الوكالة الوطنية للأمن السيبراني
National Cyber Security Agency

Guidelines for Secure Adoption and Usage of Artificial Intelligence

2024

Version 1.0

الوكالة الوطنية للأمن السيبراني
National Cyber Security Agency

DISCLAIMER / LEGAL RIGHTS

National Cyber Security Agency (NCSA) has designed and created this publication, titled "Guidelines for Secure Adoption and Usage of Artificial Intelligence" - V 1. 0 -, in order to provide guidance to organizations on how they can securely adopt Artificial Intelligence.

NCSA is responsible for the review and maintenance of this document.

Any reproduction of the present document either in part or full and irrespective of the means of reproduction; shall acknowledge NCSA as the source and owner of the "Guidelines for Secure Adoption and Usage of Artificial Intelligence".

Any reproduction concerning this document with intent of commercialization shall seek a written authorization from the NCSA. NCSA shall reserve the right to assess the functionality and applicability of all such reproductions developed for commercial intent.

The authorization from NCSA shall not be construed as an endorsement of the developed reproduction and the developer shall in no way publicize or misinterpret this in any form of media or personal / social discussions.

Organizations remain responsible for ensuring the availability, reliability, quality and safety of their products and services, regardless of whether AI technologies are used.

Adopting this voluntary Guideline will not absolve organizations from compliance with current laws and regulations. It should be noted that certain industry sectors (such as finance, healthcare, and legal) may be regulated by existing specific laws, regulations or guidelines relevant to the sector.

Document Control

Document Details	
Document ID	[IAG-NAT-SUAI]
Version	1. 0
Classification & Type	Public
Abstract	To provide guidance to organizations on how they can securely adopt Artificial Intelligence

Review/Approval

Department/Role	Reviewed/ Approved	Version	Date
National Cyber Governance and Assurance Affairs		1. 0	February 2024

Revision History

Version	Author(s)	Revision description	Date
1. 0	CSSP	Published	February 2024

LEGAL MANDATE(S)

Amiri decree No. (1) of the year 2021 regarding the establishment of National Cyber Security Agency, sets the mandate for the National Cyber Security Agency (hereinafter referred to as "NCSA"). The NCSA has the authority to supervise, regulate and protect the security of the National Critical Infrastructure via proposing and issuing policies and standards and ensuring compliance.

This Document Type has been prepared to take into consideration the current applicable laws of the State of Qatar. If a conflict arises between this document and the laws of Qatar, the latter shall take precedence. Any such term shall, to that extent be omitted from this Document, and the rest of the document shall stand without affecting the remaining provisions. Amendments, in that case, shall then be required to ensure compliance with the relevant applicable laws of the State of Qatar.

Content

1 Introduction

The recent times have witnessed progression of Artificial Intelligence (AI) technology into mainstream business and general public usage due to its potential usage, manifold applications, and limitless opportunities. Businesses have realized the potential power and impact that AI could provide in scaling up businesses, improving user experience and driving up customer satisfaction.

However, AI as a concept has been around for decades. The underlying models that define and build AI have been researched since the early 50s. But factors such as:

- The evolution of advanced techniques in machine learning, neural networks and deep learning,
- The availability of significant data sets to enable robust training,
- Advances in high performance computing enabling rapid training and development

Have led to increased interest and investment in AI, leading to development of practical AI applications that is leaving an impact on the general user.

As a technology, we have been inadvertently using different forms of AI in our day to day life and consists of but not limited to applications related to smart vision, voice recognition, video generation from images, tele-health applications, etc. However, many of these applications required a level of technicality and skills. That is no longer the case with the advent of Generative AI application such as ChatGPT which allows the common man to seek help from the tool in a natural conversational style. It broke the technological barrier and democratized power at the fingertips of a common person. .

Businesses and nations have long realized the potential of AI technology and are diligently working to harness the power of AI. For businesses, AI provides the possibilities of scaling quick, unlocking business insights from data that previously just occupied stacks of drives, and more importantly being able to take better decisions, offer better services with better quality assurance.

For nations, AI has the potential to be useful in achieving national aspirations in areas which include but are not limited to the following:

- Offer better governance to its citizens and residents,
- Enhance the efficient utilization of funds and budgets,
- Smarter policy making,
- Manage pandemics and disease prevention,
- Improve the critical infrastructure protection,,
- Enhance the defense and security mechanisms.

The State of Qatar embarked its AI journey by launching the National Artificial Intelligence Strategy in 2019. The strategy had a clear focus on several key areas which include raising awareness, increasing capacity and capability building, promoting research and development (R&D) as well as innovation, establishing governance for emerging technologies, and ensuring a common national approach aligned with the goals outlined in Qatar Vision 2030.

This was followed by the establishment of the Artificial Intelligence Committee within the Ministry of Communications and Information Technology (MCIT), in accordance with the provisions outlined in Cabinet Decision No. 10 of 2021.

These initiatives underscore Qatar's commitment to strategically integrate AI into its governance framework, reflecting a proactive approach to harnessing the transformative potential of AI technologies across sectors; and build on the long term goal of establishing Qatar as a "Knowledge Based Economy".

However, as with any other technology, AI systems come with their own set of risks. The sheer power and access to huge data sets can open it to attacks on a wider attack surface. Furthermore, processing personal data by utilizing AI may result in significant compliance risks resulting in breaching national privacy regulations. Vice versa, the power can also be used to attack others and as such the United Nations has classified AI as a dual use technology. Dual use refers to technologies that have a potential to offer greater benefit to humans if used well, but also pose huge risks to humans if not regulated and used for the wellbeing of humans. For example, nuclear science and nuclear energy can offer alternatives to clean energy, thereby, support in combating climate changes, while at the same time it can also be utilized for malevolent purposes.

This document aims to guide stakeholders in safely adopting AI technology by detailing best practices, outlining potential risks, and providing mitigation strategies to ensure a secure AI-driven ecosystem.

2 Purpose, Scope, and Usage

2.1 Purpose

The information security considerations and recommendations set out in this Guideline are intended to guide organizations that have decided to deploy AI technologies at scale.

This Guideline focuses primarily on the following broad areas:

Area 1: Build stakeholder confidence in AI through organizations' responsible use of AI to manage the risks related to information security and fair usage in an AI deployment.

Area 2: Provide guidance on the key issues to be considered and measures that can be implemented for responsible usage.

Area 3: Provide specific guidance on GenAI, highlighting its threats and possible mitigation solutions.

2.2 Scope and Considerations

This document is a national guideline, the scope covers all private and public organization in the state of Qatar using or intending to deploy AI systems, services or products. The primary focus is set on business users who are integrating or adopting AI solutions within the existing business and IT systems.

Primarily, cybersecurity and AI have three major dimensions:

1. **Cybersecurity of AI:** This involves assessing and managing the information security risks of an AI system. The AI system includes the application (front end, back end, databases, etc.), underlying infrastructure (hardware and network), and the underlying AI models and algorithms.

    a. **narrow scope:** protection against attacks on the confidentiality, integrity and availability of assets across the life cycle of an AI system.

    b. **extended scope:** complementing the narrow scope with trustworthiness features: data quality, oversight, robustness, accuracy, explainability, transparency, traceability and data privacy.

2. **AI to support cybersecurity:** AI used as a tool to scale up efforts and create advanced cybersecurity tools which may facilitate areas such as, but not limited to, advanced threat detection, behavioral analysis, predictive analysis, and speed of response. .
3. **Malicious use of AI:** adversarial use of AI to aid or create sophisticated cyber-attacks by malicious threat actors. Examples include, but not limited deep fake videos, automated social media manipulation, AI powered cyber attacks, etc.

Within the context of this document, we will focus on the **cybersecurity of AI systems.**

While the Guideline is certainly not limited in ambition, it is ultimately limited by form, purpose and practical considerations of scope. It is important to note that the guidelines were developed in accordance with the following considerations.

1. **Algorithm-agnostic**: the guidelines will not focus on specific AI or data analytics methodology, but it applies to the design, application and use of AI in general. However, due to the recent developments and increased interest, the guideline contains a set of special recommendations regarding GenAI.
2. **Technology-agnostic**: the guidelines will not focus on specific systems, software or any technology, and will apply regardless of development language and data storage methods.
3. **Sector-agnostic**: the guidelines will serve as a baseline set of considerations and measures for organizations operating in any sector to adopt. Specific sectors or organizations may choose to include additional considerations and measures or adapt this baseline set to meet their needs.
4. **Scale- and Business-model-agnostic**: the guidelines will not focus on organizations of a particular scale or size. It can also be used by organizations engaging in business-to-business or business-to-consumer activities and operations, or in any other business model.

This guideline does not replace existing cyber and information security policies, standards, guidelines, and best practices, but complements them, highlighting the areas where the existing structures should be amended to fit the new security threats and risks AI brings.

2.3 Usage

AI systems are a collation of different sub-systems such as databases, machine learning, powerful processing units, bound together by the underlying AI models and algorithms. The basic cyber security hygiene does not change and this document complements the existing national policies, standards, and guidelines.

The guideline will help AI deployers to:

- Broaden and extend the scope of their perspective on security of AI.
- Expand the scope of existing security processes to encompass the principles of AI espoused in this document to provide assurance on trust.
- Understand the importance of building trust with stakeholders while using or deploying AI systems in their business.
- Review and strengthen cybersecurity measures, risk assessment and mitigation processes.
- Understand the AI product lifecycle and the value this new technology can bring to the organization.

Organizations planning to implement AI systems, solutions or product should refer to these guidelines both before and throughout the deployment process. These guidelines offer practical recommendations to promote ethical practices and ensure seamless and secure integration of AI technologies. .

3 Key Definitions

AI deployer	Refers to companies or other entities that adopt, integrate or deploy AI solutions in their operations, such as backroom operations (e. g. , processing applications for loans), front-of-house services (e. g. e-commerce portal or ride-hailing app), or the sale or distribution of devices that provide AI-powered features (e. g. smart home appliances).
AI solutions provider	Refers to entities who develop AI solutions or application systems that make use of AI technology. These include not just commercial off-the-shelf products, online services, mobile applications, and other software that consumers can use directly, but also B2B2C applications, e. g. , AI-powered fraud detection software sold to financial institutions. They also include device and equipment manufacturers that integrate AI-powered features into their products, and those whose solutions are not standalone products but are meant to be integrated into a final product. Some organizations develop their own AI solutions and can be their own solution providers.
Artificial Intelligence (AI)	Refers to a system (hardware, software, or both) that is designed to carry out any tasks associated with human intelligence, in a manner that mimics the human mind with a certain level of autonomy.
Automated Decision Making (ADM)	Refers to the application of automated systems in any part of the decision-making process.
Data controller	The PDPPL, defines Controller as a natural or legal person who, whether acting individually or jointly with others, determines how personal data may be processed and determines the purposes of any such processing. (Article 1 of the Law No 13 for the year 2016 Personal Data Privacy Protection)
Data processor	The PDPPL, defines Processor as a natural or legal person who processes personal data for the controller. (Article 1 of the Law No 13 for the year 2016 Personal Data Privacy Protection)
Deep Learning	Is a subset of machine learning that uses neural networks, which mimics how neurons interact in the human brain, allowing it to learn complex patterns and relationships within data (e. g. , pictures, text, sounds, and other data), and requires even less human intervention, and can often produce more accurate results than traditional machine learning.

Generative AI (GenAI)	Refers to an intelligent system capable of creating new content such as audio, video, image, texts, code, etc, based on the patterns and structures of the data they have been trained on.
Graphics Processing Unit (GPU)	A specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. GPUs are used in embedded systems, mobile phones, personal computers, workstations and game consoles.
Large language model (LLM)	refers to a type of GenAI that specializes in the generation of human-like text.
Machine Learning (ML)	Is a part of AI that builds the intelligence of computer systems by improving their perception, knowledge, thinking, or actions based on algorithms that are trained on data
Multimodal Foundation Model (MfM)	Refers to a type of GenAI that can process and output multiple data types (e. g. , text, images, audio).
Personal data with special nature	The PDPPL, defines Personal data with special nature as personal data related to ethnic origin, children, health, physical or psychological condition, religious creed, marital relations, and criminal offenses. The definition should be understood broadly. (Article 16 of the Law No 13 for the year 2016 Personal Data Privacy Protection)
Sensitive data	Data that has significant value to the organization, loss, compromise, or unavailability of which may lead to financial, reputational or operational impacts for the organization
Tensor Processing Unit (TPU)	AI an application-specific integrated circuit, to accelerate the AI calculations and algorithm. Google develops it specifically for neural network machine learning for the TensorFlow software.
Use of AI	Developing or deploying (applying) an AI system, product, or service.

4 Risks, Threats, Challenges

4.1 Risks and Threats

AI systems, like any other information systems are at risk of unauthorized disclosure of information, unauthorized modification of information or loss of integrity of system (platform), non-availability of information, and regulatory violations. Besides, these AI systems also carry a huge risk of data bias, transparency and explainability limitations. [1] The following figure represents the possible categorization of major threats and risks.

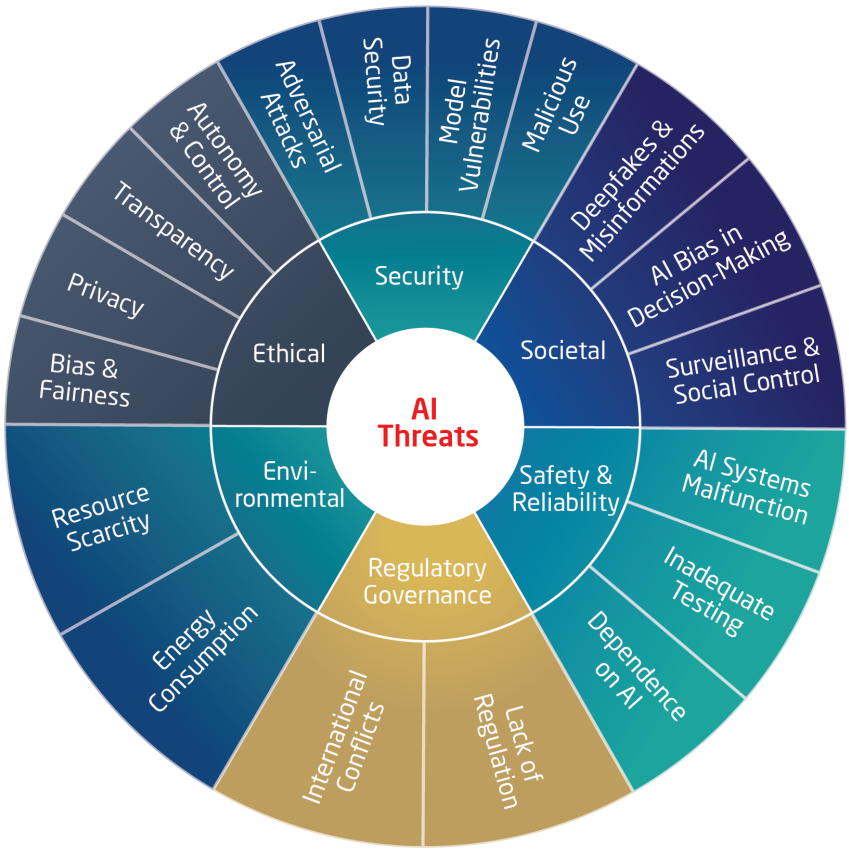

Figure 1 AI Risks, threats and challenges Non-Exhaustive

4.1.1 Unauthorized disclosure of information

Data forms the essence of any AI platform. The system thrives on "learning" from huge datasets and as such AI platforms store and process vast amounts of data, including confidential or sensitive data such as personal data, financial data and health records among other things. The sheer size of data makes it an interesting target for malicious actors. As such, data breaches pose a significant cybersecurity risk on these systems.

Several factors such as lack of cyber hygiene like weak security protocols, application security bad practices, lack of input sanitization, insufficient encryption in underlying systems, , lack of monitoring and internal threats can make the AI systems vulnerable to this risk and materialize itself in the form of ***data***

1 Onto better understand terms like transparency and explainability please refer to the Ethical and Fair AI principles in the Appendix 8. 1 section of this document.

breach / data loss / data theft. In case of a data breach, the data controller must also take into consideration the obligations derived from relevant national regulation. [2]

Data aggregation is another threat that risks potential exposure of sensitive information to malicious or non-malicious actors. AI systems thrive on data, and during the course of its operations, it might collate multiple data sets, whose information in itself may seem harmless (less sensitive and as such the confidentiality ratings may be low, for example, personal data that is not considered personal data with special nature, like national address), yet when combined with other data points, it may have the potential to create detailed databases or insights that might be potentially sensitive (higher confidentiality rating, for example if national address combined with information about children, or ethnic origin it becomes personal data with special nature) in nature. [3] Even when data is pseudonymized, AI systems might be able to use advanced pattern recognition or combine datasets to re-identify individuals without permission.

Also, AI solutions generally store data for extended periods (***data retention***) so that the model can continue referencing, analyzing and comparing it as part of its learning, predictive and other capabilities. This long-term data storage increases the risk of unauthorized disclosure. Moreover, it is a potential breach of the PDPPL obligations such as data retention, purpose limitation and data minimization.

Similarly, there is a risk of unauthorized disclosure through attacks on the AI system's machine learning model. This is referred to as the Membership Inference attack, which allows the attacker to detect data used to train a machine learning model. Usually, membership inference attacks could be staged by an attacker without having access to machine learning model's parameters, by just observing its output. Such attacks can raise security and privacy concerns if the model has been trained on sensitive information.

In generative, AI systems, the attack manifests itself by the attacker manipulating the input prompt to trigger unwanted behavior from the AI model. This manipulation by way of ***prompt injection includes jailbreaking, prompt leaking, and token smuggling*** can lead to the AI generating inappropriate responses or leaking sensitive information (resulting in breach of PDPPL regulation). These attacks can be especially potent when AI systems are used in conjunction with other systems or in a software application chain.

These are just some examples of the various threats that could be potentially used to realize the risk of loss of confidentiality.

4.1.2 Unauthorized modification of information or loss of integrity of system (platform)

Given their inherent nature, maintaining data integrity is paramount for AI systems. Any compromise on the data quality, will impact the quality of data output and its efficacy in real world. Such attacks can manifest in many ways.

Malicious data, introduced into the training data, can manipulate AI model behavior, leading to incorrect or biased predictions, inaccurate or unfair decision-making. This is also referred to as ***model poisoning***. Model poisoning attacks can be challenging to detect, because the poisoned data can be innocuous to the human eye. Detection is also complicated for AI solutions that include open-source or other external components.

Attackers can attempt to reverse engineer an AI model to replicate a proprietary trained machine learn-

2 Such as the Law No. 13 of 2016 on Personal Data Privacy Protection (PDPPL) and the Law No. 14 of 2014 Promulgating the Cybercrime Prevention Law.

3 NCA National Data Classification Policy

ing model based on its queries and responses. The attacker will use a series of custom designed queries to the model and use the responses to construct a copy of the target AI system. This is known as a ***model extraction*** attack and it could infringe on intellectual property rights and can lead to significant economic losses.

The quality of data may also be impacted due to ***data supply chain,*** which refers to the quality of data (acquired or processed) impacted due to threats related to input systems, and processes related to data collection and processing by third party vendors.

4.1.3 Non-availability of information

In a data centric system, availability is a key requirement. AI infrastructure as any other IT system is at risk of non-availability due to a variety of factors. Hardware failure of a component within the AI system, poor design of the system, resource crunch, lack of resiliency, poor operational processes could be some of the factors.

Denial of service (DoS) attacks could be one such threat. [4] By overwhelming the AI model's infrastructure with traffic, an attacker could render an AI service unusable.

4.1.4 Privacy Violations

Privacy is a fundamental right, furthermore the PDPPL sets out obligations for the entities that process personal data to protect the privacy of people.

AI systems can collect and process large amounts of data, which can raise privacy concerns, challenges regarding mandatory compliance with the PDPPL and other applicable regulation. This large-scale collection, processing, and analysis of personal data by AI systems can lead to:

Surveillance and profiling: the ability to track individuals, monitor individual behavior. AI technologies like facial recognition and social media monitoring can enable invasive surveillance and profiling of individuals that endangers rights to privacy, anonymity and autonomy resulting in targeting certain individuals or groups, violating their right to privacy, freedom of expression, and association. [5]

Regulatory non-compliance: the context and complexity of AI solutions can make it challenging to ensure that processing data is compliant with PDPPL and other privacy norms. For example: personal data is deleted when it is no longer needed or when individuals exercise their rights to request deletion; individuals right to access personal data; individuals right to receive information on personal data processing techniques.

Furthermore, there is also a possibility of unravelling personal data, from large sets of anonymized data, due to data aggregation.

Data sharing: AI platforms can involve collaboration between multiple parties or use an AI solutions provider's tools and services. On many occasions, data processors and controllers end up sharing data through multiple AI systems, solutions. This can cause data transfers across many jurisdictions, and to other entities resulting in privacy compliance challenges and regulatory violations.

4 NCSA has issued relevant guideline in 2023 January titled: "Cybersecurity Guidelines Distributed Denial of Service (DDoS) Attacks".

5 Such kind of surveillance and profiling can be a direct violation of rights, because prior the processing of personal data the individual must be informed according to PDPPL statutes.

4.1.5 Data Ethics and Bias

AI systems and services thrive on data. The more data, the better the model can be trained on these data sets and optimized to perform better. The quality of data is of essence here and a poor quality of data has the potential to throw the model off its mark completely. If the data used to train AI models is incomplete, it can lead to producing inaccurate or unexpected results, thereby introducing a **bias.**

These biases could have a significant impact, especially in the government sector, when used in systems meant for delivering public services. Besides the fact that the predictions of the system could have a negative impact on policy making, it could also lead to unrest/discontent among the public who might be affected by such decisions.

Primarily, there are two types of biases**, cognitive bias and data bias**.

Cognitive bias relates to unconscious errors in thinking that affect an individual's judgement and decision. These include discrimination or prejudice against a particular race, group, gender, or demographics usually unconscious to the person with bias. Sometimes this also seeps in through historical data that is used to train the AI system.

Data bias relates to errors introduced primarily due to quality of data. These include incompleteness of the data set, or a skewed data set not representative of all stakeholders, or a limited data set.

Examples of these biases that can negatively affect society include recognition disparities between ethnic backgrounds by facial recognition tools used by law enforcement, employment tools that associate women's names with traditionally female roles, the spread of politically motivated disinformation and perpetuation of biased worldviews, and racial bias in financial schemes among other things.

4.2 Challenges

4.2.1 The Need for Capacity and Capability Building Measures

To meet global, national and organizational AI aspirations, we increasingly need the right mix of talent to translate business needs into solution requirements, build, and deploy AI systems, integrate AI into processes, and interpret results. Management needs to understand how AI works to effectively integrate, deploy, use, and maintain AI systems.

A multi prong plan needs to be implemented to effectively deal with this challenge. In the short and mid-term, AI adopters need to train current workforces to strengthen expertise and narrow their skills gap. Many companies around the world are training developers to create AI solutions, IT staff to deploy those solutions, and employees to use AI in their day-to-day jobs. However, in the long term we need to revamp our education system to ensure that provide young children with the right temperament, skills, and career path to pursue a career in or based on AI. [6]

4.2.2 Dynamic Regulatory Landscape

Globally, the legal and regulatory environment for AI is in a perpetual state of flux. Governments, civil societies, individuals, and industries are grappling with the potential impact that AI could have on humans and human societies, and is straddled with questions like *"Will AI takeover Human Jobs?"* to *"Will AI enslave Humans some day?"*

Governments worldwide currently confront the challenge of striking a balance between utilizing AI's

6 National Artificial Intelligence Strategy for Qatar, Ministry of Communications and Information Technology 2019.

potential benefits to humankind, and controlling its potential misuse, to avoid the technology devolving into autonomous and potentially uncontrollable systems.

International institutions such as the United Nations (UN), European Union, industry associations, non-profit organizations and respective countries are evaluating the AI technology and proposing various regulations, standards and guidelines to regulate and control this technology. Since these requirements related to AI are under development globally, and technology is also evolving quickly, frequent changes and updates are expected. Organizations need to follow up on legal developments and be ready to dynamically change their business policies and processes to assure full compliance.

4.2.3 Auditability

In the traditional scheme of things, compliance and audit programs have been the backbone of maintaining a check and balance and ensuring that things are how they are supposed to be. .

However, with AI systems, the challenge stems from the concerns about explainability of the systems since inherently the systems are complex and difficult to understand. More specifically, the underlying models are often complex and involve numerous layers and parameters. These models are trained on massive amounts of data, making it challenging for humans to comprehend the exact reasoning behind their decisions. This is known as the "black box" effect.

Traditional tools available for audit and compliance may not be geared to assessing and auditing this problem.

4.2.4 Dual Use Nature of AI: AI created and AI powered malware

The UN has designated AI systems as a dual use technology on lines similar to Nuclear Technology. Even on a lower scale, commercial AI technology including GenAI technology has the capability to be misused for malicious purposes. Malicious actors can use clever techniques to bypass the security measures and controls in place to create malwares as sophisticated as if it was created by a state-sponsored threat actor. It could redefine the capabilities of a traditional script kiddie (novice) threat actor. .

AI-powered malware is trained through machine learning to be faster and more effective than traditional malware. Unlike malware that targets many people with the intention of successfully attacking a small percentage of them AI-powered malware is trained to think for itself, update its actions based on the scenario and specifically target its victims and their systems.

5 Guidelines

The goal of this guideline is to drive organizations in their journey to adopt and use AI in a safe and secure manner. So, to put that into perspective AI deployers should have strategic requirements for AI adoption that further drive decision points like why they need AI, how they go about getting the right AI solution, what they need to secure the AI solution, what value they get out of AI solution and related validation and verification that the AI solution is acceptably secure.

The following sections recommend a list of security best practices for a secure and trusted AI system that organizations can adopt and implement within existing or while building new AI systems. The guidance is designed around the established triad of people, process, technology, and follows the AI lifecycle. Organizations may use their discretion in choosing practices and controls that best suit their environment.

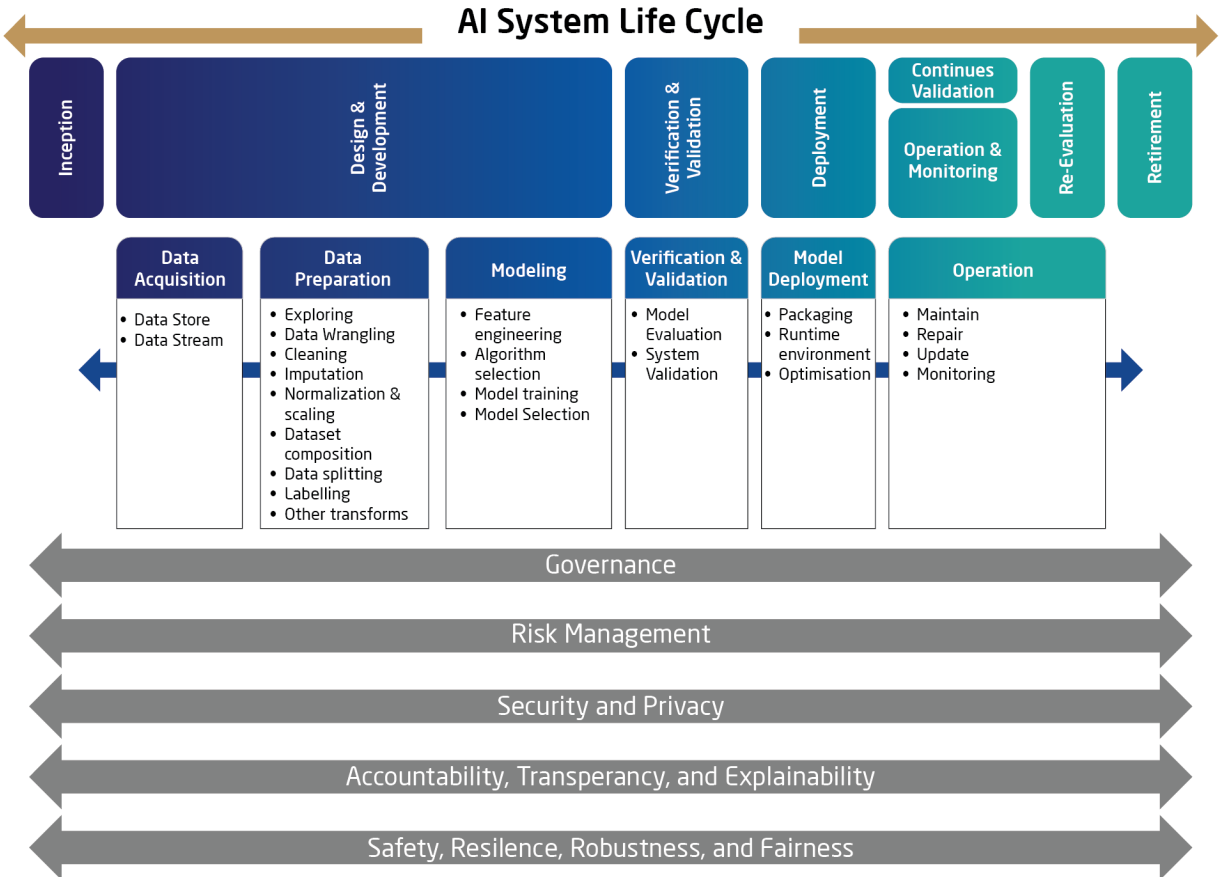

Figure 2 AI lifecycle tasks in the view of controls[7]

As mentioned earlier, this guideline does not replace existing cyber and information security policies, standards, guidelines and best practices but complements them, highlighting the areas where the existing structures should be amended to fit the new security threats and risks AI brings.

Every organization intending to use AI must understand that the most complex of information systems are eventually made of a multitude of simpler processing blocks that are put together through additional layers of hardware, and software. Additionally, AI systems can be targeted through cloud-based services that might be in use, or physical components such as the graphic processing units (GPUs) and tensor processing units (TPUs). [8]

Needless to say, basic information security hygiene is fundamental. AI deployers must have all the necessary baseline information and cybersecurity controls, governance structures and policies in place.

Organizations in Qatar should continue to adhere to the cybersecurity regulation published by NCSA such as the National Data Classification Policy and the National Information Assurance Standard along with other security and privacy policies, standards, frameworks and guidelines published from time to time. Organizations can also choose to follow international standards and frameworks such as the ISO 27000 standard group.

7 Adopted from ISO 22898-2023 AI System Lifecycle

8 GPUs and TPUs are specialized processors designed to accelerate AI workloads, and they can introduce new attack vectors. Design flaws in processors and other hardware can affect a range of products.

5.1 People

The use of AI can bring significant risks and additional obligations to the organization. To ensure effective, efficient and acceptable use of AI systems in the organization establishing AI governance is essential.

This section is intended to guide AI deployer organizations in developing appropriate internal governance structures that allow them to have oversight over AI technologies and understand how AI can affect human resources.

5.1.1 Ethical use of AI

The governance of an AI deployer is enabled by the application of its principles therefore organizations should detail a set of ethical and fair usage principles when developing or deploying AI products, services or systems.

When establishing ethical principles for AI, organizations should review their existing corporate values in light of the "Ethical and Fair AI Principles" prescribed in Appendix 8.1 of this document. The Ethical and Fair AI principles espoused in this guideline are aligned to global best practices and the National AI Strategy of Qatar[9].

5.1.2 Internal governance structures

Internal governance structures and measures help to ensure robust oversight over an organization's use of AI. The AI deployer's existing internal governance structures can be adapted to fit the new challenges brought by AI. For example, the risks associated with the use of AI can be managed within the organization's risk management structure while ethical considerations can be introduced as corporate values and managed through ethics review boards or similar structures.

An AI system might affect an organization's culture by shifting and introducing new responsibilities, roles and tasks. Responsibility for and oversight of the various stages and activities involved in AI deployment should be allocated to the appropriate personnel and/or departments.

These individuals include personnel:

- With authority to address AI risks,
- With responsibility for establishing and monitoring processes to address AI risks,
- With authority to decide on the appropriate level of human involvement in AI-augmented decision-making,
- With responsibility for maintenance, monitoring, documentation, and review of the AI models.

5.1.3 Capacity and Capability building within the organization

AI as a domain, is a relatively new area that requires multiple new skills to be developed or enhanced within an organization. These include technical skills such as machine learning, data science, neural networks, legal understanding, ethics among other skills.

Therefore, top management should take steps to improve AI-related skills among the employees. The initiative should look beyond ramping technical skills and impetus should also be given to skills related to ethics, legal, and privacy.

9 In case of a conflict, the principles prescribed by the updated National AI Strategy of Qatar or any future publication by MCIT and/or the national AI governance body shall have precedence over the principles espoused in this guideline.

5.1.4 Acceptable user policies and restrictions

AI deployers should define acceptable user policies ("AUPs") to ensure that users clearly understand how to use the AI system securely. Generally, such AUPs are targeted at end users to help them understand the do's and don'ts.

An AUP may be created using:

1. A list of possible use cases, based on potential risks, likelihood, and severity.
2. A definition of low risks and acceptable use cases.

A list employee obligations and restrictions. Organization's might find it useful to draft a list of *"Dos and Don'ts"* regarding use. For example:

- Don't input any personally identifiable information.
- Don't put any sensitive information.
- Don't input any company IP without checking with the IT security team.
- Do turn off history tracking.
- Do closely monitor outputs, for factual errors and biased or inappropriate statements.
- Don't input malicious data that unacceptably manipulates the performance and/or results of the solution's model.

The AUPs should be defined in a simple, unambiguous language.

5.1.5 Human oversight in AI decision making

Strategic decision-making factors in data from various sources e. g. , situational awareness to make an informed decision. There are many tools available for automated decision-making (ADM) and as such, AI is not a one size fits-all solution, however, the interest in AI-supported decision making is globally increasing.

A first step in determining the extent of human oversight is having a clear objective of using AI. AI deployers can decide on their commercial objectives and these can be weighed against the risk of using AI in decision-making. Moreover, it is also desirable especially for organizations operating in multiple countries to consider how AI systems interact with pre-existing societal norms, values, and expectations. Human oversight should form an integral part of the system design and performance. This may include the performance of mandatory actions and checks and rules for escalation.

Further, we should be mindful that certain regulations such as the European General Data Protection Regulation (GDPR), as also some other nascent AI regulations restrict the extent of autonomous decision making by AI in specific use cases.

We can identify three broad approaches to the various degrees of human oversight in the decision-making process. These are:

1. **Human-in-the-loop:** In this model, the human oversight is active, involved, retaining full control over the AI system. The ADM is only providing recommendations or input but the final decision is taken by a human.
2. **Human-out-of-the-loop:** In this model, there is no human oversight over the AI system and the execution of decisions. The ADM system has full control over the decision-making process and no human has the right to override them.

3. **Human-over-the-loop / human-on-top-of-the-loop:** In this model, the human oversight is a monitoring or supervisory role, with the ability to take over control when the ADM encounters unexpected or undesirable events.

The specific approach chosen is influenced by factors such as regulations and desired risk management.

5.2 Process

5.2.1 Risk management

The general principles of risk management are integrated, structured and comprehensive approach. Risk management should consider the whole system, with all its technologies and functionalities and its impact on the environment and stakeholders.

However, AI systems are complex systems and can introduce new or emerging risks for an organization with positive or negative consequences for objectives or changes in the likelihood of existing risks. As such, we need an adaptive risk management framework for AI. Such a risk management framework should be the integrated part of the AI deployer organization's risk management framework and should possess the following qualities:

Inclusive: seeking dialog with diverse internal and external groups, to communicate harms and benefits, and incorporating feedback and awareness into the risk management process.

Dynamic: The risk management framework is dynamic because:

- The nature of AI systems is itself dynamic, due to continuous learning, refining, evaluating, and validating. Some AI systems have the ability to adapt and optimize, creating dynamic changes on their own.
- Customer expectations around AI systems are high and can potentially change quickly as the systems do.
- Legal and regulatory requirements related to AI are frequently changing.

Sensitive to human and cultural factors: monitor the landscape, how AI systems or components interact with pre-existing societal patterns that can lead to impacts on equitable outcomes, privacy, freedom of expression, fairness, safety, security, employment, environment and human rights broadly.

Support continual improvement: the identification of previously unknown risks related to the use of AI systems should be considered in the continual improvement process. Organizations should monitor the AI ecosystem for performance successes, shortcomings and lessons learned and maintain awareness of new AI research findings and techniques

AI is a whole new paradigm. With its ability to influence and/or make decisions that may impact human lives, it is paramount that such systems hold to high standards of trustworthiness. The Ethical and Fair AI principles espoused in this guideline are aligned to global best practices and the National AI Strategy of Qatar[10]. The following principles of AI[11], define and articulate the trustworthiness of an AI system:

10 In case of a conflict, the principles prescribed by the updated National AI Strategy of Qatar or any future publication by MCIT and/or the national AI governance body shall have precedence over the principles espoused in this guideline.

11 Explained in details in the Appendix 8. 1 Ethical and Fair AI Principles section of this document.

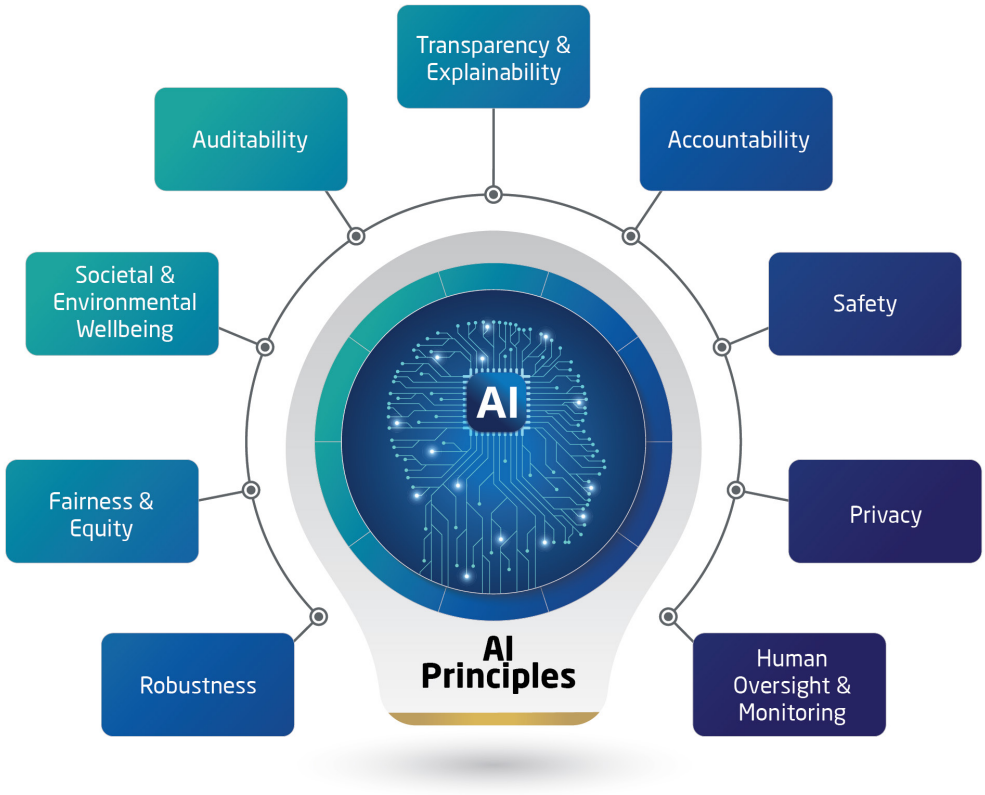

Figure 3 Ethical and fair AI principles

Building a trustworthy AI system, requires balancing each of these principles in the context of the business, the stakeholders and the AI system itself. This necessitates that Risk Management is integrated within the life cycle of an AI system, right from the plan and design phase up until the operate and monitor phase.

A detailed risk management framework is beyond the ambit of this document. However, the intent is to provide the broad contours of a pragmatic risk management framework for AI systems.

Key facts organizations should keep in mind while managing AI risks:

1. AI systems are data intensive. The data sets eventually influence how the model works and/or makes inferences, as such organizations should conduct due diligence in identifying relevant threats and risks related to data, life cycle of data, processes related to data acquisition, process, storage, output, its stakeholders etc. Organizations should also take into consideration all the requirements of the PDPPL and ensure compliance.
2. AI systems could be overly complex, as such organizations need to consider risks related to explainability of the system, regulatory risks especially around privacy, fairness, auditability and human and societal impact.
3. AI models need to be tested thoroughly over large sets of data and continuously revalidated to ensure that it remains true to its business objectives and the underlying principles.
4. Despite all the due diligence, costs, complexities, and risk management, AI systems carry a degree of un-predictiveness.
5. Based on the criticality of the system and the function being served by the AI system, it is imperative that the role and intervention of humans is clearly identified during the plan and design phase itself and effectively implemented and monitored during the life cycle of AI system.

6. Guidance when defining risk criteria:
    a. Organizations should take reasonable steps to understand uncertainty in all parts of the AI system, including the utilized data, software, mathematical models, physical extension, and human-in-the-loop aspects of the system
    b. Awareness that AI is a fast-moving technology domain. Measurement methods should be consistently evaluated
    c. Establish a consistent and dynamic approach to determining the risk level.
7. Guidance when defining risk appetite:
    a. The governing body should assess its intended use of AI as part of its risk appetite.
    b. The organization's AI capacity, knowledge level and ability to mitigate realized AI risks should be considered when deciding its AI risk appetite.
8. AI risks should be identified, quantified or qualitatively described and prioritized against risk criteria and objectives relevant to the organization and integrated into the AI life cycle.
9. AI systems demand that an organization has the ability to respond and mitigate risks in a dynamic and proactive manner since within the realm of AI risks can change rapidly. New insights and a proactive approach provide an organization with the means to respond to risk. The organization should therefore demonstrate willingness to modify or abort projects if deemed necessary.

5.2.2 AI operations management

5.2.2.1 Stakeholder interaction and communication

Management should be educated on the general understanding of AI as a technology its capabilities, potential impact on business and society, and also its implementation challenges. Building AI systems is complex, huge and potentially an expensive exercise. Management should be aligned on development cycles as also on the failure management of AI systems.

Top management should provide general information on whether and how AI is used in the organization's products and/or services. Communication should be open and easy to understand. It can take the form of a general product description or labelling / watermark on digital products.

It is advisable to disclose the way an AI decision may affect an individual (employee or client), and whether the decision is reversible. The management should inform the users if they have an option to opt-out of the AI usage. The decision review and feedback channels should be clearly marked.

5.2.2.2 Record keeping (documentation)

A key component of ensuring explainability and auditability in AI systems is documentation, including but not limited to the design, architecture, model, data sets, test and validation results. Organizations should ensure that these records are not tampered with and secured accordingly.

5.2.2.3 Security Monitoring (Log management)

When organizations are evaluating the integration of an AI system or solution into their business processes they should look for AI systems which are designed and developed with capabilities enabling the automatic recording of events ('logs') while the product is operating. Those logging capabilities should conform to recognized standards or common specifications. To the extent possible, systems should be

capable of logging logs related to its decision making, along with the operational system, application, and security logs.

5.2.2.4 Data Observability

Further to monitoring system components such as hardware, systems, applications, and security organizations should also monitor data. The science of monitoring data is known as "data observability". It provides AI deployers broad visibility over their data landscape and multilayer data dependencies, such as data pipelines, data infrastructure and data applications. By continuously monitoring, tracking, alerting, analysing and troubleshooting problems. Data observability practices aim to reduce and prevent data errors or data outages within acceptable SLAs.

Data observability generally includes observing or monitoring

a. Data content,
b. Data flow and pipeline,
c. Infrastructure and compute,
d. User, usage, & utilization,
e. Financial allocations

5.2.2.5 Disclosures

AI deployers should take proactive steps to demonstrate transparency and accountability related to the development and use of AI.

Management should effectively communicate the policies and statements related to AI usage, its intended business objectives and associated risk management among other things to the relevant stakeholders to boost their confidence on the use of AI.

5.2.2.6 Integrating AI solution

A successful integration of an AI system will focus on the business potential and not on technical feasibility and prowess. The two most important requirements of an effective integration of an AI system hinges on data and testing.

High quality data will ensure the effectiveness of the underlying model. AI deployers need to diligently look into ensuring that the data sourced for building the model is valid, is adequate in quantity and complies with the legal requirements. They should also look into securing the data with adequate controls, as well as ensure that data is secured as it moves along the data supply chain.

Organizations should ensure that adequate controls are built into the system, and the processes to protect against bias. Playbooks should be created to regularly test against such threats especially focusing on known attack techniques such as adversarial attacks[12] for example: evasion, model extraction, or model poisoning[13]. Adversarial attacks target AI models or systems in their production environment but model poisoning attacks target AI models in a development or testing environment.

12 Adversarial attacks manipulate input data to cause errors or misclassification, bypassing security measures and controlling the decision-making process of AI systems

13 In model poisoning, attackers introduce malicious data into the training data to influence the output – sometimes creating a significant deviation of behavior from the AI model.

Testing the model is a crucial part of the AI development cycle. The AI deployer needs to establish clear processes to ensure that the model is tested against business objectives at every key milestone in the AI development cycle. This alone will ensure that the model delivers as expected and that it protects the organization against established and potential biases creeping into the model.

Organizations should strive to build trust in their AI systems by making them explainable, repeatable, robust, suitable and secure. The AI model should fit for purpose and fit within organizational risk appetite and assured via security validation and verification and additionally and auditable by design.

5.2.2.7 Monitoring Data Reliability

Reliability of an AI system evolves throughout its life cycle. AI deployers should develop processes to ensure that data reliability is monitored throughout the AI lifecycle. Controls are in place to flag events when the model does not perform as predicted.

Organizations should investigate the possibility of using tools that permit an AI model to report the degree of uncertainty alongside a prediction or output. Such insights can be valuable for the human operator and drive trust in building a robust AI system.

AI deployers should also consider processes and tools possible to monitor and measure model drifts[14].

5.2.2.8 Auditability of AI Systems

Organizations need to ensure that their internal audit processes cover AI systems. Similarly, it should ensure that mechanisms to facilitate the AI system's auditability (e. g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact) are built into the AI system during the design phase.

The ability to audit AI systems will contribute to trustworthy AI. In areas where AI systems are used in critical industries and operations and where it may impact the fundamental rights of users, independent audits should be considered to demonstrate transparency of AI systems. The internal audit function should build the necessary capabilities and procure tools that will help them facilitate the successful audit of AI systems.

Various techniques such as document reviews, process walkthroughs, data analysis or security assessments could be used to perform an audit. The audit teams should be cognizant of the challenges such as complexity of AI systems and the underlying models, dynamic regulatory landscape, lack of relevant technical capabilities amongst other things. The data audit process should be able to flag potential issues with the system if any, such as data bias, ethics, regulatory compliances etc.

Compliance assessments and certifications against established and relevant standards is a good way to demonstrate auditability and build trust in AI systems.

5.2.2.9 Incident Reporting

To increase trust in AI, organizations should establish effective reporting channels for the related stake-

14 Model drift refers to the problem of decay of a model's predictive power based on the changes occurring in the environment.

holders to report inappropriate functionality, discovery of a data breach[15] or other concerns about the AI system behavior.

The channel should be easy to use and available for both internal and external users. The user reports should be monitored by a human and responded to as appropriate.

5.3 Technology

AI systems as any other information system that process and analyze data, can be vulnerable due to the underlying hardware, software (firmware, operating system, virtual ware, application, etc.), network and security protocols, insufficient encryption, lack of adequate monitoring, lax access controls and internal threats.

As mentioned earlier, basic cybersecurity hygiene cannot be overlooked. AI deployers should ensure that any information system (including AI systems) comply with information security standards. The following subheadings include technical controls specifically from their utility in an AI system perspective.

5.3.1 System Design and Architecture

Building robust and secure systems is not a game of chance. This requires astute understanding of requirements (including business, functional, technical, regulatory, and financial etc.), threat modelling to understand risks and system modelling among other things. Building a trustworthy AI system requires that principles of ethics, privacy, and security are built into the system by design.

A secure-designed AI system could protect organizations against threats such as Insecure public facing deployment[16].

5.3.2 Access Control

Access control is of vital importance in an AI system primarily due to its complexity and enormity. AI system by its nature has access to large data sets, and may have several third parties as part of its supply chain who may be contributing to various processes within the system, as such it is very important to regulate access within an AI system. It is highly recommended that AI deployers use the zero-trust concept while designing and building the AI system.

A strong access control system based on a zero-trust concept, could help organizations protect against threats such as poor access control[17] and agent excessive access[18].

5.3.3 Network Security

An AI system requires and consists of multiple modules / systems connected to each other possibly across physical and organizational boundaries. This underlying connectivity is provided by "networks".

15 AI incident management should be aligned with the data breach management to ensure that all incidents are detected and handled correctly.

16 This for example could be a model deployed directly on an unsecured inference server or made directly downloadable. Also, an Inference API or Web Service being vulnerable, unpatched and not up-to-date, and excessive permissions for service accounts on inference servers.

17 This is where the underlying tech stack has insufficient access control and the attacker is able to download the model or APIs have not been designed with access control in mind.

18 This is where a public facing agent has access to private/restricted internal APIs or a public facing agent has access to private/restricted models or an agent has access to financial systems.

As such it is crucial to ensure that the network is secure by default and design. A secure network could help AI deployers protect themselves against threats such as DoS. [19]

5.3.4 Supply Chain Security

Within the context of AI systems, we need to consider the following aspects of Supply Chain.

1. Vendors providing direct services as part of ecosystem: These are vendors who could be providing services related to data management, network management etc.
2. Vendors providing indirect services such as vendors who provide the underlying hardware / software etc.

Irrespective, organizations should ensure that vendors are committed to security, adhere to best practices and are compliant with regulatory requirements. This entails having procurement processes that support security by ensuring that a vendor's security maturity and commitment to security and resilience is one criterion used to qualify potential vendors.

Vendors committed to security will help AI deployers protect against threats such as regulatory non-compliance which will reduce surface area of attacks, amongst other things.

If the organization is working with a third-party AI solution providers may conduct a thorough assessment of their security practices and data handling procedures to ensure that they meet your organization's security standards. Specifically, organizations should ensure that data exchanged with third parties including vendors, partners is governed with a data exchange policy. Vendors/partners should ensure that data is treated in accordance with the security classification accorded to the data in line with the organization's data classification policy.

Organizations must ensure that legal agreements with vendors and contractors explicitly mention security and compliance requirements.

5.3.5 Situational Awareness

In a complex system, it is imperative that you have a clear understanding of the system, its inter-connects, the behavior of the system and its impact if it malfunctions. AI deployers need to ensure that they build in the necessary processes, systems and identify tools that would provide them with the situational awareness needed to manage and monitor an AI system. Below are some ways of how to improve situational awareness for AI systems:

5.3.5.1 Vulnerability Management

Given the fact that AI systems consists of inter-connected modules, vulnerabilities in a system may have a domino effect on other interconnected systems. As such, organizations should create maps showing how the compromise of one asset or system will affects other components of an AI systems.

5.3.5.2 Bug Bounty Program

Bug bounty programs demonstrate an organization's proactiveness in identifying potential bugs and vulnerabilities. If implemented properly, they ensure that organizations can identify potential issues in their public facing systems potentially before a malicious actor does.

19 NCSA has issued Guidelines for Protecting Against Dos or DDoS Attacks.

5.3.5.3 Threat Intelligence

Considering how large and dynamic the attack surface can be including a multitude of malicious threat actors, it would be humanly impossible to track potential threats, as such, organizations should invest in procuring or building threat intelligence services. A good threat intelligence system would help organizations gain insights into potential threats primarily in the cyber space, thereby giving them an edge in terms of being proactive in mitigating these threats.

AI deployers should also actively seek collaborations with NCSA, Ministry of Interior, sector regulators and law enforcement agencies. Such agencies have visibility over the regional and national threat landscape and obtain curated threat intelligence.

5.3.6 Threat Modelling to identify threats

AI deployers should follow a diligent threat modelling in identifying potential threats to the system. A scientific approach will ensure that organizations have analyzed the system comprehensively, identified all possible threats, and evaluated potential solutions to mitigate the identified threats. Organizations may evaluate a suitable threat model for their perusal. Some of these established threat models include:

- MITRE ATLAS model developed by Microsoft, in collaboration with MITRE
- AI Threat Ontology model developed by ETSI GR SAI 001
- Good AI Assessment (GAIA) Top 10 developed by Google
- AI Security and Privacy Guide developed by OWASP

5.3.7 Resilient AI

A robust system is one of the attributes of an AI system. Building a resilient system is the cornerstone of building a robust AI system and thereby contributes to building trust in an AI system. Among other things you could use the following controls to build and evaluate the resilience of your AI system.

5.3.7.1 Backups

A key challenge in defining a backup strategy for an AI system, is the humongous data it deals with, along with the regulatory restrictions it might carry. Since AI systems use huge data sets, which may include personal data regulated under the PDPPL, organizations need to consider the necessary regulations and its impact on the data backup strategy.

5.3.7.2 Testing

Testing, validation and verification considering the complexity and criticality of the system is crucial. This should not only include the core components such as the AI model, data sets, business logic but also the underlying systems and infrastructure and more importantly to test the complete system in entirety.

AI deployers should employ techniques such as sandboxes to test applications and create dedicated test environments for testing. Creating a digital twin might be a good idea for AI systems used in critical functions.

5.3.7.3 Simulation

Besides regular testing, conducting simulation exercises using cyber ranges and simulation platforms,

drills and exercises are good ways to test, assess and secure AI systems. Such systems and techniques could also be used to assess the AI system in the business context against established processes and people skills.

5.3.8 Securing Data

As discussed earlier data is the lifeline of an AI system. A typical AI system is a collaboration of multiple stakeholders and involves data sharing and third-party access, hence the attack-surface and the risk of unauthorized access or misuse of personal data increases.

The nature, context and complexity of AI systems can make it challenging to ensure data security. Section 4 of this guideline outlines several risks and challenges related to data aggregation, data retention and data privacy.

Among other things, organizations should consider the following to ensure data security:

1. Ensure that the underlying systems, network and infrastructure storing the data are secured using the best practices.
2. Access rights, configurations and database logs are logged and reviewed on regular basis.
3. Reevaluate open data policies in the realm of AI, to ensure that the risk of loss of confidentiality due to data aggregation is mitigated.
4. Evaluate and implement a federated learning model[20], for database design. Federated learning allows models to be trained on large amounts of data while limiting the exposure or movement of raw data and can hence be seen as a special mean of data exchange.
5. Use technology available to protect and secure data. (for example: endpoint protection systems, intrusion prevention systems, data leakage prevention and data rights management solutions)
6. Ensure the integrity by applying cryptographic hash functions to the data and storing the resulting hash values. The hash values are then signed using a digital signature algorithm. This protection makes it possible to prove and verify the correctness and integrity of data.

6 Special Recommendations on Generative AI

As with any new promising technology, organizations are keen to harness and utilize the prowess of GenAI to scale up productivity and enhance the ROIs in business. As with any other technology, GenAI comes with its own set of risks and challenges and organizations should conduct the necessary due diligence.

At a minimum, AI deployers should consider the following:

Data privacy and security: Corporate environments typically involve the processing of sensitive data such as; financial information, customer data, or trade secrets. It is important to ensure that GenAI tools are designed to meet the specific security and privacy requirements of the organization and the country, and that appropriate measures are in place to protect data (such as compliance with PDPPL).

20 Ref: ETSI GR SAI 002 V1. 1. 1 (2021-08) Securing Artificial Intelligence (SAI); Data Supply Chain Security Although not free of security threats, the approach has been shown to reduce the effectiveness of a data poisoning attacks in some cases. It allows the introduction of more and more varied training data, which helps to increase the robustness of a model, and reduces the control an attacker has over the dataset they wish to poison.

Integration with existing systems: GenAI tools may need to be integrated with existing systems such as; customer relationship management (CRM) software to provide a seamless experience for users. This requires careful planning and coordination to ensure secure and reliable integration.

Training and support: GenAI tools require training data and ongoing monitoring and support to ensure that they are providing accurate and helpful responses. AI deployers need to invest in the necessary resources to ensure that Gen AI tools are properly trained and maintained. Furthermore, organizations need to invest in developing specific skills (e. g. prompt engineering) within their organization to manage these tools.

Security Awareness: General users need to be provided with security awareness on how to use these systems. Users should be trained not to rely solely on the GenAI tool's advice or information. Users should always cross-check information with other reputable sources

User Experience: Some users may be hesitant to interact with a chatbot and may prefer to speak with a human representative. It is important to communicate the benefits of GenAI tools and provide adequate training and support to ensure user acceptance. Further, to boost transparency, it is recommended to inform your users when they are interacting with bots.

Security Hygiene: To protect their data users should follow best practices for online security, such as; using strong and unique passwords, avoiding clicking on suspicious links, downloading unknown attachments, and keeping their devices and software up to date with the latest security patches and updates. Users should be educated not to share sensitive or personal information such as passwords, credit card details, and social security numbers with the GenAI tool.

Monitor performance and usage: AI deployers should monitor the performance and usage of GenAI tools to ensure that it is providing value to users and meeting the goals of the organization. This can help in the identification of improvement areas and performance optimization.

6.1 Sensitive data leakage

Employees can easily expose sensitive and proprietary business data in the questions and queries while using from GenAI tools such as: ChatGPT, Brad, Bing AI, Dall-E, WordTune Read, Whisper, Eleven Labs. In the case of ChatGPT for the time being these questions are stored indefinitely in the OpenAI infrastructure and may be similarly stored in other vendor supplied ChatGPT versions. Additionally, these questions may be used to train third-party GPT models (by AI solutions providers) in the future further compromising the confidentiality of organizational information.

To prevent sensitive data exposure, AI deployers can do the following:

- Disallow any cut-and-paste of business content (like: emails, report, chat logs) into prompts;
- Turn off chat history and data training;
- Disallow any inputs to the GenAI product that include customer data, or any personal data;
- Mandate human review of all AI generated outputs used in customer-facing interactions;
- Organizations should use secure APIs to integrate LLMs or MfMs with their systems and applications. APIs should be secured with strong authentication mechanisms, such as OAuth or API keys, to prevent unauthorized access;

- Settings should be hardened and options to use enterprise data for training and optimization should be disabled wherever possible;
- Use sandbox or content management gateways to filter data sent to GenAI tools.

6.2 Understanding the limitations of Generative AI

As GenAI tools are not explainable and highly unpredictable, they can consistently hallucinate inaccurate and fabricated information, create offensive outputs and be prone to bias; therefore, it may not be suitable for every business use case. Furthermore, outputs generated by GenAI should be assessed for accuracy, appropriateness and usefulness before being accepted.

Before deploying the LLM or MfM management should consider whether the AI system/product is fit for the organization's purpose and evaluate the performance on a wide range of potential inputs to identify cases where performance might drop.

Human beings are special. However, AI models irrespective of how intelligent the current systems are, cannot understand ambiguity, paradox, lack context, scope, situation awareness, organizational culture for a given problem and don't have consciousness. Lastly, all AI model and systems are limited in their knowledge. Therefore, the AI deployer's management should consider the customer/client base and the range of inputs that they will be using, to ensure their expectations are calibrated appropriately.

6.3 Generative AI development

GenAI allows faster development cycles than traditional AI projects. For this reason, piloting requires a lean cycle of innovation — short experiments to test how the technology could add strategic value while mitigating the potential risks that come with it.

Success in GenAI pilots requires rapid testing, refinement and often, the elimination of use cases that do not have the anticipated effect on business value. AI deployers should define the specific use cases for the GenAI application and the data sources it will use. This will ensure that the tool is designed to meet the specific needs of the organization and is integrated with the appropriate data sources.

Organizations should provide adequate training data to GenAI Tools such as LLMs and MfMs to ensure that it is properly trained and can provide accurate and relevant responses. Organizations should also monitor the performance of the tool and adjust as necessary to improve its accuracy.

AI deployers should ensure that user data is properly protected and that access to data is restricted to authorized personnel only. Data should be encrypted in transit and at rest and appropriate access controls and audit mechanisms should be in place to ensure data privacy and security.

6.3.1 Adversarial testing

Test the AI product over a wide range of inputs and user behaviors, both a representative set and those reflective of someone with malicious intentions trying to 'break' the application.

The response provided by the LLM or MfM is based on the data it is trained on. It is essential that users validate the responses they receive from the GenAI tool and do not blindly trust the system.

6.3.2 Content management

Content filtration: To avoid undesired content, it is useful to integrate a content filtration system tailored to the deployed AI system and the business use case.

Prompt engineering: Can help constrain the topic and tone of output text. This reduces the chance of producing undesired content, even if a user tries to produce it. Providing context to the AI model by giving a few high-quality examples of desired behavior prior to input can make it easier to steer model outputs in desired directions.

Prompt engineering allows AI deployers to use public GenAI services, while protecting company's IP and leveraging private data to create precise, useful and specific responses.

Prompt limitations: To decrease the chance of misuse, AI deployers can introduce prompt limitations such as the following:

- Limit the number of output tokens.
- Limit the amount of text, and the size of the file/picture/video/recording a user can input into the prompt.
- Narrow the ranges of inputs or outputs.
- Promote inputs through validated dropdown fields, rather open-ended text/ file/picture/video/recording inputs.

7 Compliance and Enforcement

This document has been issued as a Guideline to help AI deployer organizations understand the risks associate with AI systems and how to mitigate them for an effective and secure usage.

This guideline complements the existing national regulations, policies, and standards. The document needs to be read in context with the National Data Classification Policy and the National Information Assurance Standard.

8 Appendix

8.1 Ethical and Fair Ai principles

8.1.1 Transparency and Explainability

In the context of AI, transparency refers to the ability to understand and explain how an AI system takes its decisions or predictions. Furthermore, transparency is a requirement in many privacy regulations – such as in the PDPPL. Many AI models, particularly deep learning models, are often complex and involve numerous layers and parameters. These models are trained on massive amounts of data, making it challenging for humans to comprehend the exact reasoning behind their decisions. This is known as the "black box" effect.

In high-stakes domains like healthcare, cybersecurity, defense, or criminal justice, where AI systems are increasingly being deployed to make critical decisions that impact people's lives, lack of transparency can raise serious concerns. People may be hesitant to trust AI systems if they cannot understand why a certain decision was made, leading to a lack of confidence in the technology. Moreover, when AI systems make mistakes or produce biased outcomes it becomes difficult to pinpoint the exact cause or fix the issue. Developing AI systems with transparency and explainability is essential to address these concerns. The users of AI need to be aware that they are interacting with an AI system, and should be informed of

the system's capabilities and limitations.

There is an emerging field of explainable AI with tools to understand the reason for different decisions by artificial intelligence systems in various domains.

8.1.2 Accountability

Unlike traditional software or manual decision-making processes, insteads of AI systems have an autonomous nature and their behavior can be influenced by various factors, including training data, algorithms and system configuration. Determining accountability can be challenging due to the distributed nature of responsibility for the development and deployment of AI systems. Multiple parties can be involved, such as AI developers, data providers, system integrators, end-users and even the AI algorithm.

If an AI system causes harm the lack of clear accountability frameworks can lead to difficulties in identifying who should be held responsible for the consequences. Accountability means that AI deployers must put in place governance mechanisms that outline responsibilities at each stage of AI development and deployment. This includes the proactive documentation of policies, processes, and measures implemented.

8.1.3 Safety

AI systems need to be resilient and secure. Safety means that AI systems must be proactively assessed to identify harms that could result from system use, including reasonably foreseeable misuse. Measures must be taken to mitigate the risk of harm.

Safety becomes imperative, when AI systems are used in Critical Infrastructures and have the potential to harm human lives, the surrounding environment and the national economy in a significant way.

8.1.4 Privacy

Besides ensuring full compliance with the privacy and data protection regulations, adequate data governance mechanisms must also be ensured, considering the quality, integrity and legitimized access to data. For AI systems integrated into organizational processes the terms and conditions of such integrations should be reviewed to ensure privacy and that consent is obtained for the purposes of AI training and analysis.

8.1.5 Human oversight & Monitoring

Human Oversight means that high-impact AI systems must be designed and developed in such a way as to enable people managing the operations of the system to exercise meaningful oversight. Monitoring, through measurement and assessment of AI systems and their outputs, is critical to supporting effective human oversight.

This could also be a regulatory requirement as is evident in regulations such as General Data Protection Regulation (GDPR), which gives data subjects the right to obtain human intervention in cases of automated decision-making.

8.1.6 Robustness

AI systems must be stable and resilient in a variety of circumstances, actively defending against adversarial attacks, minimizing security risks and enabling confidence in system outcomes.

8.1.7 Fairness & equity

Fairness and equity refer to the equitable treatment of individuals, or groups of individuals, by an AI system. This requires building AI systems with an awareness of the potential for discriminatory outcomes. Appropriate actions must be taken including introducing adequate checks and balances within the system, to mitigate discriminatory outcomes for individuals and groups. This is a key element to ensure that we harness the benefits of AI without inadvertently introducing bias or other unfair outcomes.

8.1.8 Societal and environmental well-being

AI systems should benefit all human beings, including future generations. It must be ensured that they are sustainable and environmentally friendly. Moreover, they should consider other living beings and their social and societal impact should be carefully considered.

Another aspect of societal wellbeing is also being able to harness the prowess of the technology for the general wellbeing of humans, environment and society.

8.1.9 Auditability

Transparency and insights into how an AI enabled decision was made would significantly drive trust in an AI system. AI deployers should strive to ensure that their AI systems are auditable. An audit of an AI system enables interested third parties to probe, understand and review the behavior of the algorithm through disclosure of information that enables monitoring, checking or criticism.

AI auditability covers amongst other things:

1. Evaluation of models, algorithms and data streams
2. Analysis of operations, results and anomalies observed
3. Technical aspects of AI systems for results accuracy
4. Ethical aspects of AI systems for fairness, legality and privacy

8.2 Acronyms

ADM	Automated Decision Making
AI	Artificial Intelligence
AUP	Acceptable User Policies
AS	Autonomous systems
CRM	Customer Relationship Management
DoS	Denial of Service
GPU	Graphics Processing Unit
ML	Machine Learning
MfM	Multimodal Foundation Model
NLP	Natural Language Processing
LLM	Large Language Model
TPU	Tensor Processing Unit

8.3 Normative References

Amiri Decree No 1 of year 2021

President of National Cyber Security Agency (NCSA) Decision No 3 of year 2022

National Data Classification Policy

Cyber Security Guidelines for DDoS Attacks

Law No. 13 of 2016 on Personal Data Privacy Protection (PDPPL)

National Data Privacy Office has issued several guidelines regarding the PDPPL compliance. You can find them from https://compliance. qcert. org/en/privacy/hub

Law No. 14 of 2014 Promulgating the Cybercrime Prevention Law

8.4 Informative References

8.4.1 Reports and frameworks

Organization for Economic Co-operation and Development (OECD). (2019) *Recommendation of the Council on Artificial Intelligence* (OECD ref. No. OECD/LEGAL/0449)

Australian Government Department of Industry, Science and Resources (2023) *Safe and responsible AI in Australia Discussion paper*

Singapore Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC). (2020) *Model Artificial Intelligence Governance Framework, Second Edition*

European Cybersecurity Agency (ENISA). (2023) Cybersecurity of AI and Standardization

European Commission High Level Expert Group on Artificial Intelligence (EU-HLEG). (2019) *Ethics Guidelines for Trustworthy AI*

European Commission High Level Expert Group on Artificial Intelligence (EU-HLEG). (2019) *Definition of AI: Main Capabilities and Disciplines*

Gartner. (2023) Applying AI – Governance and Risk Management.

8.4.2 Standards

NCSA. (2023) National Information Assurance Standard

National Institute of Standards and Technology. (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1. 0)* (NIST Standards No. NIST AI 100-1)

International Organization for Standardization. (2022) Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations (ISO Standard No. 38507:2022)

International Organization for Standardization. (2023) Information technology – Artificial intelligence – Guidance on risk management (ISO Standard No. 23894:2023)

International Organization for Standardization. (2022) Information technology – Artificial intelligence – Artificial intelligence concepts and terminology (ISO Standard No. 22989:2022)

International Organization for Standardization. (2023) Security and privacy in artificial intelligence use cases – Best practices (ISO Standard No. 27563:2023)

European Telecommunications Standards Institute. (2022) Securing Artificial Intelligence, AI Threat Ontology (ETSI Report No. ETSI GR SAI 001)

European Telecommunications Standards Institute. (2020) Securing Artificial Intelligence, Problem Statement (ETSI Report No. ETSI GR SAI 004)

European Telecommunications Standards Institute. (2021) *Securing Artificial Intelligence, Mitigation Strategy Report* (ETSI Report No. ETSI GR SAI 005)

European Telecommunications Standards Institute. (2023) Securing Artificial Intelligence, Explicability and transparency of AI processing (ETSI Report No. ETSI GR SAI 007)

8.4.3 Academic papers

Stefan Larsson (2020). *On the Governance of Artificial Intelligence through Ethics Guidelines*. Asian Journal of Law and Society No. 7, pp. 437-451.

Thilo Hagendorff (2020) *The Ethics of AI Ethics: An Evaluation of Guidelines*. Minds and Machines No. 30, pp. 99-120.

Daron Acemoglu (2021) *Harms of AI* National Bureau of Economic Research Working Paper No. 29247.

8.5 List of Figures

www.ncsa.gov.qa
info@ncsa.gov.qa 16555