In this period of growing cyber risk, AI is becoming precious to threat intelligence and cybersecurity defenses. Wide knowledge about the risks of AI bias and remaining vigilant in controlling the introduction of bias into AI-enabled security solutions will be crucial in ensuring that these tools are functional, effective, and impartial [10].

Ensuring unbiased model development and data input is crucial for promoting ethical and fair AI in cybersecurity. Organizations must strive to recognize and mitigate training data biases via bias detection, data preprocessing, and algorithmic fairness testing. By promoting inclusivity and diversity in model training and dataset collection, unbiased and equitable AI systems in decision-making processes can be developed by organizations. Excellent practices for mitigating and controlling bias in AI cybersecurity systems include the establishment of diverse and representative datasets, the development of bias detection and mitigation tools, and the promotion of accountability and transparency in AI development processes [7]. Additionally, fairness-aware techniques and algorithms can be implemented by organizations to mitigate bias in AI models to ensure impartial outcomes across different contexts and demographic groups. The adoption of these measures will help organizations build AI cybersecurity systems that are transparent, fair, and accountable, thereby encouraging confidence and trust in their use and deployment [19]. Figure 3 is the block diagram showing the taxonomy of bias mitigation.

UNDER PEER REVIEW

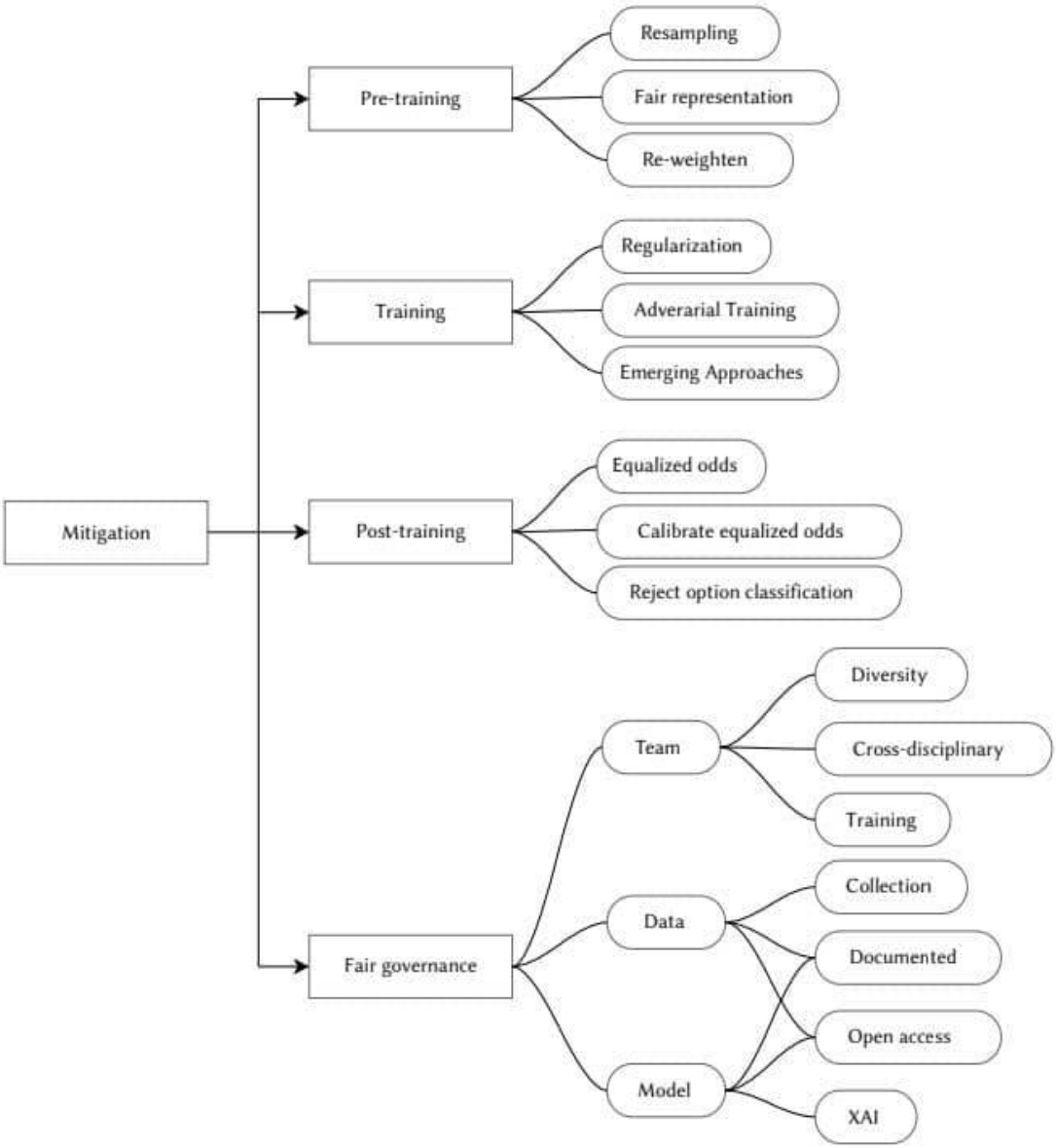


Figure 3: Taxonomy of bias mitigation [16]

## 4. CONCLUSION

The associated bias and fairness issues to artificial intelligence (AI)-driven cybersecurity have been discussed. Artificial intelligence and cyberspace are two transformative and interconnected entities possessing great capacities to revolutionize numerous areas of human life. The interconnectivity between cybersecurity and artificial intelligence as a catalyst for tackling cyber threats was discussed. Identified areas of artificial intelligence deployment in cyberspace are data analysis, cybersecurity, user experience enhancement, automation, and optimization. The main drivers of potential AI bias in cybersecurity are the data, algorithm, and cyber AI team. In conclusion, artificial intelligence-driven cybersecurity has found wide

industrial applications in different areas. However, there is a need to critically address the issues of bias and fairness attached to it to improve its efficiency.

## REFERENCES


[1] Davis JL, Williams A, Yang MW. Algorithmic reparation. Big Data and Society. 2021; vol. 8. doi: 10.1177/20539517211044808.

[2] Kennedy R. The ethical implications of lawtech," in Responsible AI and Analytics for an Ethical and Inclusive Digitized Society. 2021; pp. 198–207.

[3] Chauhan PS, Kshetri N. The Role of Data and Artificial Intelligence in Driving Diversity, Equity, and Inclusion. Computer. 2022; 55, 88–93.

[4] Gilbert TK, Mintz Y. Epistemic therapy for bias in automated decision-making," in Proceedings of the 2019 Conference on AI, Ethics, and Society, New York, NY, USA, 2019, p. 61–67, Association for Computing Machinery.

[5] Zhao C, Li C, Li J, Chen F. Fair meta-learning for few-shot classification," in 2020 IEEE International Conference on Knowledge Graph, Online, August 9-11, 2020, pp. 275–282, IEEE.

[6] Loi M, Ferrario A, Viganò E. Transparency as design publicity: explaining and justifying inscrutable algorithms. Ethics and Information Technology. 2021; vol. 23, pp. 253–263, 9. doi: 10.1007/s10676-020-09564-w.

[7] Chauhan PS, Kshetri N. The Role of Data and Artificial Intelligence in Driving Diversity, Equity, and Inclusion. Computer. 2022; 55, 88–93.

[8] Barocas S, Selbst AD. Big data's disparate impact. Calif. Law Rev. 2016; 104, 671–732.

[9] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv. Neural Inf. Process. Syst. 2016; 29, 4349–4357.

[10] Ferguson AG. Predictive policing and reasonable suspicion. Emory LJ. 2012' 62, 259.

[11] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. J. Law Technol. 2018; 31, 841–887.

[12] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data. 2017; 5, 153–163.

[13] Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018; 16, 31–57.

[14] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019.


UNDER PEER REVIEW

[15] Stolbikova V. Can Elliptic Curve Cryptography be Trusted? A Brief Analysis of the Security of a Popular Cryptosystem. ISACA Journal. 2016; vol. 3, 48-59.

[16] González-Sendino R, Serrano E, Bajo J, Novais P. A Review of Bias and Fairness in Artificial Intelligence, International Journal of Interactive Multimedia and Artificial Intelligence. 2023; Article in Press. http://dx.doi.org/10.9781/ijimai.2023.11.001

[17] Perlner G, Cooper DA. Quantum Resistant Public Key Cryptography: A survey. Proceedings of the 8th Symposium on Identity and Trust on the Internet, Gaithersburg, Maryland, USA: IDtrust, 2009.

[18] Prajapati BLimit of Privacy and Quantum Cryptography. International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET). 2018; vol. 4, no. 4, pp. 1567-1571.

[19] Kaur R, Gabrijelčič D, Klobučar T. Artificial intelligence for cybersecurity: Literature review and future research directions. Inf. Fusion. 2023; vol. 97, p. 101804. doi: 10.1016/J.INFFUS.2023.101804.

UNDER PEER REVIEW