Deploying Artificial Intelligence into Daily Life: Artificial Intelligence for Cyber Security with more Opportunities

[1]Dr. S. China Venkateswarlu,
Professor, Dept. of ECE,
Institute of Aeronautical Engineering, Hyderabad, Telangana.
cvenkateswarlus@gmail.com
[2]SHIVA SHANKAR J
Research Scholar,
Department of Electronics & Instrumentation Engineering,
Annamalai University.
shivashankar.jss@gmail.com

[3]Dr. S.Palanivel,
Associate Professor, Dept. of EIE,
Annamalai University.
s_palanivel@yahoo.com

Abstract

Deploying Artificial Intelligence into Daily Life: Summary: The chapter "Deploying Artificial Intelligence into Daily Life" focuses on the practical applications and integration of artificial intelligence (AI) technologies into various. Aspects of our everyday lives. It explores how AI is being deployed to enhance efficiency, convenience, and decision-making in different domains. With the advances in information technology (IT), the law breakers are using cyberspace and indulging in many digital violations. Developing trends of complex, distributed and Internet computing are raising important questions on information security and privacy. Cyber infrastructures are highly vulnerable to intrusions and other threats. Physical devices such as sensors and detectors are not sufficient for monitoring and protection of these infrastructures; hence, there is a need for more sophisticated IT that can model normal behaviours and detect abnormal ones. These cyber defence systems need to be flexible, adaptable and robust, and able to detect a wide variety of threats and make intelligent real-time decisions. With the pace and amount of cyber attacks, human intervention is simply not sufficient for timely attack analysis

and appropriate response. The fact is that the most network-centric cyber attacks are carried out by intelligent agents such as computer worms and viruses; hence, combating them with intelligent semi-autonomous agents that can detect, evaluate, and respond to cyber attacks has become a requirement. These so called computer-generated forces will have to be able to manage the entire process of attack response in a timely manner, i.e. to conclude what type of attack is occurring, what the targets are and what is the appropriate response, as well as how to prioritize and prevent secondary attacks . Furthermore, cyber intrusions are not localized. They are a global menace that poses threat to any computer system in the world at a growing rate. There were times when only educated specialist could commit cyber crimes, but today with the expansion of the Internet, almost anyone has access to the knowledge and tools for committing these crimes. Conventional fixed algorithms (hard-wired logic on decision making level) have become ineffective against combating dynamically evolving cyber attacks. This is why we need innovative approaches such as applying methods of Artificial Intelligence (AI) that provide flexibility and learning capability to software which will assist humans in fighting cyber crimes . AI offers this and various other possibilities. Numerous nature-inspired computing methods of AI such as Computational Intelligence, Neural Networks, Intelligent Agents, Artificial Immune Systems, Machine Learning, Data Mining, Pattern Recognition, Fuzzy Logic, Heuristics, etc., have been increasingly playing an important role in cyber crime detection and prevention. AI enables us to design autonomic computing solutions capable of adapting to their context of use, using the methods of self-management, self-tuning, self-configuration, self-diagnosis, and self healing. When it comes to the future of information security, AI techniques seem very promising area of research that focuses on improving the security measures for cyber space.

The term Artificial intelligence is used when a machine be haves like a human in activities such as problem solving or learning, which is also known as machine learning. The next generation of cyber security products is increasingly incorporating Artificial Intelligence and Machine Learning technologies. AI software on large datasets of cyber security, network, and even physical information, cyber security solutions providers aim to detect and block abnormal behaviour. There are different approaches to using AI for cyber security. Some software applications analyze raw network data to spot an irregularity, while others focus on user-entity

behaviour to detect patterns that deviate from normal. The types of data streams and the level of effort needed by analysts all vary by approach.

Keywords: Artificial Intelligence**, d**ata streams, deep learning, machine learning technologies, intelligence security, open-source software tools.

1. Introduction to Artificial Intelligence

1.1 Introduction

Artificial Intelligence one of the booming technologies of computer science which is creating a new revolution in the world by making intelligent machines. The Artificial Intelligence is now all around us. It is currently working with a variety of subfields, ranging from general to specific, such as self-driving cars, playing chess, proving theorems, playing music, Painting, etc.

AI is one of the fascinating and universal fields of Computer science which has a great scope in future. AI holds a tendency to cause a machine to work as a human.

Artificial Intelligence is composed of two words Artificial and Intelligence, where Artificial defines "man-made," and intelligence defines "thinking power", hence AI means "a man-made thinking power."

1.12 Definition

The Artificial Intelligence can be defined as “A branch of computer science by which intelligent machines can be created which can behave like a human, think like humans, and able to make decisions."

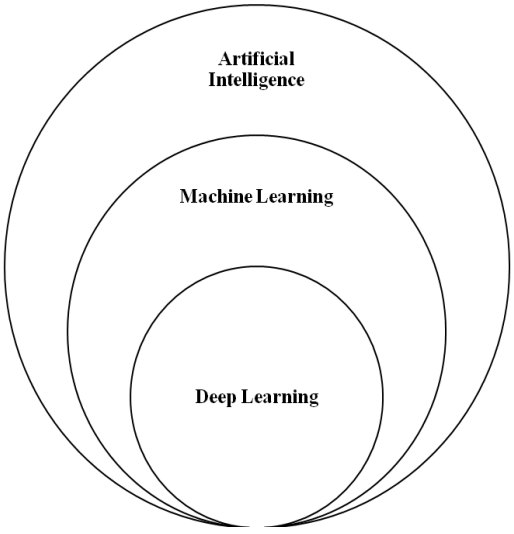

Figure 1.1. AI and its sub components

Artificial Intelligence (AI) technologies are rapidly moving beyond the realms of academia and speculative fiction to enter the commercial mainstream. Innovative products such as Apple's Siri® digital assistant and the Google search engine, among others, are utilizing AI to transform how we access and utilize information online.

Advances in Artificial Intelligence (AI) technology and related fields have opened up new markets and new opportunities for progress in critical areas such as health, education, energy, economic inclusion, social welfare, and the environment

AI has also become strategically important to national defence and securing our critical financial, energy, intelligence, and communications infrastructures against state-sponsored cyber-attacks.AI has important applications in cyber security, and is expected to play an increasing role for both defensive and offensive cyber measures. Using AI may help maintain the rapid response required to detect and react to the landscape of evolving threats.

Like every important new technology, AI has occasioned both excitement and apprehension among industry experts and the popular media. We read about computers that beat Chess and Go masters, about the imminent superiority of self-driving cars, and about concerns by some ethicists that machines could one day take over and make humans obsolete. We believe that some of these fears are over-stated and that AI will play a positive role in our lives as long AI research and development is guided by sound ethical principles that ensure the systems we build now and in the future are fully transparent and account-able to humans.

1.2 Advantages of Artificial Intelligence

The following are some main advantages of Artificial Intelligence:

High Accuracy with fewer errors: AI machines or systems are prone to less errors and high accuracy as it takes decisions as per pre-experience or information.

High-Speed: AI systems can be of very high-speed and fast-decision making; because of that AI systems can beat a chess champion in the Chess game.

High reliability: AI machines are highly reliable and can perform the same action multiple times with high accuracy.

Useful for risky areas: AI machines can be helpful in situations such as defusing a bomb, exploring the ocean floor, where to employ a human can be risky.

Digital Assistant: AI can be very useful to provide digital assistant to the users such as AI technology is currently used by various E-commerce websites to show the products as per customer requirement.

Useful as a public utility: AI can be very useful for public utilities such as a self-driving car which can make our journey safer and hassle-free, facial recognition for security purpose, Natural language processing to communicate with the human in human-language, etc.

1.3 Disadvantages of Artificial Intelligence

Every technology has some disadvantages, and the same goes for Artificial intelligence. Being so advantageous technology still, it has some disadvantages which we need to keep in our mind while creating an AI system. Following are the disadvantages of AI:

High Cost: The hardware and software requirement of AI is very costly as it requires lots of maintenance to meet current world requirements.

Can't think out of the box: Even we are making smarter machines with AI, but still they cannot work out of the box, as the robot will only do that work for which they are trained, or programmed.

No feelings and emotions: AI machines can be an outstanding performer, but still it does not have the feeling so it cannot make any kind of emotional attachment with human, and may sometime be harmful for users if the proper care is not taken.

Increase dependency on machines: With the increment of technology, people are getting more dependent on devices and hence they are losing their mental capabilities.

No Original Creativity: As humans are so creative and can imagine some new ideas but still AI machines cannot beat this power of human intelligence and cannot be creative and imaginative.

2. Introduction to Cyber Security

2.1 History

The history of cyber security began as a research project. In the 1970's, Robert Thomas, a researcher for BBN Technologies in Cambridge, Massachusetts, created the first computer "worm". It was called The Creeper. The Creeper, infected computers by hopping from system to system with the message "I'M THE CREEPER: CATCH ME IF YOU CAN." Ray Tomlinson, the inventor of email, created a replicating program called The Reaper, the first antivirus software, which would chase Creeper and delete it.

Late in 1988, a man named Robert Morris had an idea: he wanted to test the size of the internet. To do this, he wrote a program that went through networks, invaded Unix terminals, and copied itself. The Morris worm was so aggressive that it slowed down computers to the point of being unusable. He subsequently became the first person to be convicted under Computer Fraud and Abuse Act.

From that point forward, viruses became deadlier, more invasive, and harder to control. With it came the advent of cyber security.

2.12 Definition

Cyber security is the body of technologies, processes, and practices designed to protect networks, computers, programs and data from attack, damage or unauthorized access.

The term cyber security refers to techniques and practices designed to protect digital data. The data that is stored, transmitted or used on an information system. After all, that is what criminal wants, *data*. The network, servers, computers are just mechanisms to get to the data. Effective

cyber security reduces the risk of cyber-attacks and protects organizations and individuals from the unauthorized exploitation of systems, networks, and technologies.

Robust cyber security implementation is roughly based around three key terms: people, processes, and technology. This three-pronged approach helps organizations defend themselves from both highly organized attacks and common internal threats, such as accidental breaches and human error.

The attacks evolve every day as attackers become more inventive, it is critical to properly define cyber security and understand cyber security fundamentals.

Listed below are the reasons why cyber security is so important in what's become a predominant digital world:

- With each passing year, the sheer volume of threats is increasing rapidly. According to the report by McAfee, cybercrime now stands at over $400 billion, while it was $250 billion two years ago.
- Cyber attacks can be extremely expensive for businesses to endure. In addition to financial damage suffered by the business, a data breach can also inflict untold reputational damage.
- Cyber-attacks these days are becoming progressively destructive. Cybercriminals are using more sophisticated ways to initiate cyber attacks.
- Regulations such as GDPR are forcing organizations into taking better care of the personal data they hold.

Because of the above reasons, cyber security has become an important part of the business and the focus now is on developing appropriate response plans that minimize the damage in the event of a cyber attack. But, an organization or an individual can develop a proper response plan only when he has a good understanding on the fundamentals of cyber security.

2.2 The CIA Triad

Confidentiality, integrity, and availability, also known as the CIA triad, is a model designed to guide companies and organizations to form their security policies. Technically, cyber security means protecting information from unauthorized access, unauthorized modification, and unauthorized deletion in order to provide confidentiality, integrity, and availability.

2.2.1 Confidentiality

Confidentiality is about preventing the disclosure of data to unauthorized parties. It also means trying to keep the identity of authorized parties involved in sharing and holding data private and anonymous. Often confidentiality is compromised by cracking poorly encrypted data, Man-in-the-middle (MITM) attacks, disclosing sensitive data.

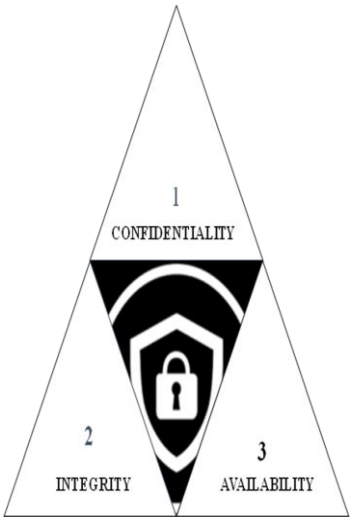

Figure 1.2. The CIA Triad

Standard measures to establish confidentiality include:

- Data encryption
- Two-factor authentication
- Biometric verification
- Security tokens

2.3 Integrity

Integrity refers to protecting information from being modified by unauthorized parties. It is a requirement that information and programs are changed only in a specified and authorized manner. Challenges that could endanger integrity include turning a machine into a “zombie computer”, embedding malware into web pages.

Standard measures to guarantee integrity include:

- Cryptographic checksums
- Using file permissions
- Uninterrupted power supplies
- Data backups

2.4 Availability

Availability is making sure that authorized parties are able to access the information when needed. Data only has value if the right people can access it at the right time. Information unavailability can occur due to security incidents such as DDoS attacks, hardware failures, programming errors, human errors.

Standard measures to guarantee availability include:

- Backing up data to external drives
- Implementing firewall
- Having backup power supplies
- Data redundancy

All cyber attacks have the potential to threaten one or more of the three parts of the CIA triad. Confidentiality, integrity, and availability all have to work together to keep the information secured. Hence it is very important to understand what the CIA Triad is, how it is used to plan and implement a quality security policy while understanding the various principles behind it.

3. AI: Perception Vs Reality

The field of AI actually encompasses three distinct areas of research:

- **Artificial Super Intelligence (ASI)** is the kind popularized in speculative fiction and in movies such as The Matrix. The goal of ASI research is to produce computers that are superior to humans in virtually every way.
- **Artificial General Intelligence (AGI)** refers to a machine that's as intelligent as a human and equally capable of solving the broad range of problems that require learning and reasoning. One of the classic tests of AGI is the ability to pass what has come to be known as "The Turing Test," in which a human evaluator reads a text-based conversation occurring remotely between two unseen entities, one known to be a human and the other a machine. To pass the test, the AGI system's side of the conversation must be

indistinguishable by the evaluator from that of the human. Most experts agree that we're decades away from achieving AGI and some maintain that ASI may ultimately prove unattainable.

- **Artificial Narrow Intelligence (ANI)** exploits a computer's superior ability to process vast quantities of data and detect patterns and relationships that would otherwise be difficult or impossible for a human to detect.

Such data-centric systems are capable of outperforming humans only on specific tasks, such as playing chess or detecting anomalies in network traffic that might merit further analysis by a threat hunter or forensic team.

The field of Artificial Intelligence encompasses a broad range of technologies intended to endow computers with human-like capabilities for learning, reasoning, and drawing useful insights. In recent years, most of the fruitful research and advancements have come from the sub-discipline of AI named Machine Learning (ML), which focuses on teaching machines to learn by applying algorithms to data. Often, the terms AI and ML are used interchangeably. In this book, how-ever, we'll be focusing exclusively on methods that fall within the machine learning space.

Not all problems in AI are candidates for a machine learning solution. The problem must be one that can be solved with data; a sufficient quantity of relevant data must exist and be acquirable; and systems with sufficient computing power must be available to perform the necessary processing within a reasonable time frame.

3.1 Machine Learning in the Security Domain

In order to pursue well-defined goals that maximize productivity, organizations invest in their system, information, network, and human assets. Consequently, it's neither practical nor desir-able to simply close off every possible attack vector. Nor can we prevent incursions by focusing exclusively on the value or properties of the assets we seek to protect. Instead, we must consider the context in which these assets are being accessed and utilized. With respect to an attack on a website, for example, it's the context of the connections that matters, not the fact that the attacker is targeting a particular website asset or type of functionality.

Context is critical in the security domain. Fortunately, the security domain generates huge quantities of data from logs, network sensors, and endpoint agents, as well as from distributed directory and human resource systems that indicate which user activities are permissible and which are not. Collectively, this mass of data can provide the contextual clues we need to identify and ameliorate threats, but only if we have tools capable of teasing them out. This is precisely the kind of processing in which ML excels.

By acquiring a broad understanding of the activity surrounding the assets under their control, ML systems make it possible for analysts to discern how events widely dispersed in time and across disparate hosts, users, and networks are related. Properly applied, ML can provide the context we need to reduce the risks of a breach while significantly increasing the "cost of attack."

3.2 **The Future of Machine Learning**

As ML proliferates across the security landscape, it's already raising the bar for attackers. It's getting harder to penetrate systems today than it was even a few years ago. In response, attackers are likely to adopt ML techniques in order to find new ways through. In turn, security professionals will have to utilize ML defensively to protect network and information assets.

We can glean a hint of what's to come from the March 2016 match between professional Go player Lee Sedol an eighteen-time world Go champion, and AlphaGo a computer program developed at DeepMind, an AI lab based in London that has since been acquired by Google. In the second game, AlphaGo made a move that no one had ever seen before. The commentators and experts observing the match were flummoxed. Sedol himself was so stunned it took him nearly fifteen minutes to respond. AlphaGo would go on to win the best-of-five game series.

In many ways, the security postures of attack and defence are similar to the thrust and parry of complex games like Go and Chess. With ML in the mix, completely new and unexpected threats are sure to emerge. In a decade or so, we may see a landscape in which "battling bots" attack and defend networks on a near real-time basis. ML will be needed on the defence side simply to maintain parity.

Of course, any technology can be beaten on occasion with sufficient effort and resources. However, ML-based defences are much harder to defeat because they address a much broader region of the threat space than anything we've seen before and because they possess human-like capabilities to learn from their mistakes.

4. AI in Cyber Security

4.1 Introduction

With the advancement in technology, cyber-crimes are also increasing and getting complex. Cyber-criminals are launching sophisticated attacks that are putting modern security systems at risk. So, the cyber security industry is also evolving to meet the increasing security demands of companies. But, these defensive strategies of security professionals may also fail at some point.

To up their game and enhance their vulnerability detection mechanisms, companies are choosing Artificial Intelligence (AI). Artificial Intelligence in Cyber Security is aiding companies to safeguard their defence mechanisms. It is also assisting them in analyzing cyber crimes better.

4.2 Impact of Artificial Intelligence on Cyber Security

Companies are focusing more on cyber security right now like never before. This is because advanced cyber security attacks have cost companies millions of dollars in data breaches. It starts with designing a multi-layered security system that will secure the network infrastructure. The first step is to install a firewall that will filter out the network traffic.

Then, antivirus software is used for cleaning out the malicious files and viruses in the infrastructure. As a part of their disaster recovery plan, regular data backups are executed.

And, this is where artificial intelligence comes in.

AI has impacted security by helping professionals to identify irregularities in the network by analyzing user actions and studying the patterns. Security professionals can now study network

data using AI and detect vulnerabilities to prevent harmful attacks. AI will help enhance the traditional security approach by the following ways –

- Advanced AI-powered security tools will be used to monitor and respond to security events
- Modern firewalls will have built-in machine learning technology that will easily detect a usual pattern in the network traffic and remove it if considered malicious
- Using the natural language processing feature in AI, security professionals can detect the origin of a cyber-attack. Natural language processing also helps in analyzing vulnerabilities
- Scanning internet data and using predictive analysis will identify malicious threats beforehand
- Higher security of conditional access and authentication

Another important revelation of Artificial Intelligence in Cyber Security is biometric login systems. These are extremely secure logins that use fingerprints, retina scans, and palm-prints. A password can be used along with this biometric information for securely logging in. This method is used in organizations for employees to log in and even in smart phones.

4.3 Applications in Cyber security

Machine learning, a very important subset of artificial intelligence, is also being used these days by corporations to enhance their security systems. Besides helping security experts in detecting malicious attacks, it has the following applications –

4.3.1 Mobile endpoint security

Machine learning is used for mobile endpoint security as smart phones, tablets, and notebooks are all prone to cyber-attacks. A company called Wandera recently launched its machine learning-powered threat detection engine called MI: RIAM. This engine has successfully detected several traces of repackaged SLocker Ransomware that targets mobile endpoints.

4.3.2 No zero-day vulnerabilities

Zero-day vulnerability is a threat that is completely new to the security professional and he or she does not yet have a solution or patch to fix it. Zero-day means that professionals have zero days to fix the issue, and they may have already been exploited by an attacker. These threats are sometimes found in unsecured IoT devices.

Machine learning algorithms can detect zero-day threats by analyzing the anomalies in network traffic. Vulnerabilities are removed and patch exploits are prevented using machine learning.

4.3.3 Improving human analysis

Machine learning helps in enhancing human analysis in cyber security activities such as vulnerability assessment, threat detection, network analysis, and endpoint security. ML algorithms can filter out suspicious data in the network and pass it on to a human security analyst. As a result, the alert detection rates can increase significantly.

4.3.4 Automating security tasks

Repetitive and boring security tasks can be reduced by machine learning. This helps professionals to focus on important jobs. Tasks like checking network traffic, interrupting threats such as ransomware, removing viruses, and analyzing network logs can be automated by machine learning.

Human security resources can also be allocated efficiently with the help of machine learning.

4.4 Companies using Artificial Intelligence in Cyber Security

AI-powered systems are used by the following companies to strengthen their security infrastructure –

Google

They are using the Deep Learning AI system on their Cloud Video Intelligence platform. Videos stored on their cloud server are analyzed by AI algorithms based on their content and context. If an anomaly is found that might be a threat, the AI algorithms send an alert.

Gmail uses machine learning to filter out spams from your mail to provide a hassle-free environment. More than 100 million spams are blocked every day.

IBM

IBM Watson uses machine learning in its cognitive training to detect threats and create cyber security solutions. AI also reduces time-consuming threat research tasks and assists in determining security risks.

5. Security of AI

5.1 Introduction

Recent advances in AI are transformative and already exceed human-level performance in tasks like image recognition, natural language processing, and data analytics. Economic factors will drive the adoption of new AI applications that disrupt almost every aspect of the enterprise both good and bad. AI-systems can be manipulated, evaded, and misled resulting in profound security implications for applications such as network monitoring tools, financial systems, or autonomous vehicles. Therefore, secure and resilient techniques and best practices are vitally important.

5.2 Specification and Verification of AI Systems

Integrated AI systems involve four components: perception, learning, decisions, and actions. These systems operate in complex environments that require each component to interact and be interdependent (e.g., errors in perception can cause an incorrect decision). Furthermore, there are unique vulnerabilities in each of the components (e.g., perception is prone to training attacks while decisions are susceptible to classic cyber exploits). Finally, the notion of correctness is not a purely logical matter; noise and uncertainty require bounds for each component to protect the system from misbehaving. There is a pressing need for formal methods to verify AI and ML

components, both independently and in concert, as it relates to logical correctness, decision theory, and risk analysis. New techniques are needed that specify what a system is expected to do and how it should respond to attack. In traditional systems, qualities that match the specification are tractable for each component. Because AI systems are so complex, their implementation and configuration are difficult to assess. Research is needed in architectural structures and analysis techniques that allow verification of these components and is part of a larger effort to develop manageable standards, best practices, tools, and methods to reason about the behaviour of a system. A new discipline and science of AI architecture could produce an AI "building code". Such a code could come from theory and experience, capture best practices, and leverage guidelines from other computer science areas. Analysis of the building code would lead to a better understanding of AI mechanisms and move the field forward. Specification and verification must also address aspects such as performance, security, robustness, and fairness. Research is needed to better understand performance tradeoffs, the operating environment, and may require a domain expert on the team. And finally, an engineer must be identified to implement, deploy, and maintain the AI system.

5.3 Trustworthy AI Decision Making

As AI systems are deployed in high-value environments, the issue of ensuring that the decision process is trustworthy, particularly in adversarial scenarios, is paramount. While there are numerous illustrations of ML vulnerabilities, science-based techniques to predict trustworthiness are elusive. Research is needed to develop methods and principles for a wide array of AI systems, including ML, planning, reasoning, and knowledge representation. Areas that need to be addressed for trustworthy decision making include defining performance metrics, developing techniques, making AI systems explainable and accountable, improving domain-specific training and reasoning, and managing training data. Threat model research must identify measurable properties that define trustworthiness so a defender can incorporate robustness, privacy, and fairness into decision-making algorithms. Given a specific threat model, the system will have to reason about adversarial interference and define requisite conditions to achieve these trustworthiness properties. Possibilities include adapting definitions from cryptography or computer security, unifying properties into a single reasoning framework, and treating them as variants of a single notion of (in)stability in ML and AI for both decision making and for security

models more broadly. Research is also needed in methods for understanding the learned reasoning of AI methods, particularly deep learning. How do certain data points influence the optimization procedures, and the reasoning, involved in ML systems? Possibilities include analysis of the optimization procedure, or the AI system outcome, if it captures both the training data and the learning method. Techniques that can estimate a training point's influence on individual predictions could also become the basis to assess the relevance of a model in a decision environment. In ML, there are approaches emerging that provide decision guarantees using a variety of techniques (e.g., convex relaxation of the adversarial optimization problem and randomized smoothing). However, the approaches are currently focused almost exclusively on supervised learning and are difficult to achieve without degrading system performance. A related area of research, AI systems that request guidance when they are uncertain, can improve trust in the eventual decision and allow the system to obtain information for future decision making. The accuracy of AI is also domain sensitive. Security vulnerabilities arise when training data is not representative of the given environment. Conversely, overly pessimistic vulnerability assessments can occur if constraints in the application domain are not considered. Research is needed on how input data is acquired, secured, maintained, and evaluated within domain-specific AI environments, and as they become a part of the full-use ecosystem. An autonomous vehicle system is trained with images and situations acquired from realistic environments and maintained constantly as its environment changes. Perception, planning, reinforcement learning, knowledge representation, and reasoning are all domain-specific vulnerabilities that need to be considered. This includes reasoning about streaming data, weighing consequences (e.g., causing a car to crash or go in the wrong direction), and adapting to unanticipated events (e.g., weather or road construction). Domain specificity research necessitates a rethinking of threat models and helps deploy and maintain AI systems in real-world environments. Researchers must also evaluate the cost/benefit ratio of collecting, protecting, and storing training data. Datasets are valuable (e.g., large network datasets can reveal everything about network vulnerabilities). Proper collection and storage can protect data and provide information for defence. But what if the data is of higher value for an adversary, should it be collected?

5.4 Detection and Mitigation of Adversarial Inputs

While AI performs well on many tasks, it is often vulnerable to corrupt inputs that produce inaccurate responses from the learning, reasoning, or planning systems. There are examples where deep learning methods can be fooled by small amounts of input noise crafted by an adversary. Such capabilities allow adversaries to control the systems with little fear of detection. As systems based on deep networks and other ML and AI algorithms become integrated into operational systems, it is critical to defend against adversarial inputs by considering more robust machine learning methods, AI reconnaissance prevention, the study of adversarial models, model poisoning prevention, secure training procedures, data privacy, and model fairness. Efforts are needed to harden learning methods against adversarial inputs. This problem is well understood in both the statistics and technical communities. Both theoretical and empirical research is needed to make the same advances for deep learning and modern ML methods without sacrificing performance or accuracy. Modern AI systems are vulnerable to reconnaissance where adversaries query the systems and learn the internal decision logic, knowledge bases, or the training data. This is often a precursor to an attack to extract security-relevant training data and sources or to acquire the intellectual property embedded in the AI. The following are possible reconnaissance prevention measures that need research:

- Increase the attacker workload and reduce their effectiveness through model inversion.
- Leverage cyber security approaches, including rate limiting, access controls, and deception.
- Study the impacts on accuracy and other aspects of algorithms and systems.
- Design reconnaissance-resistant algorithms and techniques.
- Integrate resistance into learning and reasoning optimizations.
- Embed security guarantees into the model using new multistep techniques.
- Expose the presence and goals of the attacker using the cyber security honeypot concept.

The vulnerability of an AI system is defined by the adversary's knowledge and capabilities. Research is needed to classify the different types of attacks and develop appropriate defences. Defences need to address attacks based on the type of information the attacker has access to. These models should be carefully mapped, attack and defence strategies identified,

and special research attention given to security critical domains where ML models are most at risk. (e.g., autonomous vehicles and malware detection).

AI and ML models learn how to characterize expected inputs from training data. If the training instances do not represent all possible and future situations, then the model outputs will be inaccurate. This creates a security scenario where an attacker can manipulate the model and introduce an exploitable backdoor. An adversary can control a fraction of the training set and still influence the behaviour of the model (model poisoning). ML requires as much data as possible and it is common, but also risky, to use many data sources. If even one source of data is malicious, the entire model becomes untrustworthy. To both mitigate adversarial poisoning and improve training processes, AI best practices must ensure the end-to-end provenance of training data and the detection of data that falls outside the normal input space. ML methods work well when they are used with similar data to what they were trained on and fails when the data is different (e.g., a self-driving car trained in sunny, cloudy, rainy, and snowy weather might operate poorly in sleet or hail). These are common problems because it is difficult to acquire data for all possible situations. Systems typically do not recognize abnormal data, even when a human would. The research goal is to increase the detection of anomalies, adopt training methods that amplify rare events, and allow the most effective use of existing training data and algorithms. To remain effective and accurate, ML models must be retrained frequently (e.g., social media terminology used for public sentiment analysis changes over time as vocabulary and topics of interest change). Research is needed to identify what training data to collect, when such training data is no longer relevant, and how often models should be retrained. Recent attacks have shown that an adversary can determine whether a data item was used in training a model. Because many applications require ML training using private data, this puts sensitive information at risk. Further research is needed, but advances, such as differential privacy, provide new pathways to anonymize data and prevent leaks. Finally, models will learn whatever biases and discriminatory features are present in training data. If the data reflects discrimination against a given community (e.g., in college admissions or loan approvals), that bias will appear in the outcome. Prevention of outcome bias will require scientific and technical foundations for ML

fairness to be developed. Goals must be defined and algorithmic techniques developed to measure, detect, and diagnose unfair ML training data and methods.

5.5 Engineering Trustworthy AI-Augmented Systems

New understanding of how vulnerable AI components are to adversarial action raises concerns about the safety of the entire data processing pipeline in which they are used. AI components defy conventional software analysis and can introduce new attack vectors in environments where the AI algorithms operate, implementations of AI frameworks and applications, ML models, and training data. Due to hidden dependencies in the pipeline, multiple applications can be effected. Research is needed to develop theory, engineering principles, and best practices when using AI as a component of a system. This should include threat modelling, security tools, domain vulnerabilities, and securing human machine teaming. These models need to enable iterative abstractions of attacks and refinements, be designed in accord with an AI expert, and consider data availability and integrity, access controls, network orchestration and operation, resolution of competing interests, privacy, and a dynamic policy environment. To make AI-enabled systems more trustworthy, engineering principles should be based on science, community experience, and AI component functionality research that includes redundancy (e.g., ensemble), supervisory, and other frameworks. Understanding the conditions, threats, domains, and constraints are necessary but subsidiary goals. Once overall system AI vulnerabilities are understood, traditional cyber security and robust system design can reduce the impact (e.g., to ensure AI training data is more difficult to poison); allow more redundancy and diversity to be built in (e.g., an autonomous vehicle may use lidar, radar, image processing, and map information); develop robust system architectures that can withstand AI component failures and attacks; and explore domain-specific counter measures, bounds, and safety defaults (e.g., self-driving cars with a human-driven back up braking system or an AI-controlled temperature system with upper and lower bounds). As AI technologies become ubiquitous, humans and machines will work together seamlessly to improve the efficiency and accuracy of critical tasks (e.g., helping doctors diagnose illnesses or teachers adapting to individual students' needs). The challenge is that the machine or the human's functionality can be heightened or degraded by many factors. Further research is needed to help both machine and human to sense, monitor,

and assess each other's performance and trustworthiness. What if a human cannot respond fast enough in a critical, time-sensitive, human-in-the-loop application? What if the machine and human's results disagree? Theory, techniques, and metrics are needed to support complex decisions, in real time, where the information is ambiguous or subjective, and when a late response could have grave consequences.

6. AI for Cyber security

Just as AI-systems need innovative cyber security tools and methods to improve their trustworthiness and resiliency; cyber security can use AI to increase awareness, reacting real-time, and improve its overall effectiveness. This includes self-adaptation and adjustment in the face of ongoing attacks that alter the current attacker-versus-defender asymmetries. Strategies that identify an adversary's weaknesses, use observation methods, and gather lessons learned, can use AI to categorize various kinds of attacks and inform adaptive responses (e.g., find inconsistencies quickly and know how to repair them) at scale. It is understood that a small team of expert cyber defenders can effectively protect networks used by thousands. The use of AI could extend that same level of system protection, make it ubiquitous, and also provide the domain knowledge necessary to address aspects such as quality-of-service constraints and degradation-of-system behaviours.

6.1 Enhancing the Trustworthiness of Systems

AI technologies can capture and process the enormous amount of data produced by today's technology systems. In turn, this ability provides the training data needed to drive AI-system innovation and development. AI-based reasoning, aligned with cyber security priorities, could make both fully automated and human-in-the-loop systems more trustworthy. Two potential areas are the creation and deployment of more reliable software systems and identity management. Promising research involves leveraging AI to detect errors in programs, check best practices, identify security vulnerabilities, and make it easier for software engineers to design security into their systems.

In modern development practices, code often evolves quickly. The use of AI-based “coding partners” to assist less-experienced developers and analysts in understanding large, complex software systems, and advice them on the security and robustness of proposed code changes, would be valuable. AI can also assist in securely deploying and operating software systems. Once code is developed, AI can be used to detect low-level attack vectors, inspect for domain and application configuration or logic errors, provide best practices for secure system operation, and monitor networks. Open-source software development offers a unique and high-impact opportunity for AI-based security improvements due to its widespread use by commercial and government organizations. However, due to its public nature, open source is vulnerable to malicious actions by an AI-based adversary. Another promising area of AI use is identity management and access control. Adversaries can compromise many techniques simply by stealing authorization tokens. An AI-based system could use a method based on a history of interactions and expected behaviour that is also lightweight, transparent, and difficult to circumvent. For biometric authentication systems, AI could enhance accuracy and reduce threats. However, AI monitoring of behavioural patterns could lead to privacy violations. Further research is needed to develop methods that consider both the ethical and technical aspects, and the potential for abuse of AI-assisted identity management.

6.2 Autonomous and Semiautonomous Cyber security

Unlike other successful AI applications (e.g., spam filtering), AI is likely to be used by both attackers and defenders in cyber defensive scenarios. The traditional strategy based on eliminating vulnerabilities or increasing the cost of an attack changes with the addition of AI. Both autonomous (independent of human action) and semiautonomous (human-in-the-loop) systems will need to plan for worst cases and anticipate, respond, and analyze potential and actual threat occurrences. There are multiple stakeholders affected by AI-based decisions, including data owners, service providers, and system operators. How stakeholders are consulted and informed about autonomous operations and how decision making is delegated and constrained are important considerations. Cyber defenders will likely face autonomous attacks at several levels: in a stable cyber environment, attacks could use classic deterministic

planning; where the environment is uncertain, attacks may involve planning under uncertainty; when little is known about the environment, the attacker could use AI to obtain information, learn how to attack, execute reconnaissance, and develop strategies that include a model of the victim network or system (i.e., AI-enabled program synthesis) and the cyber security product. Methods and techniques are needed to make deployed systems resistant to autonomous analysis and attack. Promising techniques include automated isolation (e.g., behavioural restrictions), defensive agility (i.e., using simulations and updates to strengthen defences), and mission-specific strategies (e.g., use of domain experts to categorize attacks and responses). Mission-driven AI systems must always incorporate the organization leader's intent into any security-related decisions (e.g., access to and operation of the system). A key research question is how to express the leader's intent. AI techniques can translate a mission briefing or operations order into something that is addressable by an autonomous decision system (e.g., dormant attackers may be left alone because rooting them out may be even more disruptive than a possible attack).

AI can also support the mission planning and execution involved in security engineering. AI can be used to identify the cyber assets that are vital for mission success, and to realize that these can change as the mission purpose or goals change. It can help identify and prioritize relevant aspects of the data, computation, information classification, and other security factors including the ongoing adaptation of the AI itself. One challenge is to orchestrate security measures designed for distinct computing resources so that their decisions do not conflict.

6.3 Autonomous Cyber Defence

As adversaries use AI to identify vulnerable systems, amplify points of attack, coordinate resources, and stage attacks at scale, defenders need to respond accordingly. Current practice is often focused on the detection of individual exploits, but sophisticated attacks can involve multiple stages before the ultimate target is compromised. Progress requires a top-down strategic view that reveals the attacker's goals and current status, and helps coordinate, focus, and manages available defensive resources. Consider the scenario of an attack on a power distribution system. A phishing email is opened on a normal workstation; a malware package

is downloaded; credentials of a system administrator who logs in to repair the workstation are acquired; the attacker moves to the power grid's operator console; the entire distribution network is disabled. Any of the individual events can be detected, but the ability to intervene before the network is shut down requires a top-down strategic approach. That strategy would include identification of adversarial goals and strategies, intelligent adaptive sensor deployment, proactive defence and online risk analysis, AI orchestration, and trustworthy AI-based defences. AI planning techniques can generate attack plans and a network of goals, sub goals, and actions that disclose an attacker's strategy. Each attack will have a plan recognizer that receives sensor data, predicts events, and posits defensive responses. AI is trained on search heuristics to derive a single optimal plan; however, a complete set of attack plans is required. Managing plan generation is a major challenge that warrants several possible approaches: use Monte Carlo techniques to generate a representative subset of attack plans; interleave plan generation and plan recognition; and effectively represent the attacker's strategies and tactics. Other considerations include the efficient storage and maintenance of hypotheses and heuristics, and the integration of intelligent and adaptive sensors/detectors to help establish the top-down plan-recognition process. Using a top-down strategic approach to the power distribution scenario means that a plan is generated when the attack is still in its early stages and allows the defender to take actions to prevent the shutdown. These defensive actions might be costly (e.g., shutting down certain machines that provide useful services) or inconvenient (e.g., raising the level of protection in a firewall) and thus require a cost benefit assessment. Reasoning needs to be automated (with possible human-in-the-loop supervisors) because events are extremely time sensitive. As ML and AI systems improve the performance of individual cyber security tools, coordination and orchestration between multiple tools becomes increasingly important. Successful execution may require that models include interactions with other systems. These systems may involve different goals and objectives, cyber security tools, and intent and state of mind of human actors.

6.4 Predictive Analytics for Security

Cyber security will benefit from predictive analytics that process information (both internal and external) to assess the likelihood of a successful attack. Initial work has developed techniques for identifying adversarial operations early in the attack's lifecycle by using data

streams (such as dark web traffic) or distributed logs of cyber-relevant activity. Work has also begun to identify patterns and linkages among datasets that tie together the cyber and human domains, taking advantage of a priori knowledge (e.g., from classified sources) to augment, discover, and track new activities and campaigns. Further research is needed to uncover adversary intent, capability, and motivation of human operators, especially when a system's defences are being tracked. Beyond just detection and the success/failure factor, information about attacks can help protect sources and methods and provide new insights to improve resilience over time. Focus areas include data sources, operational security, and successful adaptation. Obtaining the clean, labelled, real data required for predictive analytics is challenging. Some options include lowering the "labelled" threshold to leverage smaller datasets; capturing and using poisoning resilient data; identifying new cyber-attack early-warning signals using unconventional data streams; and making synthetic training data more realistic. When diverse datasets and AI analytics are used to monitor, track, and counter cyber attacks, false flags can lead to misattribution or even collateral damage. Therefore, AI analysis for cyber attacks may require a higher standard of validation than other intelligence problems. Research is needed to perform multimodal analysis; cross-validation; and identify risks, potential flaws, or gaps in the data sets or the reasoning. AI analysis can also provide new insights that help reduce operator error in both human-in-the-loop and human-on-the -loop contexts, provide more confidence in the outcomes, and help large systems adapt over time. Such analysis might consider the internal state of the system, how regularly patches are applied, what security controls exist (including the human operators), and the level of situational awareness. The analysis would provide scenarios that characterize and prioritize the adversaries' goals, threat level, and likelihood of success and include the prediction's rationale and identify the exploitable weaknesses.

6.5 Applications of Game Theory

There has been significant research into game-theory models that can be used to understand attack plans and reason about potential defences. But because an adversary's actions are still not easily observable, and information is not perfect, more research is needed. In cyber security settings, the "game" can change quickly due to adversarial actions (e.g., a new attack tool or capability), a shifting game environment, players with different incentives, or

irrational players. Also, equilibrium concepts may not make sense, and optimality concepts will need to be derived to apply non-cooperative game theory to cyber security. Non-cooperative game-theory models are appropriate for modelling many different cyber security scenarios; however, there may be instances where different players (e.g., coalition partners) need to cooperate to achieve their goals against an adversary. In some networks it may make sense to treat collections of assets as coalitions, or to consider cooperative orchestration of multiple AI systems (e.g., among different Internet service providers) and teams of AI experts. Additional research is needed on uncertainty planning in a mixture of cooperative and non-cooperative environments. This should also address, in the context of human-machine teaming, how multimodal information is incorporated for more effective decision support. Conversely, game-theory models must assume certain attacker capabilities and incentives. By analyzing data related to attacker tools, AI could provide adversarial modelling including capabilities and incentives. Probabilistic modelling using AI tools may help assess the security of a system (i.e., the extent to which defences will protect the system against a specific set of threats). Game-theory models can be dual use. It is possible that a model can be used for cyber offense and cyber defence. More research is needed to model offense and defence scenarios where there is significant uncertainty, equilibrium is not optimal, attacker action visibility is poor, and the game's action space and assumptions are constantly evolving.

6.6 Human-AI Interfaces

As threats grow more complex and severe, not only is coordination between AI-cyber security systems important, but coordination and trust between human-AI interfaces becomes critical. From enterprise IT to self-driving cars, problems arise when individual system components maximize their own goals without consideration of system-level objectives. Attackers can induce a module to behave in a manner that is locally optimal but globally pathological. Moreover, in an era where information can be misinformed, misattributed, or manipulated, good decision making requires hybrid approaches that leverage and orchestrate the unique human and AI capabilities and perspectives. Human-machine teaming, building trust between systems and humans, and providing decision-making assistance are three important research areas to consider. Human-machine teaming needs to be designed so

humans can understand, trust, and explain the outcomes. Users must be trained to supply goals, feedback, and well-formatted and relevant data, and to know where they fit in the decision-making process. Research is needed on how to incorporate humans to maximize outcomes and minimize latency and negative consequences. AI is often used to automatically shut down suspicious activity to allow time for human decision making. Will this still work as AI is applied to critical systems such as the electrical utility grid, where even a short shutdown could be extremely widespread, disruptive, or dangerous? One solution would be to slow AI systems to accommodate humans in the loop. This would reduce agility, but it could also allow humans to intervene and replace failing components. In a diverse human-AI system environment, interactions must be managed with a goal to reduce human error, increase safety, and provide accountability.

Stakeholders who adopt and use an AI system must understand and trust its operation. The right level of trust requires that humans can identify a system's state and predict its behaviour under various circumstances. Over trust could lead to reluctance to overrule a misbehaving system; under trust could lead to the abandonment of an otherwise effective system. Determining the right level of trust requires human-readable, rule-based specifications based on approximating system behaviour, and consideration of cognitive and other biases. Research literature cites AI systems that can generate extremely convincing fake video and audio that humans will trust. Research must include decision-making assistance such as training human operators to withstand data falsification attacks, and AI-models that can predict failure modes and adapt when humans make erroneous decisions.

7. Science and Engineering Community Needs

7.1 Research Test beds, Datasets, and Tools

To establish the AI community standards and metrics required to safely deploy future AI systems, more investment is needed in research testbeds and datasets. Threat detection mechanisms must be tested and evaluated for critical AI application domains (e.g., autonomous vehicles, medical diagnosis) to incentivize adoption. Possibilities include the creation and maintenance of realistic simulation environments and diverse domain-specific

datasets. The complexity of both the AI system and the AI-threat landscape require testbeds and datasets that evaluate capabilities and defences in a comprehensive, principled, and sustainable manner. They should be modular (to facilitate use across different disciplines) and open source; foster innovation, collaboration, and reproducibility; and continually re-evaluate cross-layer interaction.

7.2 Education, Job Training, and Public Outreach

Education and outreach efforts should focus on fostering the necessary workforce and developing an informed public that understands the usefulness, limitations, best practices, and potential dangers of AI technology. AI should be integrated into primary, secondary, and university education that brings together the disciplines of computer science, data science, engineering, and statistics. The teaching of AI should be considered as part of the accreditation process.