SSR Journal of Artificial Intelligence (SSRJAI)

Volume 2, Issue 2, 2025 Journal homepage: https://ssrpublisher.com/ssrjai/ ISSN: 3049-0413
Email: office.ssrpublisher@gmail.com

Security Implications of Artificial Intelligence in Machine Learning Systems

M.A. YA'A[1], Mohammed Idris[2] & Dr. GT Obadiah[3]

[1]Center for Cyberspace, Department of Cybersecurity, Nasarawa State University, Keffi, Nigeria

[2]Doctorate Candidate, Security and Strategic Studies, Institute Of Government and Department Studies, Nasarawa State University, Keffi

[3]Dean, Faculty of African Languages and Development, Omni University, Imo State, Nigeria

Received: *15.08.2025* | ***Accepted:*** *30.08.2025* | ***Published:*** *07.09.2025*

***Corresponding author:** M.A. YA'A

DOI: *10.5281/zenodo.17080669*

Abstract **Original Research Article**

The AI (Artificial intelligence) and the ML (machine learning) are growing up day by day. Today there are a lot of systems that uses these types of technologies to do tasks of every nature, from medical to military, from agriculture to industries. Also, robotics uses ML to train the machines. But if on one side the AI systems are growing up to do "good tasks" often they are trained to do also "bad tasks" that can influence the concept of security not only digital but also physical and political. This article would summarize and explain the security of the AI systems mainly referencing to the results written in the report: "The malicious use of Artificial intelligence. The rapid integration of Artificial Intelligence (AI) and Machine Learning (ML) systems into critical sectors such as finance, healthcare, defense, and governance has amplified both opportunities and risks. While AI-driven ML models provide adaptive solutions, enhanced decision-making, and predictive accuracy, they also introduce novel security vulnerabilities that traditional systems were not designed to handle. This study examines the security implications of Artificial Intelligence in Machine Learning systems, highlighting how adversarial attacks, data poisoning, model inversion, and algorithmic manipulation can compromise trust, confidentiality, and integrity. It explores the dual-use dilemma, where the same AI algorithms designed to secure systems can be exploited by malicious actors to launch sophisticated cyberattacks. Additionally, the research addresses the ethical and policy challenges of deploying AI-driven ML in sensitive domains, emphasizing the risks of bias, opacity, and lack of accountability in automated decision-making. The work draws attention to the importance of robust security frameworks, adversarial resilience, explainable AI (XAI), and regulatory oversight as critical strategies for safeguarding machine learning ecosystems. By integrating perspectives from cybersecurity, data science, and policy, this paper contributes to an interdisciplinary understanding of AI security challenges and their implications for global digital infrastructure. The findings underscore that ensuring the security of AI-enabled machine learning systems is not just a technical necessity but a societal imperative to maintain trust, safety, and stability in an increasingly automated world.

Keywords: Artificial Intelligence Security, Machine Learning Systems, Adversarial Attacks, Data Poisoning, Model Inversion, Algorithmic Manipulation, Cybersecurity, Explainable AI (XAI), Trust and Accountability, Digital Infrastructure.

Citation: Ya'a, M. A. (2025). Security implications of artificial intelligence in machine learning systems. *SSR Journal of Artificial Intelligence (SSRJAI), 2*(2), 122-135.

1. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) are among the most transformative technologies of the 21st century, reshaping industries, governance, and everyday life *(Taesoo Kim, Dawn Song, Walker Michael, 2017)*. Their growing integration into security frameworks has created unprecedented opportunities for enhancing resilience against threats. AI-driven systems can detect anomalies, analyze vast datasets in real time, and automate decision-making processes that would otherwise be impossible for humans to handle at scale. From cybersecurity to national defense, finance, healthcare, and critical infrastructure, AI and ML are increasingly relied upon to strengthen security mechanisms. However, alongside these opportunities come significant challenges and risks that demand urgent scholarly attention.

The reliance of AI/ML systems on large datasets introduces vulnerabilities related to privacy, data integrity, and unauthorized access. Malicious actors can exploit these weaknesses through adversarial attacks, data poisoning, and model inversion, thereby compromising entire security architectures. Moreover, the "black box" nature of many machine learning models reduces transparency and accountability, raising ethical and legal questions when decisions affect human lives and societal trust. Dual-use concerns further complicate the landscape, as the same AI innovations developed for defense or protection can be misused for offensive cyber operations, surveillance abuse, or large-scale misinformation campaigns *(Aker, C. and Kalkan, S. 2017).*

In recent years, notable cases of AI misuse in cybersecurity and governance highlight the urgency of addressing these challenges. For instance, adversarial examples in image recognition have shown how easily machine learning models can be deceived, raising concerns about their application in security-critical contexts such as biometric authentication and autonomous defense systems. Similarly, the growing sophistication of AI-powered malware and phishing attacks demonstrates how AI can amplify cyber threats rather than mitigate them *(Allen, G. and Chan, T. 2017).* These realities underscore that AI in security is a double-edged sword: while it provides powerful tools for defense, it simultaneously expands the attack surface for malicious exploitation.

Given this dual nature, the study of the security implications of AI in machine learning systems becomes critical for ensuring their safe and ethical deployment *(Crootof, 2015).* Scholars, practitioners, *(Goodman, B. and Flaxman, S., 2016)* and policymakers must not only assess the technical vulnerabilities but also examine the ethical, legal, and societal consequences. Key issues such as fairness, bias, privacy, human oversight, and global inequalities require rigorous exploration to ensure that AI-driven security does not undermine the very rights and freedoms it is intended to protect.

This research is therefore motivated by the urgent need to balance innovation with caution. It seeks to investigate the potential benefits and risks of AI in security contexts, examine ethical considerations, and propose frameworks for responsible governance *(Crawford, 2016*). By adopting a multidimensional approach that integrates technical, ethical, and policy perspectives, the study aims to contribute to the development of resilient, transparent, and accountable AI security systems *(Booth, 2017).*

In conclusion, the introduction of AI into machine learning security systems represents both a remarkable advancement and a profound challenge. The outcome of this research will provide insights into how organizations, governments, and societies can harness the power of AI responsibly, ensuring that its adoption enhances security without compromising human rights, trust, and ethical values.

2. OBJECTIVES OF THE RESEARCH

The main objectives on "Security Implications of Artificial Intelligence in Machine Learning Systems" are briefly stated:

- To examine the potential benefits of integrating Artificial Intelligence (AI) into Machine Learning (ML) systems for enhancing security measures, particularly in threat detection and response.
- To identify the vulnerabilities and risks, such as adversarial attacks, data poisoning, and misuse, that AI-enabled ML systems may introduce into security frameworks.
- To analyze the ethical, legal, and social implications of deploying AI/ML security systems, with emphasis on privacy, transparency, fairness, and accountability.
- To evaluate existing policies, governance mechanisms, and best practices that guide the secure and responsible application of AI in security contexts.

• To recommend strategies for strengthening resilience, promoting explainability, and ensuring ethical oversight in the adoption of AI-driven ML systems for security purposes.

3. METHODOLOGY AND ANALYSIS

This study adopts a mixed-methods research design, combining both qualitative and quantitative approaches to comprehensively explore the security implications of Artificial Intelligence (AI) in Machine Learning (ML) systems.

3.1 Research Design

a. Qualitative: Literature review and expert interviews to explore ethical, legal, and social implications of AI in security contexts.

b. Quantitative: Survey and experimental testing of AI/ML security models to evaluate vulnerabilities and effectiveness.

3. 1.1. Population and Sampling: Target population includes cybersecurity professionals, AI/ML researchers, policymakers, and IT practitioners. Purposive sampling will be used for expert interviews, while stratified random sampling will guide survey respondents to ensure representation.

3.1.2 Data Collection Methods: Primary Data: Structured questionnaires, semi-structured interviews, and simulation experiments of ML security models against adversarial attacks. Secondary Data involves Scholarly articles, industry reports, government policies, and case studies on AI security incidents.

3.1.3. Instruments: Questionnaires designed to assess perceptions of AI/ML risks and benefits.

Experimental setups using existing ML models (e.g., neural networks, decision trees) tested for vulnerability to adversarial inputs.

3.1.4 Ethical Considerations: Informed consent, data confidentiality, and responsible disclosure of sensitive security findings will be strictly observed.

3.2 Data Analysis

i. Qualitative Analysis: Thematic analysis will be applied to interview transcripts and literature findings, identifying recurring patterns related to ethical concerns, accountability, and governance.

ii. Quantitative Analysis: Descriptive statistics (mean, frequency, percentages) will summarize survey results. Inferential statistics (Chi-square tests, regression analysis) will be used to test hypotheses on the relationship between AI adoption and security vulnerabilities. Experimental results will be analyzed by comparing system accuracy and resilience before and after exposure to adversarial attacks.

iii. Triangulation: Results from qualitative and quantitative methods will be cross-validated to improve reliability and provide a holistic understanding of AI security implications.

4. HYPOTHESES

The hypotheses on "Security Implications of Artificial Intelligence in Machine Learning Systems" are:

- **H1:** The integration of Artificial Intelligence (AI) in Machine Learning (ML) systems significantly enhances the efficiency and accuracy of security threat detection compared to traditional security approaches.
- **H2:** AI-driven Machine Learning systems are more vulnerable to adversarial attacks and data poisoning than conventional rule-based security systems.
- **H3:** The lack of transparency and explainability in AI/ML security systems reduces trust and accountability among users and stakeholders.
- **H4:** Ethical concerns such as privacy violations, bias, and misuse significantly influence the adoption and regulation of AI in security applications.
- **H5:** International collaboration and robust governance frameworks positively moderate the risks associated with dual-use and misuse of AI-enabled security technologies.

5. THEMATIC ANALYSIS AND LITERATURE REVIEW

Thematic Analysis and literature review of the research centered on scholarly Notes and journals of authors and scholars that focused their findings on AI Framework and Security on Machine learning.

5.1. Thematic Analysis

A thematic review of literature on AI and security reveals five dominant themes: (a) opportunities in AI-driven security, (b) vulnerabilities and adversarial risks, (c) ethical and legal concerns, (d) governance and accountability, and (e) dual-use and global inequalities.

i. Opportunities in AI-driven Security: AI-powered ML systems enhance threat detection, intrusion prevention, and real-time monitoring. Studies show that AI improves predictive analytics by identifying attack patterns and mitigating risks before escalation *(Sharmeen, 2020)*. Anomaly detection systems using deep learning, for example, have been effective in countering zero-day threats *(Nguyen, 2022).*

ii. Vulnerabilities and Adversarial Risks: Despite strengths, AI/ML models remain vulnerable. Adversarial attacks, data poisoning, and model inversion can manipulate system outputs *(Goodfellow, 2015). Biggio and Roli (2018)* emphasize that reliance on large datasets introduces exploitable weaknesses, raising security concerns for biometric authentication and financial systems.

iii. Ethical and Legal Concerns: Ethical debates focus on bias, privacy, and accountability. AI surveillance systems often operate with minimal oversight, raising human rights issues *(Crawford & Calo, 2016).* Algorithmic bias in predictive policing disproportionately affects marginalized groups, thereby reducing trust in AI security applications *(O'Neil, 2016).*

iv. Governance and Accountability: Governance structures for AI remain underdeveloped. Bryson and Winfield (2017) argue that ethical principles like transparency and accountability should be embedded at the design stage. The EU's AI Act and OECD principles attempt to establish frameworks, but global consensus is lacking *(European Commission, 2021).*

v. Dual-use and Global Inequalities: Dual-use risks dominate AI security debates. Bostrom and *Yudkowsky (2014)* caution that AI developed for defense may be weaponized for cybercrime or military aggression. The unequal distribution of AI technologies further deepens digital divides, creating global security imbalances *(Roberts, 2021).*

5. 2. Literature Review

AI enhances intrusion detection, malware analysis, and fraud prevention through automation. Its ability to process massive datasets surpasses human capacity, making it essential in modern cybersecurity *(Sharmeen, 2020).*

5.2.1 Vulnerabilities of AI/ML Models.

Goodfellow (2015) demonstrated that deep learning models are easily deceived by small perturbations, raising concerns about reliability in high-stakes domains. *Biggio and Roli (2018)* further highlight the persistent risk of data poisoning.

i. Ethical and Social Implications: *Crawford and Calo (2016) and O'Neil (2016)* underscore how AI surveillance and biased datasets can compromise fairness and privacy. These ethical challenges threaten public trust and require stronger oversight mechanisms.

ii. Governance and Policy Gaps: *Bryson and Winfield (2017)* stress embedding ethics into AI development. The EU's Artificial Intelligence Act (European Commission, 2021) provides a regional governance approach, but no comprehensive global standard exists.

iii. Dual-use Dilemma and Misuse: *Bostrom and Yudkowsky (2014)* highlight risks of AI misuse, especially in cyber warfare. *Roberts et al. (2021)* show that technological imbalances between nations intensify global inequality and raise ethical questions about access and security.

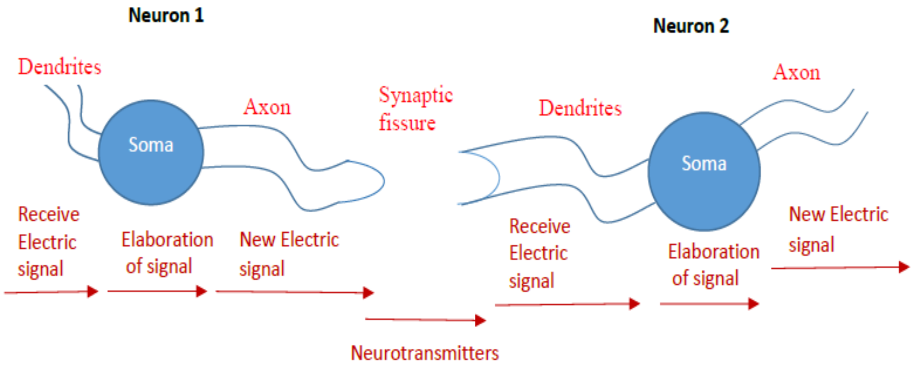
5.3 Synthesis of Literature

The reviewed studies converge on one insight: AI in ML systems is both a security enabler and a security risk. While it strengthens defenses, it also creates vulnerabilities and ethical dilemmas. There is broad agreement that responsible innovation, robust governance, and international collaboration are essential for ensuring AI contributes positively to global security.

6. DISCUSSION

AI refers to the use of digital technology to create systems that are capable of performing task commonly thought to require intelligence. While Machine learning refers to the development of digital systems that improve their performance on a given task over time through experience. Malicious use refers to as all practices that are intended to compromise the security of individuals, groups or a society.

Task: work that an AI system must do to reach a goal. Neural network: Mathematical model that tries to emulate the biological neural network of a man where every neuron (connected to each other by synapses that transmit the signals in the network) elaborates the external stimulus to change its configuration to take decisions.
How an AI system usually work? In this segment, it is described how a neural network (that represent the start for an AI system) works. As previously stated, a neural network is a mathematical model that tries to emulate the biological neural network of a man where every neuron (connected to each other by synapses that transmit the signals in the network) elaborates the external stimulus to change its configuration to take decisions. In medicine a neural network can be summarized as in figure 1 (Neurotransmitter).



Every neuron in a human brain is composed by Dendrites, a Soma and an Axon. Every neuron receives an electric signal through its dendrites, this signal will be elaborated in the soma part of the neuron and it will be propagated to the next neuron through the Axon. To reach the next neuron, the signal is transmitted into the synaptic fissure with an emission of neurotransmitters (fluids) that opens a canal to permit the circulation of ions into the dendrites of the second neuron. These ions cause the transmission of the signal (they alter the electric charge of the dendrites) into the dendrites of the second neuron that goes on in the neural network. A neuron is connected to more than one neuron so, the signals that arrives from the dendrites are more than one. The elaboration in the human brain is a sum of the signals received. Every neuron, after the elaboration, emits only one signal. The signal that can be emitted by every neuron may be exciter, inhibitor or null.
From an AI point of view, as the definition suggests, a neural network must be a mathematical model to represent the medical's neural network. An AI neural network works

like a biological neural network. Starting from a set of input signals, the network elaborates the data in a parallel computation, deploying a graph that terminates once an output unit is reached. All the neural networks are not programmed directly but they use a ML approach to be trained.

Let's start to see how an AI neural networks is built. The unit used in a neural network is the neuron. This unit receives, as the biological network, a certain number of signals and outputs only one signal. The neuron in the next figure receives 3 signals (Hi). Every signal is characterized by one weight (w). This weight can be positive (to induct an exciter signal), negative (to induct an inhibitor signal), >1 (to induct an amplifier synapse), <1 (to induct an attenuator synapse). The weight identifies how a signal is important for the network. Other than the signal, the neuron receives a bias weight (wb) that is used to regulate the work of the neuron. This weight is always connected with a signal (Hb) equal to 1.

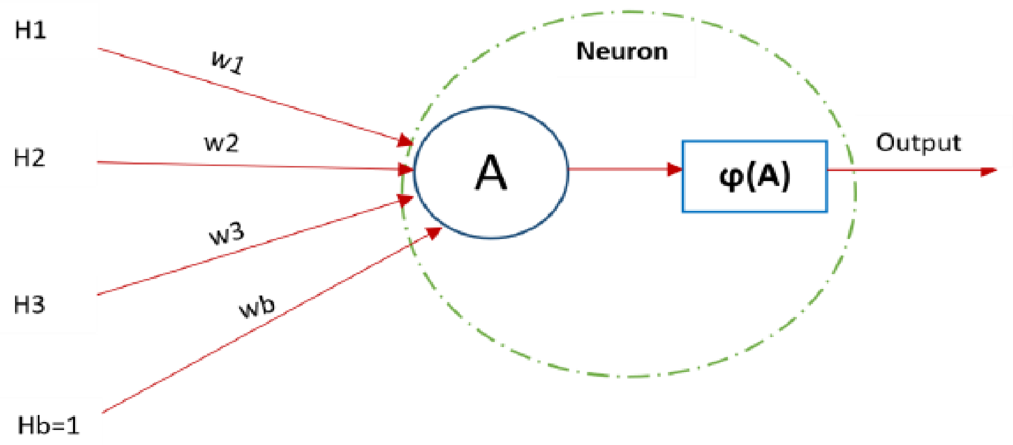


Figure 2: Network signal: Sourced – Authors' Compilation

Every neuron has an internal activation function that usually is the weighted sum of its signals and a function that determines the output signal (also called activity). The internal activation function (A) of this neuron is:
$A = H1 * w1 + H2 * w2 + H3 * w3 + Hb * wb$

To activate a neuron and to calculate the output is used the activity function (ϕ). The activity function most used in AI systems are:Threshold function (or Heaviside function), Piecewise-linear function, Sigmoid function. The threshold function is a function that outputs a value that is 1 or 0. This function is determined with the equation:

$$\varphi(\mathrm{v}) = \begin{cases} 1 \; if \; v \geq 0 \\ 0 \; if \; v < 0 \end{cases}$$

Figure 3: Threshold Function Equation

The Piecewise-linear function is a function that outputs a value that is 1, v or 0. This function is determined with the equation:

$$\varphi(\mathrm{v}) = \begin{cases} 1 \; if \; v \geq \frac{1}{2} \\ v \; if \; -\frac{1}{2} < v < \frac{1}{2} \\ 0 \; if \; v \leq -\frac{1}{2} \end{cases}$$

Figure 4: Piecewise-linear function

The Sigmoid function is determined with the equation:

$$\boldsymbol{\varphi(v) = \frac{1}{1 + e^{-a*v}}}$$

Figure 5: Sigmoid Function

Where “a” is a parameter that determines the slope of the function. Often the neural networks are not composed by only one layer of neurons but they have a multi-layer to perform a lot of data operations. For every layer of neurons added the final approximations will be more precise but requires more calculation.

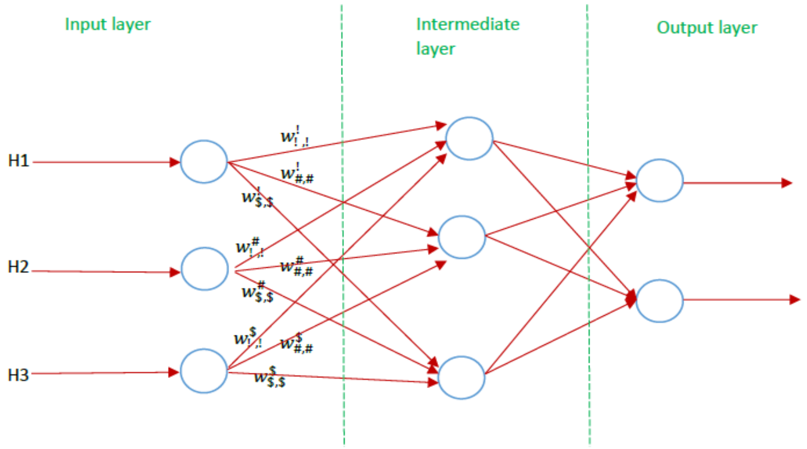


Figure 6: Multi-layer Network Layer

Other than the approximation function described before as activity function, is possible to train our neural network to do some tasks. To do it, is possible to define a train algorithm that, based on some function, will drive the neural network to the best solution for a problem. Some examples of these algorithm can be classified in various types:

Training based on error’s correction: Every neuron receives a signal and creates an output based on the problem to resolve. Knowing the best result of the problem, is possible to calculate the error between the result of the neurons and the best solution of the system. Once calculated the various errors from the neurons in the network is possible to use the gradient method to define the weight to propagate to the next neurons. At the end it’s possible to have the best solution based on the input gave in the first layer.

i. **Training based on memory:** In a memory are stored all the past resolution of a problem. When a new input goes in the system, the system reacts to find the nearest resolution stored in memory to classify the new input.

ii. **Pattern recognition:** Is a process where every signal is assigned to a category. After a training where a network “sees” a lot of pattern of a type, the network will associate similar pattern to the pattern saw in the first phase. For example, if a network is trained to recognize a person, when a photo of that person is showed to the network the system will recognize it also if the person is in a photo that is not stored in memory. Security-Relevant properties of AI: As the training algorithm become more and more precious, the development of AI systems will increase day by day. AI systems today are used for offensive and defensive tasks. So it’s important to identify some properties that are security relevant in a system of this type. Is possible to define efficient and scalable an AI system?

An AI system, as described in the report, is efficient if, trained correctly, can complete certain tasks more quickly and cheaper than a man. An AI system is scalable if, adding computing power, can resolve more instances of a certain task.

Are there any vulnerability in systems of this type? Yes, there are a lot of flaws in these systems. Data poisoning attacks, adversarial examples and exploitation of flaws in the design of autonomous system goals are notified in these years.

Often, also there are a lot of code errors that can drive the systems to do wrong tasks that can be devastating. Think about an automatic car that had errors in the training process and doesn't recognize the red color of a traffic-lights…it could be dangerous! Is possible to become anonymous with AI?

One of the main problems of the malicious use of AI is the anonymity. As documented in a lot of

Novels, often a man trains a machine to do malicious tasks maintaining its anonymity. One example of this is "sweep bot" a cleaning robot created by an anonymous user, that had a bomb inside.

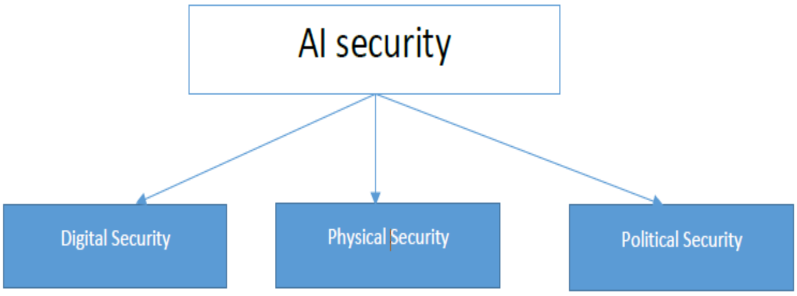


Figure 7: AI Security Category and Architecture (Three different "Security-zones")

Malicious uses of AI can be multiple; therefore, it's a common approach to divide AI security in three different "security zones": digital security, physical security and political security. Digital security is related to the digital attacks related to the AI. Often AI is used in cybersecurity to offense and to defense a system. Related to this topic is possible to notify attacks like the "spear phishing attack" (An automated social engineering attack), large scale attacks, and machine learnings models to avoid detection (a machine that creates and executes commands on a system ad-hoc to avoid detection). Also, in the cybersecurity defense the AI is involved. In fact, AI can be used to implement active-defense: a mechanism that can learn from the attacks received and classified to classify the future attacks.

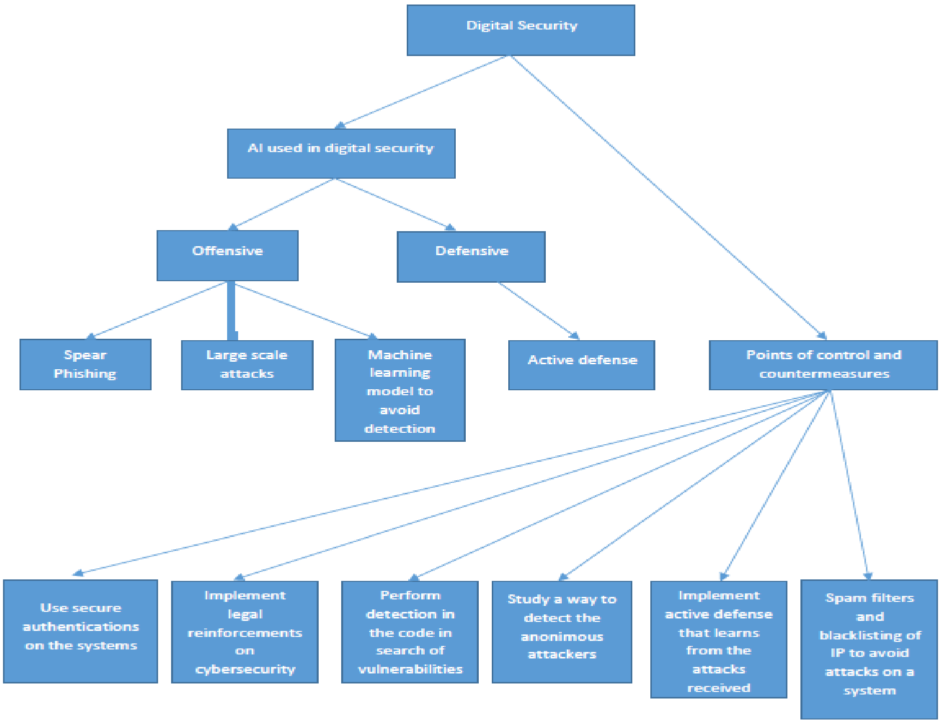

Digital Security
AI used in digital security
Offensive
Defensive
Spear Phishing
Large scale attacks
Machine learning model to avoid detection
Active defense
Points of control and countermeasures
Use secure authentications on the systems
Implement legal reinforcements on cybersecurity
Perform detection in the code in search of vulnerabilities
Study a way to detect the anonimous attackers
Implement active defense that learns from the attacks received
Spam filters and blacklisting of IP to avoid attacks on a system

Figure 8: Digital Security

Sourced: Authors' Compilation, August 27, 2026

Physical Security is related to the use of AI in weapon systems. In fact, there are a lot of robots system and drones that uses AI to kill people. These systems are often trained to recognize a certain person and to kill him automatically. A system like this guarantees the anonymity of the terrorist and may have devastating effects. One famous example of an attack like this is the “sweep bot”, a cleaning robot trained to kill the prime minister. Systems of this type encourage terrorism.

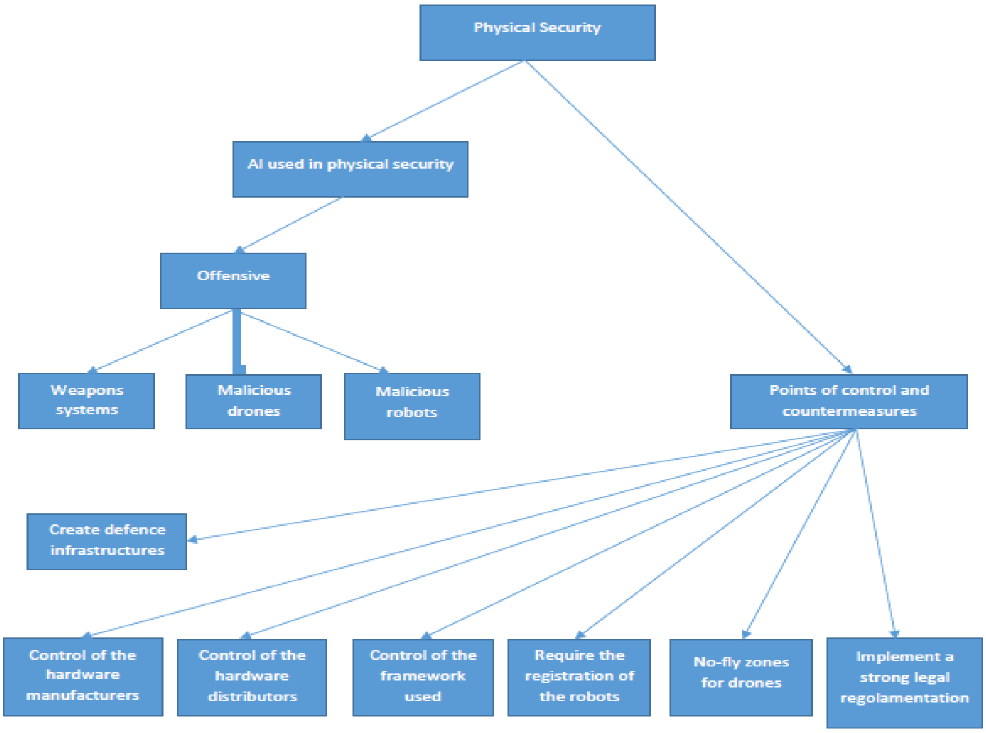

Physical Security
AI used in physical security
Offensive
Weapons systems
Malicious drones
Malicious robots
Points of control and countermeasures
Create defence infrastructures
Control of the hardware manufacturers
Control of the hardware distributors
Control of the framework used
Require the registration of the robots
No-fly zones for drones
Implement a strong legal regolamentation

Figure 9: Physical Security

Sourced: Authors’ Compilation, August 27, 2025

Political Security is related to the use of AI to alter the political landscape of a nation. The AI is used to analyze big data that came from the social networks. In fact, the use of AI on the data that came from the network can masquerade people with political views to spread political messages to cause dissent. Also, is possible to use some social engineering attacks to convince fooling humans to change public opinion. AI is also used to widespread fake news with very realistic videos, to change the voting intentions of a person, to filter some information to a person and to monitor surveillance in the authority regimes.

The widespread of the AI systems and the growing precision of the ML algorithms are expanding the attack’s landscape. In fact, the software that being developed is more complex than the classical software, so often is vulnerable (vulnerabilities are discovered daily). Also, these systems use a lot of data for ML algorithm that must be always controlled because they are a sort of “professors” for the machine and if train the machine in a wrong way, the system doesn’t complete correctly its task or may be vulnerable. Some others security issues came from the dual use of the AI systems. An AI system can be used to offence. The widespread of these types of systems may enable everyone to create an intelligent weapon system that can kill a person maintaining anonymous the attacker. So, security issues are nowadays multiple and is difficult to create a 100% secure AI system.

Mitigate security issues

Four high-level recommendations can be identified to mitigate security issues.

Recommendation #1	Policymakers should collaborate closely with technical researchers to investigate, prevent and mitigate the potential uses of AI

Recommendation #2	“Researchers and engineers in artificial intelligence should take the dual-use of their work seriously, allowing measure related considerations to influence research priorities and norms, and proactively reaching out to relevant actors when harmful applications are foreseeable” (Ref. 1)
Recommendation #3	“Best practices should be identified in research areas with more mature methods for addressing dual-use concerns, such as computer security, and imported, where applicable to the case of AI” (Ref. 1)
Recommendation #4	“Actively seek to expand the range of stakeholders and domain experts involved in discussions of these challenges” (Ref. 1)

Table 1: AI Security Mitigation Issues

Starting from those 4 recommendations, in the report are identified 4 priority research areas where to invest to increase security in AI.

These areas are related to:

i. The application of cybersecurity in AI to discover vulnerabilities and improve the knowledge (creation of red teams, formal verification of the code, creation of public log when a vulnerability is discovered, forecasting security-relevant capabilities, creation and distribution of security tools, implementation and control of the hardware).

ii. Explore the different openness models (Pre-publication of the risk assessment in technical areas of special concern, creation of central access licensing modules, sharing regimes that favour safety and security, sharing norms applied to other AI models).

iii. Promoting a culture of responsibility (Education for scientist and engineers, creation of ethical statements and standards, whistleblowing measures, nuanced narratives).

iv. Developing technological and policy solutions (privacy protection, coordinated use of AI for public-good security, monitoring AI relevant resources) *(Cooper, D.M. 2013).*

Compliance classes

It could be possible to assign compliance classes to AI systems in the same way of IoT systems to create a compliance plan. AI systems, like IoT systems, may involve different classes of security related to the CIA principles (Confidentiality, Integrity, and Availability). Those classes identify how an AI system, if compromised, could be dangerous *(Dao, James, 2013).*

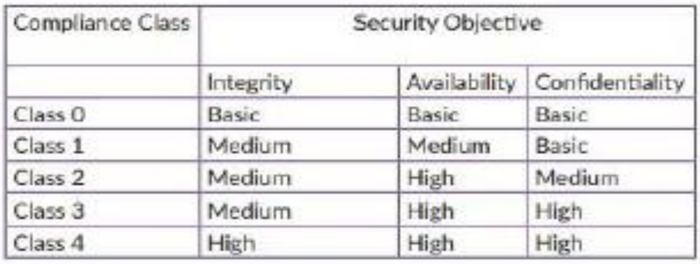
“Class 0: where compromise to the data generated or level of control provided is likely to result in little discernible impact on an individual or organisation.

Class 1: where compromise to the data generated or level of control provided is likely to result in no more than limited impact on an individual or organisation.

Class 2: in addition to class 1, the device is designed to resist attacks on availability that would have significant impact an individual or organisation, or impact many individuals, for example by limiting operations of an infrastructure to which it is connected.

Class 3: in addition to class 2, the device is designed to protect sensitive data including sensitive personal data.

Class 4: in addition to class 3, where the data generated or level of control provided or in the event of a security breach have the potential to affect critical infrastructure or cause personal injury.



Compliance Class	Security Objective		
	Integrity	Availability	Confidentiality
Class 0	Basic	Basic	Basic
Class 1	Medium	Medium	Basic
Class 2	Medium	High	Medium
Class 3	Medium	High	High
Class 4	High	High	High

Table 2: Data Compliance

Where the definitions of the levels of integrity, availability and confidentiality are as follows:

• *Integrity:* o *Basic - devices resist low level threat sources that have very little capability and priority* o *Medium -*

devices resist medium level threat sources that have from very little, focussed capability, through to researchers with significant capability

-High - devices resist substantial level threat sources

• Availability: o Basic - devices whose lack of availability would cause minor disruption

-Medium – devices whose lack of availability would have limited impact on an individual or organisation

-High – devices whose lack of availability would have significant impact to an individual or organisation, or impacts many individuals • Confidentiality:

-Basic – devices processing public information.

Medium – devices processing sensitive information, including Personally Identifiable

Information, whose compromise would have limited impact on an individual or organisation o High - devices processing very sensitive information, including sensitive personal data whose compromise would have significant impact on an individual or organisation." (Ref. 2)

Security and AI systems

As introduced by the faculty summit of Microsoft in 2017 the AI systems can enable the computer security and the computer security can enable AI systems.

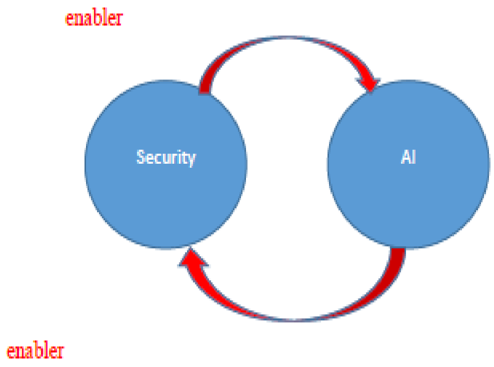


Figure 10: Security and AI System

Authors' Compilation

In fact, to improve computer security, nowadays, must be used AI systems that learn from the "experience" (Big data) to classify vulnerabilities. But, to the other side and AI system must be classified "secure" and controlled by the computer security.

Three levels of security

To control an AI system, it must be necessary to check the security at 3 levels:

1) Software level

2) Learning level

3) Distributed level

The first level is the classical software level. It includes the static code analysis, the programming vulnerabilities, the language vulnerabilities.

The learning level is the ML level. It must be necessary to control the data inserted in the database and how the machine reacts to some input of data.

The distributed level is a level used when an AI system is composed by many instances that resolve different tasks to join a final single decision. It must be necessary that every instance took the right decision to have a right final decision *(Crootof, R. and Renz, F. 2017).*

Software level

To control the software level, it can be used the classical process to integrate Quality and Security in SDLC.

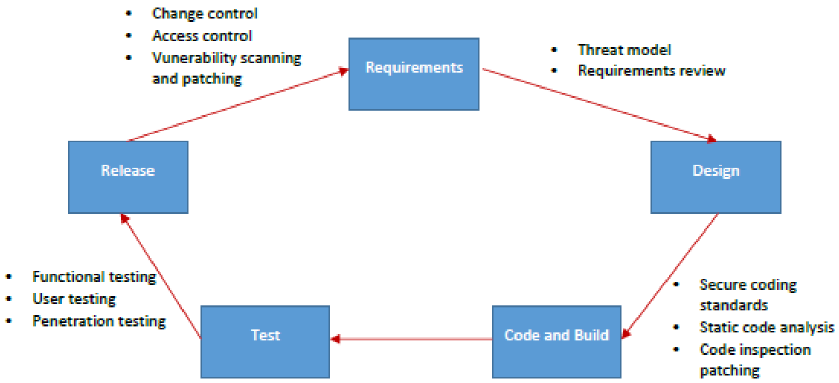
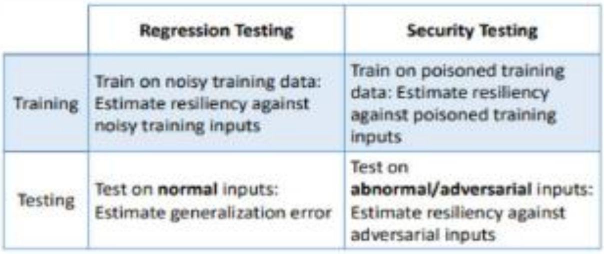


Figure 11: Software Development Life Cycle; Authors' Compilation, August 27, 2025

Some tasks as vulnerability scanning and patching, static code analysis or code inspection may be implemented using an AI system that may do the task better than a human.

Learning level

To be an AI system, the machine must learn with a ML algorithm how to do a task. So, it must be validated the ML algorithm of the machine. To control at this level the machine must receive a lot of different data as input and it must be checked how the machine reacts and how is the error of the machine. To do so are necessary regression testing and security testing. As defined in the Microsoft summit:



	Regression Testing	Security Testing
Training	Train on noisy training data: Estimate resiliency against noisy training inputs	Train on poisoned training data: Estimate resiliency against poisoned training inputs
Testing	Test on **normal** inputs: Estimate generalization error	Test on **abnormal/adversarial** inputs: Estimate resiliency against adversarial inputs

Table 3: AI Security Learning, Training and Testing Level

The problem today is that that test often fails because there are a lot of adversarial schemes that an AI system encounters during is lifecycle. For example, is possible for an automatically driven car knows all the possible images that may encounter in the world. So, it can be possible that, in front of rare situations, cannot recognize some patterns and does some wrong behaviour that may be fatal.

Distributed level

This level is related to the control of the local instances of a distributed AI system. It consists to validate all the results that came from the different instances of the system. To do so, is necessary to check all the instances at the previous two levels.

7. CONFLICT OF INTEREST

The research on Security Implications of Artificial Intelligence in Machine Learning Systems may involve potential conflicts of interest arising from academic, industrial, and governmental stakeholders. Since AI and ML technologies are increasingly adopted in security-sensitive domains such as cybersecurity, defense, finance, and surveillance, competing interests may influence the development, deployment, and evaluation of such systems.

First, corporate interests in commercializing AI-driven security products may create tension with ethical responsibilities, as organizations might prioritize profit and market advantage over transparency, fairness, or safety. Similarly, governmental interests in using AI for surveillance or national security may conflict with human rights, privacy, and civil liberties.

Second, research sponsorship and funding sources can bias outcomes. When studies are funded by technology companies or security agencies, there is a risk of emphasizing the strengths of AI solutions while downplaying potential vulnerabilities, ethical risks, or societal implications.

Third, academic conflicts may arise where researchers working on AI security also serve as consultants for private firms or governments, potentially influencing objectivity. This could result in selective reporting, suppression of negative results, or biased policy recommendations.

Finally, dual-use concerns in AI security research—where innovations meant for protection could also be exploited for malicious purposes—pose a conflict between advancing knowledge and preventing misuse. Researchers must balance open scientific communication with responsible disclosure to minimize risks.

8. ETHICAL CONSIDERATION

The following are some of the Ethical consideration of this Research.

1. **Data Privacy and Protection:** AI and ML systems depend on vast datasets, which may contain sensitive personal or organizational information. Ethical concerns arise when such data is improperly collected, shared, or used without consent. Researchers and practitioners must ensure compliance with data protection laws (e.g., GDPR, NDPR) and adopt anonymization or encryption techniques to safeguard privacy.
2. **Bias, Fairness, and Discrimination:** Machine learning algorithms can unintentionally reinforce biases present in training data. This may lead to unfair treatment of individuals or groups in security applications (e.g., facial recognition or predictive policing). Ethically, system designers must implement fairness-aware models and conduct bias audits to ensure inclusivity.
3. **Transparency and Explainability**: AI-driven security systems often operate as “black boxes,” making their decisions difficult to interpret. Ethical responsibility demands that these systems provide explainable outputs, so stakeholders can understand how security-related decisions are made and challenge them when necessary.
4. **Accountability and Responsibility:** Security breaches caused by AI-driven systems raise questions of responsibility: is it the developer, the deploying organization, or the AI itself? Ethical frameworks must clearly define accountability lines, especially in high-risk areas like national security, healthcare, or finance.
5. **Dual-Use and Misuse Risks:** AI research for security can be repurposed for malicious activities such as cyberattacks, surveillance abuse, or automated hacking. Ethical practice requires careful dissemination of research findings, adoption of usage restrictions, and adherence to responsible innovation principles.
6. **Human Autonomy and Oversight:** Over-reliance on autonomous AI security systems may reduce human decision-making power. Ethical safeguards must ensure that humans remain in control, with AI serving as a support tool rather than a replacement in critical security contexts.
7. **Global Inequality and Access:** Advanced AI security systems are often controlled by resource-rich nations or corporations, potentially creating security imbalances. Ethically, there is a duty to consider equitable access and to avoid exacerbating digital divides.

9. ACKNOWLEDGEMENT

I wish to express my deepest gratitude to Almighty God for the strength, wisdom, and guidance throughout this research journey. My sincere appreciation goes to my supervisors and lecturers for their invaluable support, constructive feedback, and encouragement, which shaped this work. I also extend my heartfelt thanks to colleagues, cybersecurity professionals from Center for Cyberspace, Department of Cybersecurity (Prof. Gilbert Aimufua; Director, Center for Cyberspace Studies, Nasarawa State University, keffi, Nigeria), Dr. M.O. Adenomon of the Department of Stastitics, Umar Musa; A PhD Colleague, Centre for Cyberspace Studies, Department of Cybersecurity, Nasarawa State University, Keffi, Nigeria, and AI researchers who shared insights and experiences that enriched this study. Special appreciation goes to my family and friends for their patience, love, and moral support. Without their prayers and motivation, this research on AI and security implications would not have been possible.

10. CONCLUSION

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into security systems presents both transformative opportunities and profound challenges. On one hand, these technologies enhance the detection of anomalies, enable predictive analytics, and automate responses to threats with remarkable speed and efficiency. They have become indispensable in strengthening cybersecurity infrastructures, safeguarding critical assets, and ensuring resilience against evolving attacks. However, the study of their security implications also uncovers significant risks. AI and ML models remain vulnerable to adversarial attacks, data poisoning, model inversion, and system manipulation, which can compromise entire security architectures. Moreover, the reliance on vast datasets introduces privacy concerns, while issues of bias, opacity, and lack of accountability challenge ethical and societal trust.

The findings suggest that AI-enabled security systems are double-edged: they not only provide defensive strength but also create new attack surfaces that malicious actors can exploit. Dual-use risks further complicate the ethical debate, as tools designed for protection can equally serve offensive or harmful purposes. These realities underscore that technological advancement alone is insufficient. For AI in security to be sustainable and trustworthy, it must be guided by ethical principles, robust governance, and continuous risk assessment. Ultimately, the successful deployment of AI in ML security systems requires balancing innovation with caution, automation with human oversight, and efficiency with accountability. Thus, the future of AI in security depends not merely on technological breakthroughs but on collective responsibility to ensure its safe, equitable, and ethical use.

11. RECOMMENDATION

To address the security implications of AI in Machine Learning systems, several actionable recommendations are essential. First, governance and regulation must be prioritized. Policymakers should establish clear legal and ethical frameworks that guide AI adoption in security-sensitive sectors, ensuring compliance with data protection, transparency, and accountability standards. Second, technical safeguards should be strengthened. Developers must embed security-by-design principles, conduct adversarial testing, and employ robust defenses such as encryption, federated learning, and anomaly detection to reduce vulnerabilities.

Third, explainability and auditability should be mandatory features in AI security systems. Transparent and interpretable models will enhance trust, facilitate error detection, and ensure fair decision-making. Fourth, human oversight must not be eliminated. AI should serve as a support tool, while final decisions—especially in critical areas like defense, finance, and healthcare—remain under human control to preserve accountability and ethical judgment.

Fifth, capacity building and awareness are crucial. Training security professionals, researchers, and policymakers on the risks and benefits of AI will ensure informed decision-making and prevent misuse. In addition, international collaboration is vital. Global cooperation among governments, industries, and academia can help standardize best practices, share threat intelligence, and prevent monopolization or weaponization of AI technologies.

Finally, research institutions should adopt responsible innovation practices, carefully balancing openness with security to prevent dual-use risks. By integrating these recommendations, stakeholders can create resilient, ethical, and trustworthy AI systems that not only advance security but also safeguard human rights and societal well-being.

REFERENCES

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning, Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), The Cambridge Handbook of Artificial Intelligence (pp. 316–334). Cambridge University Press.

Bryson, J. J., & Winfield, A. F. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. Computer, 50(5), 116–119. https://doi.org/10.1109/MC.2017.154

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538(7625), 311–313. https://doi.org/10.1038/538311a

European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).

Nguyen, T. T., Pathirana, P. N., Ding, M., & Seneviratne, A. (2022). Artificial intelligence in the battle against cybersecurity threats. IEEE Access, 10, 8572–8591. https://doi.org/10.1109/ACCESS.2022.3142526

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. AI & Society, 36(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2

Sharmeen, S., Hossain, M. S., Muhammad, G., & Alamri, A. (2020). Towards artificial intelligence-driven cybersecurity techniques for 6G networks. IEEE Network, 34(6), 224–230. https://doi.org/10.1109/MNET.011.2000176

University of Cambridge, Future of Humanity insititute, University of Oxford, Centre for the study of existensial risk, Center for a New American security, electronic frontier foundation, OpenAI, February 2018. "The malicious use of Artificial Intelligence: Forecasting, Prevention and Mitigation"

IoT Security Foundation, 2016. "IoT Security Compliance Framework"

Infosecurity Magazine, January 2017. "How AI is the Future of Cybersecurity", Ryan Kh

Taesoo Kim, Dawn Song, Walker Michael, 2017. "AI and security" Microsoft Faculty Summit

Aker, C. and Kalkan, S. 2017. "Using Deep Networks for Drone Detection" arXiv preprint server

Allen, G. and Chan, T. 2017. "Artificial Intelligence and National Security" Harvard Kennedy

School Belfer Center for Science and International Affairs

Booth, S., Tompkin, J., Gajos, K., Waldo, J., Pfister, H., Nagpal, R. 2017. "Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security"

Crootof, R. 2015. "The Killer Robots are Here: Legal and Policy Implications"

Cylance. 2017. "Black Hat Attendees See AI As Double-Edged Sword" The Cylance Team

Carbon Black, 2017. “Beyond the Hype: Security Experts Weigh in on Artificial Intelligence, Machine Learning, and Non- Malware Attacks”

CNAS. 2017. “Artificial Intelligence and Global Security Summit,” Centre for New American Security”

Crawford, K. and Calo, R. 2016. “There is a blind spot in AI research” Nature, October 13, 2016

Crootof, R. and Renz, F. 2017. “An Opportunity to Change the Conversation on Autonomous Weapon Systems,” Lawfare

Cooper, D.M. 2013. “A Licensing Approach to Regulation of Open Robotics” paper presented at We Robot, April 2013

Dao, James. “Drone Pilots Are Found to Get Stress Disorders Much as Those in Combat Do” New York Times

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde- Farley, D., Ozair,

S., Courville, A. and Bengio, Y., 2014. “Generative Adversarial Networks” In Advances in Neural information Processing Systems 2014

Goodman, B. and Flaxman, S., 2016. “European Union regulations on algorithmic decision making and a right to explanation"

IEEE Standards Association, 2017. “The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems