رؤيـــة VISION
2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

AI Ethics Principles

Version 1.0

August 2022

Table of Contents

Introduction

Due to the fast growth of practices and technologies around Artificial Intelligence (AI), the use of AI has expanded to several industries such as health, education, entertainment, etc. AI helps make entities decision-making processes more efficient, accurate and faster by predicting future patterns. AI can be used to analyze data, including big data, by developing and operating systems with advanced models and algorithms that help improve the quality of processes. Due to the increasing interest in these technologies, many entities in both the public and private sector, in addition to non-profit entities, have developed digital solutions based on AI to address existing challenges using creative and innovative methods, thus making the role of AI more essential to maintain the competitive capabilities of these entities.

In accordance with KSA's commitment to human rights and its cultural values, as well as aligning to international standards and recommendations on the ethics of Artificial Intelligence and with reference to the Council of Ministers' Resolution No. (292) dated 27/04/1441 AH., stating in Paragraph (1) of Article 10 that NDMO is mandated to develop policies, governance mechanisms, standards, and controls related to data and Artificial Intelligence and monitor compliance therewith upon issuance, NDMO has analyzed global practices and standards to develop this AI Ethics Framework which aims to:

- Support the Kingdom's efforts towards achieving its vision and national strategies related to adopting AI technology, encouraging research and innovation, and driving economic growth for prosperity and development.
- Develop AI Ethics policies, guidelines, regulations, and frameworks.
- Enable regulatory authorities to develop their own policies or codes of conduct, execute their plans, and utilize AI technology to forecast the future and make decisions in line with the national vision and strategies.
- Govern data and AI models to limit the negative implications of AI systems (economically, psychologically, socially, etc.) and potential threats (security, political, etc.).
- Help entities adopt standards and ethics when building and developing AI-based solutions to ensure responsible use thereof.
- Protect the privacy of data subjects and their rights with respect to the collection and processing of their data.
- Coordinate and work with regional and international organizations to align and adopt best practice AI Ethics application and governance

Definitions

For the purposes of this document, the following words and phrases, wherever mentioned herein, shall have meanings ascribed thereto, unless the context requires otherwise:

Terminology	Definition
Adopting Entity	Any Public Entity, business, or individual that is required to comply with the present document.
Artificial Intelligence	Artificial intelligence or AI is a collection of technologies that can enable a machine or system to sense, comprehend, act, and learn.
AI System	A set of predictive models and advanced algorithms that can be used to analyze data and predict the future or facilitate decision-making for projected future events.
AI System Assessor	Any natural or legal person that audits AI systems to achieve certain goals.
AI System Developer	Any natural or legal person that develops AI systems by building predictive models using data and algorithms to achieve certain goals.
AI System Lifecycle	The cyclical process that AI projects are expected follow to be able to design, build, and produce a robust and safe system that delivers business value and insights through adhering to a standard and structured way of managing AI model delivery and implementation.
AI System Owner	Any natural or legal person that applies or uses AI systems to achieve certain goals.
Authorized User	An individual that is permitted, appropriately cleared and has a requirement to access an information system to perform or assist an established and preset role or responsibility on the system's functionalities and components.
CCO/CO	The Chief Compliance Officer (CCO) or Compliance Officer (CO) reports to the head of the Entity or Chief Data Officer and is responsible for developing and maintaining ethical and compliance standards that entities follow whenever expanding operations, recruiting, acquiring further assets, etc.
CDO	The Chief Data Officer (CDO) is responsible for the development and execution of the Data Management & Data Governance and oversee the implementation of data management practices across the Public Entities.
Code of Ethics & Conduct	A code of ethics and conduct is a guide of principles designed to help professionals conduct business honestly and with integrity upholding fundamental human rights and values.
Data	A collection of facts in a raw or unorganized form such as numbers, characters, images, video, voice recordings, or symbols.
Data Governance	Data governance is the process of managing the availability, usability, integrity, and security of data in organizations and systems, based on data standards and policies that also control data usage.
Data Sample	The data used to build, train, and test predictive models and AI algorithms to reach specific results.

Terminology	Definition
Data Subject	An individual to whom the personal data belongs, his representative, or whoever has legal guardianship over him.
End-User	Any natural or legal person that consumes or makes use of the goods or services produced by AI systems.
EO	The Ethics Officer (EO) is responsible for the ethics compliance within an entity.
KPI	Key Performance Indicator
KSA	Kingdom of Saudi Arabia
National Regulatory Authority	Any independent governmental or public entity assuming regulatory duties and responsibilities for a specific sector in the Kingdom of Saudi Arabia under a legal instrument.
NDMO	National Data Management Office
Personal Data	Every data – of whatever source or form – that would lead to the identification of the individual specifically, or make it possible to identify him directly or indirectly, including: name, personal identification number, addresses, contact numbers, license numbers, records, personal property, bank account and credit card numbers, fixed or removing pictures of the individual, and other data of personal nature.
SDAIA	Saudi Data and Artificial Intelligence Authority
Sensitive Data	Any personal data that indicates or includes a reference to a person's ethnic or tribal origin; religious, intellectual or political beliefs; membership of civil associations or institutions; criminal and security data; biometric identifying data; genetic data; credit data; health data; location data; and data that indicates that one or both of an individual's parents are unknown.
Third Party	A natural or legal person, public entity, agency or body other than the following main participants of AI system End-user, AI System Owner, AI System Developer and AI System Assessor.

Scope

This AI Ethics Framework shall apply to all AI stakeholders designing, developing, deploying, implementing, using, or being affected by AI systems within KSA, including but not limited to public entities, private entities, non-profit entities, researchers, public services, institutions, civil society organizations, individuals, workers, and consumers. Certain aspects of this framework may be waived for specific use cases on good cause shown for AI systems built for the purposes of maintaining public health or safety, protecting life or health of a particular individual(s), or looking after vital public interests of the Kingdom.

AI System Lifecycle

The AI System Lifecycle is the cyclical process that AI projects follow. It defines each step that an organization is expected to follow to take advantage of AI to derive practical business value. It is a standard way of representing the tasks based on best practices in implementing and managing AI models which makes it a great candidate for embedding AI ethics.

The AI System Lifecycle is split into four steps, all of which have equal importance, and the relevant activities are explained below.

Plan and Design:

- Define the problem
- Support your problem with a data-driven approach
- Select a framing approach on technology and system which governs AI
- Conduct feasibility assessment for the selected approach
- Define KPI's

Prepare Input Data:

- Gather data
- Discover and assess data
- Cleanse and validate data
- Transform data into AI model input features

Build and Validate:

- Execute the defined approach
- Train and test the model
- Tune the hyperparameters of the model
- Validate model performance

Deploy and Monitor:

- Deploy the model to the AI system
- Create versioning structure
- Monitor the production model performance periodically
- Assess if there is a need to change the design according to results of periodic reviews

AI Ethics practices cut across the AI System Lifecycle. It is therefore important to ensure that ethical principles and their corresponding controls are embedded and reflected within the activities of the mentioned AI System Lifecycle phases. As a result, this Framework is structured based on AI Ethics Principles, which are supported by controls and are categorized based on AI System Lifecycle steps to showcase an organized and comprehensive approach.

AI Ethics Principles and Controls

The KSA AI Ethics principles have been defined in line with the global benchmarks and KSA's cultural values. The principles in this section are supported by guiding controls across the AI System Lifecycle. Controls outlined in this document have been anchored by these principles, which in turn should guide any future evolution of the KSA AI Ethics Agenda.

Principle 1 – Fairness

The fairness principle requires taking necessary actions to eliminate bias, discrimination or stigmatization of individuals, communities, or groups in the design, data, development, deployment and use of AI systems. Bias may occur due to data, representation or algorithms and could lead to discrimination against the historically disadvantaged groups.

When designing, selecting, and developing AI systems, it is essential to ensure just, fair, non-bias, non-discriminatory and objective standards that are inclusive, diverse, and representative of all or targeted segments of society. The functionality of an AI system should not be limited to a specific group based on gender, race, religion, disability, age, or sexual orientation. In addition, the potential risks, overall benefits, and purpose of utilizing sensitive personal data should be well motivated and defined or articulated by the AI System Owner.

To ensure consistent AI systems that are based on fairness and inclusiveness, AI systems should be trained on data that are cleansed from bias and is representative of affected minority groups. AI algorithms should be built and developed in a manner that makes their composition free from bias and correlation fallacy.

Plan and Design:

At the initial stages of setting out the purpose of the AI system, the design team shall collaborate to pinpoint the objectives and how to reach them in an efficient and optimized manner. Planning the design of the AI system is an essential stage to translate the system's intended goals and outcomes. During this phase, it is important to implement a fairness-aware design that takes appropriate precautions across the AI system algorithm, processes, and mechanisms to prevent biases from having a discriminatory effect or lead to skewed and unwanted results or outcomes.

Fairness-aware design should start at the beginning of the AI System Lifecycle with a collaborative effort from technical and non-technical members to identify potential harm and benefits, affected individuals and vulnerable groups and evaluate how they are impacted by the results and whether the impact is justifiable given the general purpose of the AI system.

A fairness assessment of the AI system is crucial, and the metrics should be selected at this stage of the AI System Lifecycle. The metrics should be chosen based on the algorithm type (rule-based, classification, regression, etc.), the effect of the decision (punitive, selective, etc.), and the harm and benefit on correctly and incorrectly predicted samples.

Sensitive personal data attributes relating to persons or groups which are systematically or historically disadvantaged should be identified and defined at this stage. The allowed threshold which makes the assessment fair or unfair should be defined. The fairness assessment metrics to be applied to sensitive features should be measured during future steps.

Prepare Input Data:

Following the best practice of responsible data acquisition, handling, classification, and management must be a priority to ensure that results and outcomes align with the AI system's set goals and objectives. Effective data quality soundness and procurement begin by ensuring the integrity of the data source and data accuracy in representing all observations to avoid the systematic disadvantaging of under-represented or advantaging over-represented groups. The quantity and quality of the data sets should be sufficient and accurate to serve the purpose of the system. The sample size of the data collected or procured has a significant impact on the accuracy and fairness of the outputs of a trained model.

Data accuracy and awareness should be adopted when managing and classifying data (labelling, annotating, and organizing) to avoid the introduction of human judgement and bias and ensure data fairness.

Sensitive personal data attributes which are defined in the plan and design phase should not be included in the model data not to feed the existing bias on them. Also, the proxies of the sensitive features should be analyzed and not be included in the input data. In some cases, this may not be possible due to the accuracy or objective of the AI system. In this case, the justification of usage of the sensitive personal data attributes or their proxies should be provided.

Build and Validate:

At the build and validate stage of the AI System Lifecycle, it is essential to take into consideration implementation fairness as a common theme when building, testing, and implementing the AI system. Model building and feature selection will require engineers and designers to be aware that the choices made about grouping or separating and including or excluding features as well as more general judgements about the reliability and security of the total set of features may have significant consequences for vulnerable or protected groups. During the selection of the champion model, the fairness metric assessment should be considered. The champion model fairness metrics should be within the defined threshold for the sensitive features. The optimization approach of fairness and performance metrics should be clearly set throughout this phase. The fairness assessment should be justified if the champion model does not pass the assessment.

Causality-based feature selection should be ensured. Selected features should be verified with business owners and non-technical teams.

Automated decision-support technologies present major risks of bias and unwanted application at the deployment phase, so it is critical to set out mechanisms to prevent harmful and discriminatory results at this phase.

Deploy and Monitor:

Well-defined mechanisms and protocols should be set in place when deploying the AI system to measure the fairness and performance of the outcomes and how it impacts individuals and communities. When analyzing the outcomes of the predictive model, it should be assessed if represented groups in the data sample receive benefits in equal or similar portions and if the AI

system disproportionately harms specific members based on demographic differences to ensure outcome fairness.

The predefined fairness metrics should be monitored in production. If there is any deviation from the allowed threshold, it should be investigated whether there is a need to renew the model.

The overall harm and benefit of the system should be quantified and materialized on the sensitive groups.

Principle 2 – Privacy & Security

The privacy and security principle represents overarching values that requires AI systems; throughout the AI System Lifecycle; to be built in a safe way that respects the privacy of the data collected as well as uphold the highest levels of data security processes and procedures to keep the data confidential preventing data and system breaches which could lead to reputational, psychological, financial, professional, or other types of harm. AI systems should be designed with mechanisms and controls that provide the possibility to govern and monitor its outcomes and progress throughout its lifecycle to ensure continuous monitoring within the privacy and security principles and protocols set in place.

Plan and Design:

The planning and design of the AI system and its associated algorithm must be configured and modelled in a manner such that there is respect for the protection of the privacy of individuals, personal data is not misused and exploited, and the decision criteria of the automated technology is not based on personally identifying characteristics or information. The use of personal information should be limited only to that which is necessary for the proper functioning of the system. The design of AI systems resulting in the profiling of individuals or communities may only occur if approved by Chief Compliance and Ethics Officer, Compliance Officer or in compliance with a Code of Ethics and Conduct developed by a National Regulator Authority for the specific sector or industry. The security and protection blueprint of the AI system, including the data to be processed and the algorithm to be used, should be aligned to best practices to be able to withstand cyberattacks and data breach attempts.

Privacy and security are usually defined as non-functional requirements for AI systems. Privacy and security legal frameworks and standards should be followed and customized for the particular use case or organization. An important aspect of privacy and security is data architecture; consequently, data classification and profiling should be planned to define the levels of protection and usage of personal data. Security mechanisms for de-identification should be planned for the sensitive or personal data in the system. Furthermore, read/write/update actions should be authorized for the relevant groups.

Prepare Input Data:

The exercise of data procurement, management, and organization should uphold the legal frameworks and standards of data privacy. Data privacy and security protects information from a wide range of threats.

The confidentiality of data ensures that information is accessible only to those who are authorized to access the information and that there are specific controls that manage the delegation of authority. Designers and engineers of the AI system must exhibit the appropriate levels of integrity to safeguard the accuracy and completeness of information and processing methods to ensure that the privacy and security legal framework and standards are followed. They should also ensure that the availability and storage of data is protected through suitable security database systems. All processed data should be classified to ensure that it receives the appropriate level of protection in accordance with its sensitivity or security classification and that AI system developers and owners are aware of the classification or sensitivity of the information they are handlining and the associated requirements to keep it secure. All data shall be classified in terms of business requirements, criticality, and sensitivity in order to prevent unauthorized disclosure or modification. Data classification should be conducted in a contextual manner that does not result in the inference of personal information. Furthermore, de-identification mechanisms should be employed based on data classification as well as requirements relating to data protection laws.

Data backups and archiving actions should be taken in this stage to align with business continuity, disaster recovery and risk mitigation policies.

Build and Validate:

Privacy and security by design should be implemented while building the AI system. The security mechanisms should include the protection of various architectural dimensions of an AI model from malicious attacks. The structure and modules of the AI system should be protected from unauthorized modification or damage of any of its components. The AI system should be secured to maintain the integrity of the information that it processes. The AI system should be secure so that it remains continuously functional and ready for use by authorized users and keep confidential and private information secure even under hostile or adversarial conditions. Furthermore, appropriate safeguards should be put in place to ensure that automated decision-making AI systems uphold relevant data privacy and security requirements. The AI System should be tested to ensure that the combination of available data does not reveal the sensitive data or break the anonymity of the observation.

Deploy and Monitor:

After the deployment of the AI system, when its outcomes are realized, there must be continuous monitoring to ensure that the AI system is privacy-preserving, safe and secure. The privacy impact assessment and risk management assessment should be continuously revisited to ensure that societal and ethical considerations of regularly evaluated. AI System Owners should be accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system. The components of the AI system should be updated based on the continuous monitoring and privacy impact assessment.

Principle 3 – Humanity

The humanity principle highlights that AI systems should be built using an ethical methodology to be just and ethically permissible, based on intrinsic and fundamental human rights and cultural values to generate beneficial impact on individual stakeholders and communities, in both the long and short-term goals and

objectives to be used for the good of humanity. Predictive models should not be designed to deceive, manipulate, or condition behavior that is not meant to empower, aid, or augment human skills but should adopt a more human-centric design approach that allows for human choice and determination.

Plan and Design:

It is essential to design and build a model that is based on the fundamental human rights and cultural values and principles that are applied within and on the AI system's decisions, processes, and functionalities. The designers of the AI model should define how the AI system will align to fundamental human rights and KSA's cultural values while designing, building, and testing the technology; as well as how the AI system and its outcomes will strive to achieve and positively contribute to augment and complement human skills and capabilities.

Prepare Input Data:

To ensure that AI models embody a human-centric build and design requires adhering to practices of responsible and ethical data management frameworks and processes to be followed according to best practices and data regulations within KSA. Data must be properly acquired, classified, processed, and accessible to ensure respect for human rights, KSA's cultural values and preferences.

Build and Validate:

When constructing AI systems, designers and engineers should prioritize building AI systems and algorithms that allow and facilitate decision making with an outlook of aligning to human rights and KSA's cultural values. The automated decisions that result from AI systems should not act in a partial and standalone manner without considering broader human rights and cultural values in its final outcomes and results. To achieve this, designers should enable AI systems with the appropriate parameters and algorithm training to attain outcomes that advance humanity.

Deploy and Monitor:

Periodic assessments of the deployed AI system should be conducted to ensure that its results are aligned to set human rights and cultural values, accuracy key performance indicators (KPIs), and impact on individuals or communities to ensure the continuous improvement of the technology.

Designers of AI models should establish mechanisms of assessing AI systems against fundamental human rights and cultural values to mitigate against any negative and harmful outcomes resulting from the use of the AI system. If any negative and harmful outcomes are found, the owner of the AI system should identify the areas that need to be addressed and apply corrective measures to recursively improve the functioning and outcomes of the AI system.

Principle 4 – Social & Environment Benefits

The social and environmental benefit principle embraces the beneficial and positive impact of social and environmental priorities that should benefit individuals and the wider community that focus on sustainable goals and objectives. AI systems should neither cause nor accelerate harm or otherwise adversely affect human beings but rather contribute to empowering and complementing social and environmental progress

while addressing associated social and environmental ills. This entails the protection of social good as well as environmental sustainability.

Plan and Design:

AI systems have a significant impact on communities and the ecosystems that they live in; hence AI System Owners should have a high sense of awareness that these technologies may have disruptive and transformative effects on society and the environment. The design of AI systems should be approached in an ethical and sensitive manner in line with the values of prevention of harm to both human beings and the environment. When planning and designing AI systems, due consideration should be given to preventing and helping address social and environmental issues in a way that will ensure sustainable social and ecological responsibility.

Prepare Input Data:

The processes and policies that govern data management should be followed when preparing the categorization and structuring of data that will feed into the AI system. The data pertaining to the social and environmental topics should be accessible to the public data infrastructure and must clearly articulate the social benefit of the data presented.

Build and Validate:

The models and algorithms must have, as their ultimate goal, a result linked to a socially recognized end, with the ability to demonstrate how the expected results relate to that social or environmental purpose through transformative and impactful benefits where applicable. It is best practice to measure and maintain acceptable levels of resource usage and energy consumption during this phase setting the tone that AI systems not only strive to foster AI solutions that address global concerns relating to social and environmental issues but also practice sustainable and ecological responsibilities.

Deploy and Monitor:

After the deployment of the AI system, the AI System Owner should ensure that continuous assessment of the human, social, cultural, economic and environmental impact of AI technologies are carried out with full cognizance of the implications of the AI system for sustainability as a set of constantly evolving goals across a range of dimensions against the priority objectives that were set at the Plan and Design phase as well as foster and encourage the power of AI solutions in addressing areas of global concern aligning to sustainable development goals.

Principle 5 – Reliability & Safety

The reliability and safety principle ensures that the AI system adheres to the set specifications and that the AI system behaves exactly as its designers intended and anticipated. Reliability is a measure of consistency and provides confidence in how robust a system is. It is a measure of dependability with which it operationally conforms to its intended functionality and the outcomes it produces. On the other hand, safety is a measure of how the AI system does not pose a risk of harm or danger to society and individuals. As an illustration, AI systems such as autonomous vehicles can pose a risk to people's lives if living organisms are not properly

recognized, certain scenarios are not trained for or if the system malfunctions. A reliable working system should be safe by not posing danger to society and should have built-in mechanisms to prevent harm.

The risk mitigation framework is closely related to this principle. Potential risks and unintended harms should be minimized in this aspect.

The predictive model should be monitored and controlled in a periodic and continuous manner to check if its operations and functionality are aligned to the designed structure and frameworks in place. The AI system should be technically sound, robust, and developed to prevent malicious usage to exploit its data and outcomes to harm entities, individuals or communities. A continuous implementation/continuous development approach is essential to ensure reliability.

Plan and Design:

There is a great need to design and develop an AI system that can withstand the uncertainty, instability, and volatility that it might encounter. Planning to set out a robust and reliable AI system that works with different sets of inputs and situations is essential to prevent unintended harm and mitigate risks of system failures when positioned against unknown and unforeseen events. Establishing a set of standards and protocols for assessing the reliability of an AI system is necessary to secure the safety of the system's algorithm and data output. It is essential to keep a sustainable technical outlay and outcomes generated from the system to maintain the public's trust and confidence towards the AI system. The documentation standards are essential to track the evolution of the system, foresee possible risks and fix vulnerabilities. All critical decision points in the system design should be subject to sign-off by relevant stakeholders to minimize risks and make stakeholders accountable for the decisions.

Prepare Input Data:

Adequate steps and actions should be taken to measure the data sample's quality, accuracy, suitability, and credibility when dealing with the data sets of an AI model. This is essential to ensure the accuracy of data interpretation by the AI system, the consistency of avoiding misleading measurements, as well as ensuring the relevance of the AI system's outcomes to the purpose of the model.

It is crucial for the build and validate step to test how the system behaves under outlier events, extreme parameters, etc. In this step, stress test data should be prepared for extreme scenarios.

Build and Validate:

To develop a sound and functional AI system that is both reliable and safe, the AI system's technical construct should be accompanied by a comprehensive methodology to test the quality of the predictive data-based systems and models according to standard policies and protocols.

To ensure the technical robustness of an AI system rigorous testing, validation, and re-assessment as well as the integration of adequate mechanisms of oversight and controls into its development is required. System integration test sign-off should be done with relevant stakeholders to minimize risks and liability.

Automated AI systems involving scenarios where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life and death decisions should trigger human oversight and final determination. Furthermore, AI systems should not be used for social scoring or mass surveillance purposes.

Deploy and Monitor:

Monitoring the robustness of the AI system should be adopted and undertaken in a periodic and continuous manner to measure and assess any risks related to the technicalities of the AI system (an inward perspective) as well as the magnitude of the risk posed by the system and its capabilities (an outward perspective).

Principle 6 – Transparency & Explain3ability

The transparency and explainability principle is crucial for building and maintaining trust in AI systems and technologies. AI systems must be built with a high level of clarity and explainability as well as features to track the stages of automated decision-making, particularly those which may lead to detrimental effects towards data subjects. It follows that data, algorithms, capabilities, processes, and purpose of the AI system need to be transparent and communicated as well as explainable to those who are directly and indirectly affected. The degree to which the system is traceable, auditable, transparent, and explainable is dependent on the context and purpose of the AI system and the severity of the outcomes that may result from the technology. AI systems and their designers should be able to justify how the rationale behind its design, practices, processes, algorithm, and decision or behaviors are ethically permissible, nondiscriminatory, and nonharmful to the public.

Plan and Design:

When designing a transparent and trusted AI system, it is vital to ensure that stakeholders affected by AI systems should be fully aware and informed of how outcomes are processed. They should further be given access and explanation to the rationale for decisions made by the AI technology in an understandable and contextual manner. Decisions should be traceable. AI system owners must define the level of transparency for different stakeholders on the technology based on data privacy, sensitivity, and authorization of the stakeholders. The AI system should be designed to include an information section in the platform to give an overview of the AI model decisions as part of the overall transparency application of the technology. Information sharing as a sub-principle should be adhered to with end-users and stakeholders of the AI system upon request or open to the public, depending on the nature of the AI system and target market. The model should establish a process mechanism to log and address issues and complaints that arise to be able to resolve them in a transparent and explainable manner.

Prepare Input Data:

The data sets and the processes that yield the AI system's decision should be documented to the best possible standard to allow for traceability and an increase in transparency. The data sets should be assessed on the context around their accuracy, suitability, validity, and source. This has a direct effect on the training and implementation of these systems since the criteria for the data's organization, and

structuring must be transparent and explainable in their acquisition and collection adhering to data privacy regulations and intellectual property standards and controls.

Build and Validate:

Transparency in AI is thought about in two perspectives, the first is the process behind it (the design and implementation practices that lead to an algorithmically supported outcome) and the second is in terms of its product (the content and justification of that outcome). Algorithms should be developed in a transparent way to ensure that input transparency is evident and explainable to the end-users of the AI system to be able to provide evidence and information on the data used to process the decisions that have been processed. Transparent and explainable algorithms ensure that stakeholders affected by AI systems, both individuals and communities, are fully informed when an outcome is processed by the AI system by providing the opportunity to request explanatory information from the AI system owner. This enables the identification of the AI decision and its respective analysis which facilitates its auditability as well as its explainability. If the AI system is built by a third party, AI system owners should make sure that an AI Ethics due diligence is carried out and all the documentation are accessible and traceable before procurement or sign-off.

Deploy and Monitor:

Upon deployment of the AI system, performance metrics relating the AI system's output, accuracy and alignment to priorities and objectives, as well as its measured impact on individuals and communities should be documented, available and accessible to stakeholders of the AI technology. Information on any system failures, data breaches, system breakdowns, etc. should be logged and stakeholders should be informed about these instances keeping the performance and execution of the AI system transparent. Periodic UI and UX testing should be conducted to avoid risk of confusion, confirmation of biases, or cognitive fatigue of the AI system.

Principle 7 – Accountability & Responsibility

The accountability and responsibility principle holds designers, vendors, procurers, developers, owners and assessors of AI systems and the technology itself ethically responsible and liable for the decisions and actions that may result in potential risk and negative effects on individuals and communities. Human oversight, governance, and proper management should be demonstrated across the entire AI System Lifecycle to ensure that proper mechanisms are in place to avoid harm and misuse of this technology. AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. The designers, developers, and people who implement the AI system should be identifiable and assume responsibility and accountability for any potential damages the technology has on individuals or communities, even if the adverse impact is unintended. The liable parties should take necessary preventive actions as well as set risk assessment and mitigation strategy to minimize the harm due to the AI system. The accountability and responsibility principle is closely related to the fairness principle. The parties responsible for the AI system should ensure that fairness of the system is maintained and sustained through control mechanisms. All parties involved in the AI System Lifecycle should consider and action these values in their decisions and execution.

Plan and Design:

This step is crucial to design or procure an AI System in an accountable and responsible manner. The ethical responsibility and liability for the outcomes of the AI system should be attributable to stakeholders who are responsible for certain actions in the AI System Lifecycle. It is essential to set a robust governance structure which defines the authorization and responsibility areas of the internal and external stakeholders without leaving any areas of uncertainty to achieve this principle. The design approach of the AI system should respect human rights, and fundamental freedoms as well as the national laws and cultural values of the kingdom. Additionally, organizations can put in place additional instruments such as impact assessments, risk mitigation frameworks, audit and due diligence mechanisms, redress, and disaster recovery plans. It is essential to build and design a human-controlled AI system where decisions on the processes and functionality of the technology are monitored, executed, and are susceptible to intervention from authorized users. Human governance and oversight establish the necessary control and levels of autonomy through set mechanisms.

Prepare Input Data:

An important aspect of the Accountability and Responsibility principle during Prepare Input Data step in the AI System Lifecycle is data quality as it effects the outcome of the AI model and decisions accordingly. It is, therefore, important to do necessary data quality checks, clean data and ensure the integrity of the data in order to get accurate results and capture intended behavior in supervised and unsupervised models. Data sets should be approved and signed-off before commencing with developing the AI model. Furthermore, the data should be cleansed from societal biases. In parallel with the fairness principle, the sensitive features should not be included to the model data. In the event that sensitive features need to be included, the rationale or trade off behind the decision for such inclusion should be clearly explained. The data preparation process and data quality checks should be documented and validated by responsible parties. The documentation of the process is necessary for auditing and risk mitigation. Data must be properly acquired, classified, processed, and accessible to ease human intervention and control at later stages when needed.

Build and Validate:

Model development of the AI system and algorithm should consist of the selection of features, hyperparameter tuning and performance metric selection. To achieve this, the technical stakeholders who build and validate models should be responsible for these decisions. Assigning the appropriate ownership and communicating responsibilities will set the tone for accountability that would aid in steering the development of the AI system on good reasons, solid interference, and will allow the intervention of human critical judgement and expertise. The decisions should be supported with quantitative (performance measures on train/test datasets, consistency of the performance on different sensitive groups, performance comparison for each set of hyperparameters, etc.) and qualitative indicators (decisions to mitigate and correct unintended risks from inaccurate predictions). The appropriate stakeholders and owners of the AI technology should review and sign-off the model after successful testing and validation of user acceptance testing rounds have been conducted and completed before the AI models can be productionized.

Deploy and Monitor:

The responsibility and associated liability in the Deploy and Monitor step should be set clearly. The outcomes and decisions set in the build and validate step should be monitored continuously and should result in periodic performance reports. Predefined triggers/alerts should be defined for this step on the data and performance metrics. Setting of these triggers is a rigorous process and each trigger should be assigned to the appropriate stakeholder. These triggers/alerts can be defined as part of the risk mitigation or disaster recovery procedure and may need human oversight.

Roles and Responsibilities

The AI Ethics Framework defines the following roles and responsibilities both at the national and Entity level.

National Level

NDMO

NDMO shall act as the KSA AI Ethics initiatives owner and operator, coordinating AI Ethics activities at the national level. It shall provide the KSA AI Ethics strategic direction and develop the national regulations, standards, and guidelines to ensure AI Ethics are effectively managed and published across the Kingdom and yields the targeted national impact.

NDMO, as the national regulator of the data agenda, shall monitor compliance against this AI Ethics Principles with the support of the Regulatory Authorities.

As such, NDMO's responsibilities shall include:

- **AI Ethics Development –** Develop, issue, and update this AI Ethics principles and framework. This document should be revisited regularly to address possible changes affecting AI Ethics, associated regulations, communities, and environment.
- **AI Ethics Adoption Plan Development –** Develop supporting material and provide continuous guidance to Adopting Entities to facilitate the adoption of this Framework.
- **AI Ethics Advisory –** Support Adopting Entities on complying with this Framework and answer any queries related to the AI Ethics and compliance with this document.
- **AI Ethics Compliance Measurement –** Measure compliance of Adopting Entities on a regular basis based on the defined compliance mechanism directly or through sector regulators (Refer to the 'Compliance' section for more details) and audit AI Ethics activities when required.
- **AI Ethics Education and Awareness –** Conduct and monitor communication and training initiatives to drive AI Ethics awareness and adoption at the national level.
- **AI Ethics Performance –** Analyze AI Ethics KPI's and impact at the national level and synthesize improvement opportunities to be communicated to relevant stakeholders.
- **AI Ethics Compliance Monitoring –** Conduct investigations, audits and monitor compliance against this Framework with the support of the National Regulatory Authorities.

- **Compliance Portal –** Design, build, and maintain the Compliance Portal to ensure Adopting Entities can publish, manage, and update their reports and include AI Ethics guiding/supporting material. The portal shall also be used to report any breaches or irregularities with the deployed AI systems with potential adverse effects on communities or the environment.

National Regulatory Authority

NDMO may delegate oversight and enforcement of AI Policy for particular sectors to National Regulatory Authorities. If a National Regulatory Authority is appointed for this role, the responsibilities of National Regulatory Authority shall include:

- Acting as NDMO proxy for the responsibilities: AI Ethics Adoption Plan Development, AI Ethics Framework Advisory, AI Ethics Compliance Measurement, AI Ethics Education and Awareness and AI Ethics Performance.
- Publishing additional sector specific guidelines or Codes of Conduct to support implementation of AI Ethics Framework.

Entity Level

All Adopting Entities shall have the primary responsibility for ensuring that their AI Ethics documents are published in compliance with this AI Ethics Principles. As such, Entities shall designate individuals who will be responsible for carrying out the AI Ethics activities as outlined below.

Head of the Entity / Chief Data Officer (CDO) - The head of the Entity or their designate shall be accountable for the AI Ethics practice within a Private Entity, while CDO shall be accountable for the AI Ethics practice within a Public Entity. Responsibilities include:

- **AI Ethics Plan Approval –** Approve and oversee the implementation of the AI Ethics Plan within the Entity.
- **AI Ethics Roles Assignment –** Designate different roles with regards to AI Ethics.
- **AI Ethics Yearly Report Approval –** Approve the AI Ethics annual report prepared by the Chief Compliance Officer or Compliance Officer (CCO/CO).
- **Issue Resolution –** Take or delegate the necessary actions for resolution of the issues that are raised by CCO/CO.
- **Point of Contact for NDMO –** Act as the first point of contact between the Entity and the NDMO. The Head of the Entity or CDO shall resolve any pending issues around AI Ethics for their respective Entity and escalate them to NDMO whenever necessary.

Chief Compliance Officer (CCO) / Compliance Officer (CO) - The CCO or CO is the strategic lead of the AI Ethics practice. In the Public Entities, CO is positioned under CDO and reports directly to CDO. The **CCO/CO** shall exercise the responsibilities in consultation with relevant officers that cover Data Management, Data Governance, Data Privacy, Open Data, Analytics functions. The responsibilities of the CCO/CO shall include:

- **AI Ethics Strategic Planning -** Oversee the development of the AI Ethics Plan and present it to the head of the Entity. The CCO/CO shall also review the performance of AI Ethics to identify improvement opportunities and feed into the AI Ethics Plan.

- **AI Ethics Supervision –** Review the AI Ethics identification and prioritization activities, monitor AI Ethics KPI's for in-house and third-party systems and ensure maintenance activities are being performed.
- **AI Ethics Compliance –** Ensure compliance of the Entity's AI Ethics activities with national regulations, including but not limited to Data Classification, Data Privacy, and Freedom of Information. Make sure that third-party systems comply with These Principles through contractual guarantees.
- **Consultation with Ethics Office** – CCO/CO shall where applicable consult with the Ethics Office when there are any cases or grievances.
- **Point of Contact for NDMO –** Act as the second point of contact between the Entity and the NDMO (regulator).

Responsible AI Officer (RAIO) - The **RAIO (or AI Ethics Officer)** is the operational lead of Responsible AI within the entity. This role can be appointed by the CDO in the Public Entities. The RAIO should work collaboratively with other officers functioning around Data Management, Data Governance, Data Privacy, Open Data, Analytics. Responsibilities include:

- **AI Ethics Planning –** Develop the AI Ethics Plan, including the AI Ethics prioritization methodology, and set targets and KPIs to be agreed on with the CCO/CO or head of Entity/CDO.
- **AI Ethics Management -** Manage AI Ethics activities within the Entity, in particular:
    - The identification of AI Ethics actions
    - The prioritization of AI Ethics actions
    - The update, maintenance, and quality review of AI Ethics actions
- **AI Ethics Education and Awareness -** Educate and raise awareness across Entity's employees on AI Ethics and support national awareness campaigns in coordination with the CCO/CO.

AI System Developer – The AI System Developer has the following responsibilities:

- Define the requirements and system objectives clearly regarding AI Ethics Principles and Controls.
- Define the potential benefits and harms of the AI systems to the organization as well as end-users.
- Design a robust and reliable AI system that can withstand unknown and unforeseen inputs/events.
- Design an AI system that will not discriminate any groups or individuals in the decision-making process.
- Design an AI system that complies with data protection laws and standards as well the associated provisions on automated decision making.
- Design feedback mechanisms to enable the communication process with end-users. Allow end-users to request access to data, provide for opt-in and opt-out options if the AI system uses personal data.
- Make sure that the AI system is respectful and protects the fundamental human rights and cultural values of the kingdom and nurtures social and environmental benefits.
- Design the AI system whose objectives are parallel with the ethics and code of conduct of the organization.
- Document the whole design process and set standards or follow existing standards for the documentation process to support reproducibility and recovery of the AI system.

- Define roles and responsibilities of the stakeholders. Include necessary sign-off and critical decision points throughout the AI System Lifecycle.
- Design a system which incorporates human oversight and define the humans in-the-loop and out-of-the-loop processes.
- Define KPI's of the system on accuracy, performance, risk, fairness, privacy, and security, etc.
- Create clear specifications that address concerns about the AI Ethics principles during the procurement process if the AI system is third-party product.
- Define sensitive features which their usage in the algorithms may lead to discriminatory outcomes or decisions.
- Not use sensitive features as a data sample during the development and training of AI systems or predictive models.
- Take steps to ensure that data sampling is unbiased and test the bias across the sensitive features.
- Take steps to ensure that the data sample is diversified and representative of all segments of society or targeted segments, is fair and will not result in unfair any discrimination.
- Prepare a detailed record of all data analysis activities, including the data of all processes and procedures done on each data set.
- Conduct a fairness assessment. All decisions which conflict with the fairness principles should be justified with business context, cost-benefit analysis, performance trade-off, etc.
- Build AI systems that are understandable and explainable to end-users.
- Take the necessary actions and adequate steps to verify the accuracy of data interpretation.
- Prove the fairness of important decisions by providing the possibility of checking key factors that lead to taking any decisions that could affect individuals' vital interests.
- Provide a manual intervention methodology that allows individuals to track or object to the stages of important decision-making related to their vital interests
- Prepare a mechanism that includes a set of standards necessary for assessing the reliability of AI systems in predicting and taking future decisions.
- Adopt a comprehensive methodology to test the quality of predictive data-based systems, models and AI algorithms according to standard practices.
- Take steps to ensure the data sample's quality, accuracy, and relevance to the purpose of building predictive models and AI systems.
- Incorporate re-assessment into the AI System Lifecycle and monitor the KPI's on the production system and provide for remediation mechanisms.

AI System Owner – The AI System Owner has the following responsibilities:

- Prepare policies and guidelines to support and enable ethical use of AI in accordance with development and deployment best practices as well as third-party product procurement processes.
- Put in place relevant contractual clauses for third-party service providers to undertake to comply with the requirements of these Principles.
- Adhere to the National Data Governance Policies issued by NDMO and approved by SDAIA's Board of Directors.

- Obtain NDMO's approval, upon coordinating with the regulatory authority, before analyzing data classified as confidential according to the Data Classification Policy.
- Ensure that data analysis is exclusive to restricted and general classification levels, provided that there is a need to process data for analyzing them, with appropriate di-identification in place.
- Take the necessary actions and adequate steps to ensure the quality, accuracy, and validity of data to be analyzed, credibility of their sources, and suitability of their collection methods.
- Provide suitable channels that enable individuals to obtain explanations for important results and decisions that affect their vital interests and object to or request proof of fairness for such decisions.
- Prepare guides on explaining the work mechanism of predictive models or AI algorithms used, data to be analyzed, targeted segments, and the factors influencing important results and decisions.
- Take the necessary steps to ensure the machine does not assume control and that AI systems do not take important decisions on behalf of the concerned people or influence their decisions without their prior consent or their ability to question or object to the results.
- Prepare and document data retention policy and procedures according to the specified purposes and relevant laws and legislations.
- Discard and destroy data safely, including archived data and backups, in accordance with the data disposal policy adopted by the entity and in accordance with the relevant laws and policies.
- Prepare a procedural guide that indicates the steps necessary to assess potential risks and implications of analyzing data using predictive models and AI algorithms to measure the accomplishment of overall goals with the least possible impact on individuals' privacy.
- Prepare a procedural guide to indicate steps to assess the impact of bias in results to ensure that the data set to be analyzed is diversified and representative of all used segments.
- Restrict the use of data analysis results to the purpose for which they were intended which must be in line with the relevant laws, regulations, and policies.
- Ensure compliance to data protection laws.
- Train relevant staff dealing with AI systems.
- Enable end-users to request access to data, object to decisions and request explanation on certain decisions or the working of the AI system.
- Define risk management strategy to mitigate risks related with the execution of the AI Ethics actions.

AI System Assessor – The AI System Assessor has the following responsibilities:

- Review communications channels and interactions with stakeholders to provide disclosure and effective feedback channels.
- Conduct periodical reviews on the AI Ethics process and documentation.
- Continuously review AI Ethics KPI's.
- Publish audit reports on the AI Ethics Assessment of the organization which covers the AI development and deployment process as well as third-party AI product procurement process.

Compliance

NDMO, as the national regulator of the AI Ethics agenda, shall monitor compliance against this AI Ethics Framework with the support of the Regulatory Authorities. NDMO will also award appreciation badges for the low-risk applications and the developed ones inside and outside the kingdom to encourage AI Ethics principles adoption. Badges will be electronically provided through NDMO website where the developers can register to have the badge depending on their level of Compliance. Compliance levels are:

There are four levels of AI ethics badges (bronze, silver, gold, and platinum). The entity wishing to obtain badges for its products must register or update their registration on the site or in the experimental environment (sandbox) and explain the capabilities of the product or how it works, and provide the office with all the supporting documents that shows the compliance level to AI ethics principles based on the Compliance checklist.

The application or product receives the Bronze badge if it is registered on the NDMO website and provides NDMO with the basic product information. The product gets the silver, gold, or platinum badge according to the Compliance level that will be announced on the site.

We encourage public and private entities to review the principles and register on NDMO website and the experimental environment to support them in assessing compliance to their AI ethics principles, and provide the compliance report that approves a certain product comply with AI ethics principles to have the appreciation special badges.

The compliance report should contain:

• The product or entity progress towards compliance with the checklist mentioned in the annexes of this document.

• The internal or external evaluation Results to the AI ethics.

• The product or entity objectives and AI ethics key performance indicators.

• The compliance level to the AI ethics and the badge obtained by the product or entity.

NDMO can assist other authorities in reviewing the annual reports and making recommendations to the Saudi Data and Artificial Intelligence Authority Board of Directors regarding general compliance with AI ethics. NDMO can also assisting entities in resolve disputes related to the publication of the AI ethics report.

Annexure

Annexure A: AI Ethics Tools

Non-Technical Tools

- **Risk Management:** AI risk management is the accounting and addressing of risks posed by developing or using AI systems. The assessment defines and provides thresholds for risk exposure and triggers risk mitigation processes and actions. A four-tiered risk framework recognizes the varying levels of risk posed by AI systems to one's health, safety, and/or fundamental rights and sets out proportionate

requirements and obligations per risk level. In addition, risk management is used to classify the categories and levels of risks associated with the development and/or use of AI technologies. These risk tiers are classified as minimal or no risk, limited risk, high risk and unacceptable risk. AI systems that pose 'minimal or no risk', such as spam filters, will be permitted with no restrictions but AI system providers will be encouraged to adhere to voluntary codes of conduct. AI systems that pose 'minimal or no risk', such as spam filters, will be permitted with no restrictions but AI system vendors will be encouraged to adhere to voluntary codes of conduct. AI systems that pose a 'limited risk', such as chatbots, should be subject to transparency obligations (e.g., technical documentation on function, development, and performance) and may similarly choose to adhere to voluntary codes of conduct. The transparency obligations will serve, amongst others, to allow AI system owners to make informed decisions about how they integrate an AI software into their products and/or services. AI systems that pose a 'high risk' to fundamental rights will be required to undergo pre and post conformity assessments. AI systems that pose an 'unacceptable risk' to the safety, livelihoods, and rights of people, such as for social scoring, exploitation of children, or distortion of human behavior whereby physical or psychological harms are likely to occur, are not allowed. Risk management should be directly interlinked into AI initiatives, so that oversight is concurrent with internal development of AI technology. Risk management of AI systems affects a wide range of risk types including data, algorithm, compliance, operational, legal, reputational, and regulatory risks. Risk management subcomponents, such as model interpretability, bias detection, performance monitoring, are built in so that oversight is constant and consistent with AI development activities. In this approach, standards, testing, and controls are embedded into various stages of the AI System Lifecycle, from design to development and post-deployment.

- **AI Fairness Position Statement:** A fairness position statement allows the owner of the AI technology to clearly state the fairness criteria that has been employed by the AI system and explain the rationale and logic behind it in a direct and non-technical language. To implement a fair AI system in a sustainable way, choosing the right fairness objectives is key to set the tone of the AI model in terms of its ethical standards and regulatory requirements. This is done by sharing the reasons and underlying fairness values expressed throughout the model as well as the decision-making process of the AI model to communicate and reach the wider audience. This document would be made accessible and available to the public and affected individuals and communities.

- **Ethical Impact Assessment:** AI has accelerated innovation in how business is conducted and executed by practitioners, and therefore it is imperative to evolve AI system ethical impact assessments to identify areas that need adjustment and recalibrating to design the AI model into an ethically accepted technology to maximize its positive impact on complimenting human capabilities and skillsets. The objective of impact assessments is to evaluate and analyze the level of ethical impact of the AI technology on individuals or communities in both direct and indirect manners which enables the owner of the AI system to address identified issues and strengthen areas where improvements and adjustments are required. It is also essential to be able to assess the ethical risks that the AI system is projecting, analyze the discriminatory harm impact, accurate representation on the ethical impact of the system through a diversified and multi-stakeholder analysis, as well as act as a facilitator to address whether a model should move to production or deployment. One of the purposes of the ethical

impact assessment is to help build public confidence around the AI system and demonstrate consideration and due diligence to wider the public audience.

- **Privacy and Security Standards:** Privacy and security standard are in place to help companies improve their information security strategy by providing guidelines and best practices based on the company's industry and type of data they maintain. In the following table, some examples are given for privacy and security standards:

No	Standard Name	Link
1	ISO/IEC 27000 Family (International Organization for Standardization)	https://www.iso.org/isoiec-27001-information-security.html
2	ISO 31000 Family (International Organization for Standardization)	https://www.iso.org/iso-31000-risk-management.html
3	NIST Cybersecurity Framework (National Institute of Standards and Technology)	https://www.nist.gov/cyberframework/framework
4	NIST AI Risk Management Framework (National Institute of Standards and Technology) – Work in Progress	https://www.nist.gov/itl/ai-risk-management-framework
5	CIS Controls (Center for Internet Security Controls)	https://www.cisecurity.org/controls/
6	PCI-DSS (Payment Card Industry Data Security Standard)	https://www.pcisecuritystandards.org/pci_security/
7	COBIT (Control Objectives for Information and Related Technologies)	http://www.isaca.org/resources/cobit

Technical Tools

- **Architecture for Trustworthy AI:** The requirements which are stated in AI Ethics Principles and Controls section should be reflected in the design of the AI system architecture. The AI system architecture should set the rules and restrictions across the AI System Lifecycle. The principles and controls are generic and addressing them to the particular use cases or AI systems should be done with the sense-plan-act theoretical approach. Adapting the architecture to AI Ethics entails the integration of the three steps of this approach:
    - **Sense: T**he system should be developed such that it recognizes all environmental elements necessary to ensure adherence to the requirements
    - **Plan:** The system should only consider plans that adhere to the requirements
    - **Act:** The system's actions should be restricted to behaviors that realize the requirements.

  The technical objectives of accuracy, reliability, safety, and robustness must be prioritized to ensure that the AI System functions safely. AI System Developers should build a system that will operate accurately and reliably, in accordance with the intended design, even when confronted with unexpected changes, anomalies and disturbances.

- **Algorithm Assessment:** The objective of the algorithm assessment is to ensure that individuals or communities are informed about the use of algorithms and the weights and counterweights that exist to manage their use. The AI maturity of enterprises and government agencies are different, and this assessment will show the improvement areas for the respective entities or government agencies to improve descriptions on how algorithms inform or impact decision making, particularly in those cases where there is a degree of automatic decision making or where the algorithms support decisions that have a significant impact on individuals or groups. This assessment will balance the human oversight for efficient AI systems.

- **Fairness Assessment:** It is a set of diagnostic methods that helps you compare how fair models and label markers perform for specific groups. It checks whether the model's result is regularly overestimated or underestimated for one or more groups compared to others. Additionally, it evaluates how well the diversity of data is represented for each group. In the following table, fairness assessment tool examples are included

No	Tool Name	Link
1	Google Model Card Toolkit	https://github.com/tensorflow/model-card-toolkit
2	IBM AI Fairness 360	https://aif360.mybluemix.net/
3	Microsoft Fairlearn	https://fairlearn.org/
4	Google What-if Tool	https://pair-code.github.io/what-if-tool/
5	Aequitas Bias and Fairness Audit Toolkit	http://aequitas.dssg.io/
6	Veritas Fairness Assessment Tool	https://github.com/mas-veritas2/veritastool

- **AI Methods Explanation Report:** As explained under Transparency and Explainability section, it should be explained why the system behaves the way it does and how it takes decisions. Although some training methods have superior performance, they work as black-box, and it is a challenge to interpret the results. Small deviations in the data could lead to dramatic deviations and changes on the outcome. The report should explain the behavior of the system as well as ensure the deployment of reliable technology. The report should help to better understand the AI system's underlying mechanism as well as interpretation of the outcomes.

  AI System stakeholders should consider the trade-off between performance, materialization, and explanation methods. In some cases, explanation methods increase complexity or require sacrificing performance. The cost-benefit analysis should be conducted and the level of the explainability should be justified based on this analysis.

- **Algorithm Auditing:** Unexpected algorithmic behaviors can be detected with algorithm audits. In general, algorithm audits are done on an ad-hoc basis, and it is important to standardize algorithm auditing process with supporting AI algorithms. The process should be systematic and continuous. Regulating and auditing AI systems for ethical compliance is more complicated than regulating and auditing human decision making or processes. AI systems should be designed with due consideration of AI Ethics principles and controls. Auditing mechanisms should follow the same principles and controls in alignment with these principles.

- **Safety Self-Assessment:** Safety considerations of accuracy, reliability, security, and robustness should be considered at every step of AI System Lifecycle. AI system safety self-assessments should be continuously logged and documented in a way that allows review and re-assessment. The performance of AI system safety self-assessments should be conducted by relevant stakeholders at each stage of the workflow. They should evaluate how the design and implementation practices line up with the safety objectives of accuracy, reliability, security, and robustness.

- **Data Protection Methods:** These methods help to protect data by applying data transformation methods, especially for the data that is classified as sensitive. Following data protection methods are given as example and they should be after data classification.

    - **Data De-Identification** is the process of eliminating Personally Identifiable Data (PII) from any document or other media, including an individual's Protected Health Information (PHI).

    - **Data Anonymization** is a kind of data sanitization process that intends to protect the privacy of individuals. It is the process of removing PII from data sets to maintain the anonymity of individuals whom the data describe. It is often the preferred method for making structured medical datasets secure for sharing.

    - **Data Masking** is a technique that removes or hides information, replacing it with realistic replacement data or fake information. The objective is to create a version that can't be decoded or reverse engineered. There are a number of ways to change the data, including encryption, character shuffling, and word or character replacement.

    - **Data Pseudonymization** is a way of masking data that ensures it is not possible to attribute personal data to a specific person, without using additional information subject to security measures. It is an integral part of the EU General Data Protection Regulation (GDPR), which has several recitals specifying how and when data should be pseudonymized.

    - **Data Encryption** is a method of data masking, used to protect it from cybercriminals, others with malicious intent, or accidental exposure. The data might be the contents of a database, an email note, an instant message, or a file retained on a computer.

    - **Data Tokenization** is a process of substituting personal data with a random token. Often, a link is maintained between the original information and the token (such as for payment processing on sites). Tokens can be completely random numbers or generated by one-way functions (such as salted hashes).

Annexure B: AI Ethics Tools Mapping to AI System Lifecycle

Tool Type	Tool	Plan & Design	Prepare Input Data	Build & Validate	Deploy & Monitor
Non-Technical	Risk Management	✓	✓	✓	✓
Non-Technical	Fairness Position Statement	✓			

Non-Technical	Ethical Impact Assessment	✓			✓
Non-Technical	Privacy and Security Standards	✓	✓	✓	✓
Technical	Architecture for Trustworthy AI	✓			
Technical	Algorithm Assessment			✓	✓
Technical	Fairness Assessment			✓	
Technical	AI Methods Explanation Report			✓	✓
Technical	Algorithm Auditing			✓	✓
Technical	Safety Self-Assessment	✓	✓	✓	✓
Technical	Data Protection Methods	✓	✓	✓	✓

Annexure C: AI Ethics Checklist

AI System Lifecycle Phase 1: Plan & Design

Number	Checklist Requirements	Binding for Third-party	Principles
PD.1	Did you design the appropriate level of human oversight for the AI system and use case?	Yes	Accountability & Responsibility
PD.2	Does your AI system design prevent overconfidence in or overreliance on the AI system with necessary human intervention mechanism?	Yes	Accountability & Responsibility
PD.3	Did you define human oversight processes with the appropriate KPI's and assign responsibility to the relevant parties?	No	Accountability & Responsibility
PD.4	Did you design an operation and governance strategy to abort or intervene the system when the system doesn't work in an intended way?	No	Accountability & Responsibility
PD.5	Did you consider the liability and Data Subject protection requirements, and take them into account?	Yes	Accountability & Responsibility
PD.6	Did you define thresholds of the KPI's and did you put governance procedures or autonomous actions in place to trigger alternative/fallback plans?	No	Accountability & Responsibility

PD.7	Did you provide training and education to help develop accountability practices?	No	Accountability & Responsibility
PD.8	Did you ensure that the AI Ethics Governance structure is compliant with the proposed governance mechanism in National AI Ethics Policy?	No	Accountability & Responsibility
PD.9	Did you ensure AI Ethics Governance structure includes internal or external audit mechanisms?	No	Accountability & Responsibility
PD.10	Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, covering both input data as well as for the algorithm design?	Yes	Fairness
PD.11	Did you identify sensitive personal data attributes relating to persons or groups which are systematically or historically disadvantaged? If so, the permissible limit that makes the assessment fair or unfair must be determined.	No	Fairness
PD.12	Did you define fairness assessment KPI's?	No	Fairness
PD.13	Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use?	No	Fairness
PD.14	Did you conduct an impact analysis on how the AI system affects the fundamental human rights and cultural values? Did you list any potential negative effects on fundamental human rights and cultural values and the solutions or recovery mechanisms?	Yes	Humanity
PD.15	Did you put measures in place to ensure that the AI system does not lead to people being deceived or unjustifiably impaired in their freedom of choice?	Yes	Humanity
PD.16	Did you align your AI system with relevant standards or policies (for example ISO, IEEE, Data Privacy Law) or widely adopted protocols for daily data management and governance?	Yes	Privacy & Security
PD.17	Did you follow established protocols, processes and procedures to manage and ensure proper data governance? Did you ensure that National Data Management and Personal Data Protection Standards are followed?	Yes	Privacy & Security
PD.18	Did you ensure that data access control meets security, privacy and compliance requirements? Did you design log mechanism for audit and debug purposes?	Yes	Privacy & Security

PD.19	Did you design a Risk Management strategy for your AI system? Did you include risk tiers, KPI's, risk assessment and mitigation procedures?	No	Reliability & Safety
PD.20	Did you assess whether there is a likelihood that the AI system may cause damage or harm to users or third parties? Did you assess the potential damage, impacted audience and severity?	Yes	Reliability & Safety
PD.21	Did you assess whether there is a likelihood that the AI system may unintentionally give wrong results or inaccurate predictions, fail or feed societal biases?	Yes	Reliability & Safety
PD.22	Did you consider the potential impact or safety risk to the environment, living creatures or society, in addition to the Data Subjects?	Yes	Social & Environment Benefits
PD.23	Did you assess what the system's business model is aligned with organization's vision and mission as well as code of conduct?	Yes	Transparency & Explainability
PD.24	Did you design an interpretable AI system where the data, algorithms, outcomes and decisions are transparent and explainable to the related parties?	Yes	Transparency & Explainability
PD.25	Did you design User Experience with human psychology in mind to avoid risk of confusion, confirmation bias or cognitive fatigue?	Yes	Transparency & Explainability
PD.26	Was there any trade off assumption?	Yes	Fairness
PD.27	Have you established a measurement or assessment mechanism for privacy impact?	Yes	Privacy & Security
PD.28	Has the data management approach been reviewed based on human-centric values and according to data regulations within the KSA?	Yes	Humanity

AI System Lifecycle Phase 2: Prepare Input Data

Number	Checklist Requirements	Binding for Third-party	Principles
PID.1	Is there an established mechanism that flags issues related to data privacy or protection in the process of data collection and processing?	Yes	Privacy & Security
PID.2	Has the data been reviewed in terms of scope and categorization?	No	Privacy & Security

PID.3	Has the data been reviewed to check if personal data is evident within the dataset? Is there an established mechanism that allows the AI model to train without or with minimal use of personal or sensitive data?	No	Privacy & Security
PID.4	Is there an established mechanism that controls the usage of personal data (such as valid consent and possibility to revoke, when applicable)?	Yes	Privacy & Security
PID.5	Are there processes to ensure that AI systems are secure and keep information safe, confidential, and private as well as, the integrity of the processed information even under hostile or adversarial conditions?	Yes	Privacy & Security
PID.6	Has the quality and source of the acquired data been assessed through set processes?	No	Privacy & Security
PID.7	Has there been an assessment on whether an analysis can be performed post training and testing the data?	No	Transparency & Explainability
PID.8	Has diversity and inclusion of the dataset at hand been considered or reviewed?	No	Fairness
PID.9	Is there an established mechanism that measures whether the integrity, quality, and accuracy of data collection and its sources have been evaluated and data is up to date?	No	Accountability & Responsibility
PID.10	Has an analysis process been developed for the proxies of the sensitive features?	Yes	Fairness
PID.11	Has the team evaluated the classification, processing, and access to data to ensure that it has been properly acquired?	Yes	Humanity
PID.12	Has the data and AI models been validated to include respect for human rights, values, and cultural preferences in Saudi Arabia?	Yes	Humanity
PID.13	Did you classify the data using NDMO recommendations? If you use other standards, please mention them.	Yes	Social & Environment Benefits
PID.14	Are there appropriate procedures to measure the quality, accuracy, relevance, and credibility of a data sample when dealing with data sets for an AI model?	Yes	Reliability & Safety

AI System Lifecycle Phase 3: Build & Validate

Number	Checklist Requirements	Binding for Third-party	Principles

BV.1	Has the behavior of the system been tested against unexpected situations and environments? Is there a defined fallback plan if the AI model encounters an adversarial attacks or other unexpected situations? Has the fallback plans been tested and confirmed?	Yes	Reliability & Safety
BV.2	Are there defined processes that outline procedures to describe actions to be taken when an AI system fails in different contexts? Have the processes been tested?	Yes	Reliability & Safety
BV.3	Are there defined processes that outline procedures to describe when an AI system fails in different contexts? Have the processes been tested?	Yes	Reliability & Safety
BV.4	Is there an established mechanism of communication to assure the (end-)users of the system's reliability?	Yes	Reliability & Safety
BV.5	Are there clear and understandable definitions on explaining why the outcomes of the AI system took a certain decision?	No	Transparency & Explainability
BV.6	Has the model been built in a simple and interpretable manner?	No	Transparency & Explainability
BV.7	Has an examination of the AI model's interpretability been successfully completed after the model's training?	No	Transparency & Explainability
BV.8	Has there been a research exercise done relating to the use of available technical tools to be able to improve the understanding of the data, model, and its performance?	No	Fairness
BV.9	Are there established processes and quantitative analysis to test and monitor for potential biases and the overall fairness of the system during the development of the system? Are there mechanisms in place to protect any individuals or groups who might be disproportionately affected by negative implications?	No	Fairness
BV.10	Are there any established mechanisms that assess whether the AI system encourages humans to develop attachment and empathy towards the system? Are there are mechanisms that ensure that the AI systems' social interaction is simulated and that is has no capacity for "feelings"?	No	Social & Environment Benefits
BV.11	Have the stakeholders approved the successful tests and validated rounds of user acceptance testing prior to the production of AI models?	Yes	Accountability & Responsibility
BV.12	Did you use any sensitive data/ attributes in the model? If so , justify the use of sensitive personal data attributes or their proxies?	Yes	Fairness

BV.13	Are there AI approaches and algorithms that allow and facilitate decision-making alignment with human rights and KSA's cultural values?	Yes	Humanity

AI System Lifecycle Phase 4: Deploy & Monitor

Number	Checklist Requirements	Binding for Third-party	Principles
DM.1	In case of a chatbot or other communication systems, are the end-users aware that they are interacting with a non-human correspondent?	Yes	Accountability & Responsibility
DM.2	Has the team assessed the AI system's vulnerabilities to potential attacks, revelation of sensitive data or breaking the confidentiality?	Yes	Privacy & Security
DM.3	Are there mechanisms to measure if the system is producing an unacceptable amount of inaccurate predictions?	No	Accountability & Responsibility
DM.4	Is there a set strategy in place to monitor and measure if the AI system is meeting the goals, purposes, and intended applications?	No	Reliability & Safety
DM.5	Are the persons that are accessing the data qualified with the necessary competences to understand the details of data protection requirements?	No	Privacy & Security
DM.6	Are there mechanisms in place to assess the level of influence the AI system may have on end-users' decision making?	No	Transparency & Explainability
DM.7	Is there a process set in place, which is clear and explainable, to inform end-users of the reasons, criteria, and benefits behind the outcomes and results of the AI system? Are there clear steps of communication on how and to whom issues can be raised?	No	Transparency & Explainability
DM.8	Is there a process set in place to collect and consider the end-users' feedback and adoption to the system?	Yes	Transparency & Explainability
DM.9	Are there established processes and quantitative analysis to monitor biases and the overall fairness of the system during the deployment of the system?	Yes	Fairness
DM.10	In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?	No	Fairness
DM.11	Are there established mechanisms to ensure fairness in your AI systems?	No	Fairness

DM.12	Is the information about the AI system accessible to end-users of assistive technologies?	No	Fairness
DM.13	Are there established mechanisms to measure the social and environmental impact of the AI system's deployment and use?	Yes	Social & Environment Benefits
DM.14	Are there established mechanisms to ensure the application of fundamental human rights?	Yes	Accountability & Responsibility
DM.15	Are there established processes for third parties or workers to report potential vulnerabilities, risks or biases in the AI system?	Yes	Accountability & Responsibility
DM.16	Are there established mechanisms to demonstrate your compliance to the principles set out in this document ?	No	Accountability & Responsibility
DM.17	Are there established mechanisms that allow for redress in case of the occurrence of any harm or adverse impact?	Yes	Accountability & Responsibility
DM.18	Are there established mechanisms that provide information to end-users/third parties about opportunities for redress?	Yes	Accountability & Responsibility
DM.19	Are there continuous monitoring techniques to ensure that the AI system maintains privacy and security?	Yes	Privacy & Security
DM.20	Are there periodic assessments of the deployed artificial intelligence systems to ensure the implementation of fundamental human rights and cultural values of the Kingdom?	Yes	Humanity