The Guiding Manual on the Ethics of Artificial Intelligence Use

in Member States of the Gulf Cooperation Council (GCC)

Version 1.0 | November 2023

Contents

Introduction

The term Artificial Intelligence (AI) refers to computer systems and devices that simulate human intelligence by exhibiting behavior that mimics human cognitive abilities and working patterns. These capabilities primarily include analysis, inference, cumulative learning, decision-making, and responding to changing situations that the machine has not been explicitly programmed for. This is the core principle behind the functioning of AI.

AI's ability to simulate and surpass the way humans perceive and interact with the world, supported by various forms of machine learning that recognize data patterns, has produced many positive outcomes. At the same time, it has also introduced significant risks to human societies, ecosystems, human life, and even the human mind.

The use of AI technologies affects many sectors and areas of life, including the economy, labor, education, culture, health, and social well-being. It is widely acknowledged that AI can offer substantial opportunities to improve quality of life across many fields. For example, AI has helped facilitate education and make knowledge accessible to various segments of society, especially during the pandemic, which would have been extremely difficult to navigate without the use of AI technologies.

However, several issues and challenges must be considered when working with AI, as they may compromise the protection of human rights and personal data. For instance, the use of AI could threaten the right to equality and non-discrimination, which are among the most fundamental pillars of human rights. In addition, AI technologies may either pose a threat to or strengthen individual rights, such as the right to privacy. When used appropriately, these technologies can serve as safeguards for enjoying fundamental rights and personal and political freedoms. Therefore, AI should be treated with caution as a double-edged sword.

As such, the risks and threats arising from the growing use of AI, which are expected to increase in the coming years, may impact governments, economies, and societies at the local, regional, and international levels. This highlights the urgent need for national and international responses to minimize or mitigate these risks and address their consequences.

Based on this understanding, the Guiding Manual on the Ethics of Artificial Intelligence Use was developed to support innovation while protecting the vital interests and national security of GCC states from potential risks associated with AI and other technologies. This helps ensure optimal benefit from technological advancements and enables GCC countries to pursue global leadership across all sectors.

Trustworthy AI was the foundational goal in creating this document. It is essential to build AI systems that earn public trust, as people will only fully benefit from these technologies when

both the systems and the individuals behind them are reliable. This document was developed with reference to the Recommendation on the Ethics of Artificial Intelligence issued by UNESCO in 2021.

Values and principles

The objective of outlining values and guiding principles is to advance artificial intelligence technology across all sectors in the GCC, whether technical, social, or political. While good intentions may guide the use of AI, they do not eliminate the possibility of unintended harm caused by AI systems. Therefore, GCC countries are encouraged to use this document as a reference to achieve trustworthy AI that aligns with the United Nations' recommendations on AI ethics.

These values and guiding principles establish a general framework for the development of AI systems focused on protecting human dignity and rights, preserving the environment, and promoting coexistence among all living beings. They are grounded in a commitment to using this technology in service of humanity and the public interest, with the aim of improving human well-being and freedom. Although AI systems provide significant opportunities, they also present risks that must be managed appropriately and proportionately.

This is the approach that GCC countries are encouraged to adopt in order to become a leading region in advanced and ethical technology. Safe AI will benefit the citizens of the region in ways that are consistent with fundamental values, including respect for Islamic Sharia, human rights, and the rule of law.

Through this document, the GCC aims to increase investment opportunities for AI system developers and provide them with a competitive advantage by integrating AI into products and services that serve humanity. It also aims to build user trust in AI products, since the absence of a clear and comprehensive ethical framework for the safe application of this technology limits public and societal confidence in its development and use.

Therefore, the values and principles outlined in this document should be respected and promoted when necessary through the development of new legislation, regulations, and guidelines, or by amending existing ones in accordance with the local laws and regulations of member states.

Values

Value One: Respecting, protecting, and promoting human dignity, freedom, and autonomy

Human dignity is based on the idea that every individual possesses intrinsic value, which should never be diminished or violated by others, including through the use of new technologies such as AI systems. In this context, respecting human dignity means treating all people with the respect they deserve as moral beings, rather than as objects to be sorted, recorded, conditioned, or manipulated. Therefore, AI systems should be developed in a way that respects, serves, and protects the physical and mental well-being of individuals, including their personal and cultural sense of identity, and they should also meet their basic needs, and enable those interacting with AI to maintain full and effective self-determination, including the ability to choose and define their own preferences.

Value Two: Respect for Islamic Sharia, the Constitution, and strengthening Gulf unity

AI systems should uphold and promote the principles of Islamic Sharia, national constitutions and legislation, prevailing laws, and the applicable local standards and procedures in GCC countries. They should also respect the diversity of values and lifestyles of individuals and not undermine these principles or limit their effectiveness. Where appropriate, AI systems should support cooperation and integration among GCC states. They must also commit to avoiding any operation that would undermine the fundamental obligations underlying the rule of law, mandatory regulations, legal procedures, or equality before the law for all members of society.

Value Three: Environmental protection and promotion of sustainability

Digital transformation initiatives face major challenges, including their environmental impact on human and other ecosystems. GCC countries should therefore consider the environmental risks resulting from the operation and training of AI systems using electricity, such as carbon emissions or the use of fresh water for cooling. Measures should be taken to ensure that ecosystems are not degraded and that their sustainability is maintained for future generations.

Value Four: Promoting peaceful use for the well-being of GCC citizens

Human well-being means that AI systems should improve the quality of life for both individuals and communities. When provided with massive data input, these systems can perform repetitive tasks with high accuracy, free from fatigue or boredom. Entities responsible for implementing AI systems must ensure that such systems are not used in

ways that threaten human coexistence, security, or safety within GCC societies. AI systems must not cause or worsen harm or negatively affect humans. This means that AI systems and their operating environments must be safe, secure, technically robust, and protected against misuse. This document also recommends giving special attention to vulnerable groups such as people with disabilities, the elderly, and others, and ensuring that their needs are incorporated in the development of AI systems.

Principles

Principle One: Human decision-making

AI systems should support human autonomy and decision-making, as outlined in the values previously mentioned. This requires that AI systems act as enablers of a democratic, prosperous, and just society by supporting individuals and communities, strengthening fundamental rights, and allowing for human oversight of AI systems.

Like any technology, AI can both enable and restrict fundamental rights. For example, it can benefit individuals by helping them track their personal data, improving access to education, speeding up disease detection, or providing appropriate treatments, thus supporting the right to education and healthcare. However, due to the vast reach and analytical power of AI systems, they can also negatively impact fundamental rights.

In cases where such risks are present, a fundamental rights impact assessment should be conducted. This should take place before the system is developed and include an evaluation of the level of harm and risk to all segments of society to ensure human safety and security before implementation begins.

Accordingly, AI systems should empower individuals to make better and more informed choices aligned with their personal goals. These systems must avoid shaping or influencing human behavior through mechanisms that are difficult to detect, such as exploiting unconscious processes, unfair manipulation, deception, or conditioning. All of these can undermine individual autonomy.

The autonomy of users must be the primary principle guiding system functions. A key element of this is granting users the right not to be subject to decisions made solely by automated processing if such decisions could result in legal effects or significantly affect them.

Recommendations and proposals for implementing this principle
• **Activating governance and human oversight:** This document recommends activating human oversight to ensure that AI systems do not undermine human autonomy or cause other negative effects. Oversight can be implemented through governance mechanisms, which may include decisions not to deploy the AI system in specific situations, setting levels of human discretion during system use, or ensuring the ability to override decisions made by the system. In addition, it is important to confirm that public implementing bodies have the capacity to exercise oversight in line with their legal authority. Supervision mechanisms of varying degrees may be required to support safety and control measures, depending on the AI system's field of application and the potential risks involved. The less human oversight is possible over an AI system, the greater the need for comprehensive evaluation tests and stricter governance.

Principle Two: Safety and harm prevention

One of the defining features of artificial intelligence is its technical ability to generate decisions and control outputs. This requires that AI systems be developed within a preventive framework that addresses risks and ensures they operate reliably as intended, while minimizing unintended or unexpected harm and preventing unacceptable damage. This also applies to potential changes resulting from the operation of AI systems in environments that may interact with them in ways that threaten the physical or mental safety of humans and other living beings.

AI systems share characteristics with all software systems—they may contain vulnerabilities that can be exploited by hackers. Attacks may target data, the AI model itself, or its infrastructure. These attacks can alter the behavior of the system, causing it to make unexpected decisions or shut down in ways that jeopardize the safety of humans and other creatures.

Therefore, this document recommends providing adequate security protection for the data used in AI systems, including training and operational models. It also emphasizes the importance of protecting the system from extremist or terrorist groups that may exploit it for harmful purposes. Steps must be taken to prevent such misuse and to mitigate any resulting damage.

Recommendations and proposals for implementing this principle
• **Establishing a rollback mechanism:** This document recommends including a rollback mechanism in AI systems to address high-risk harms that may threaten humans, living beings, or the environment. This includes mitigating unintended consequences and errors. It is also necessary to develop processes for identifying and assessing potential risks associated with AI usage across different application

areas, and to define a rollback system as a safety measure to prevent ongoing harm. When development processes or the system itself are expected to pose particularly high risks, rollback and safety procedures must be developed and tested proactively.

- **Assessing the accuracy of predictions, recommendations, and decisions:** This document emphasizes the importance of maintaining a high level of accuracy in AI systems' ability to make correct judgments or to generate accurate predictions, recommendations, or decisions based on available data or models. A clear and well-designed evaluation process can help mitigate, reduce, or correct unintended risks caused by inaccurate outputs. In cases where inaccurate predictions cannot be avoided, estimating the potential error rate is essential to assist decision-makers in determining whether to deploy or withhold the system, especially in situations where AI directly affects human lives.

- **Testing and verifying reliability:** The document stresses the need to verify the reliability of AI systems and their ability to consistently reproduce results. An AI system is considered reliable if it functions correctly with a variety of inputs and in different scenarios. This is critical during the performance evaluation stage to prevent unintended harm. Developers must ensure, during testing, that the system's behavior does not become erratic when the same inputs are repeated under identical conditions.

Principle Three: Justice, fairness, and non-discrimination

All stakeholders should be involved throughout the lifecycle of AI systems and empowered in their development. In addition, individuals and communities must have equal and fair access to the benefits of AI applications. This means ensuring equal access through inclusive design processes and equal treatment for all, which is closely linked to the principle of justice.

Member states should adopt the principle of universal accessibility during the design stages of AI systems in a way that enables all individuals to use AI products or services, regardless of their age, gender, abilities, or characteristics. Access to this technology should also be made available to persons with disabilities, the elderly, and other vulnerable groups.

Recommendations and proposals for implementing this principle

- **Auditing datasets used for system training:** This document recommends ensuring that datasets used to train and operate AI systems are free from unintended biases that may affect specific groups and result in direct prejudice or discrimination against individuals or communities. Such biases could lead to reinforced marginalization in system outputs. Harm can also arise from the

deliberate exploitation of biases or through unfair competition. Bias and discrimination are not limited to data deficiencies but can also stem from the way AI systems are developed, including algorithm design.

- **Establishing oversight processes to ensure fairness and non-discrimination:** The document calls for implementing oversight processes to clearly and transparently analyze, assess, and manage the purpose, limitations, requirements, and decisions of AI systems before deployment. In addition, employing individuals from diverse backgrounds, cultures, and specializations should be encouraged to ensure a variety of perspectives are considered.

- **Addressing the digital divide:** This document emphasizes the need to tackle digital and knowledge gaps both within and between countries throughout the lifecycle of any AI system. This includes access to technology, data, and the quality of that access, in line with national legal frameworks. It also involves bridging gaps in knowledge, skills, connectivity, and community participation to ensure that all individuals are treated fairly.

- **Engaging stakeholders:** The document stresses the importance of early engagement and continuous consultation with stakeholders throughout the AI system lifecycle, as they are the most likely to be directly or indirectly affected by these systems. Gathering regular feedback from stakeholders even after deployment is also considered valuable to support AI implementation and operation.

- **Ensuring diversity in access to AI systems:** The document recommends offering AI systems through diverse methods and channels, rather than relying on limited technologies or platforms. Design principles should aim to accommodate the widest possible range of users, following recognized accessibility and universal access standards. This will help ensure fair access and active participation for all individuals.

Principle Four: Privacy and data protection

Privacy is closely linked to the principle of harm prevention and is considered a fundamental right that is particularly affected by AI systems. GCC countries should work to prevent harm to privacy through proper governance of data usage. This includes ensuring the quality, integrity, and relevance of the data used in relation to the specific field in which AI systems are deployed, along with access protocols and the ability to process data in ways that safeguard privacy.

Recommendations and proposals for implementing this principle

- **Regulating the use of digital records:** This document recommends regulating the use of digital records that allow AI systems to analyze human behavior in order to infer preferences or tendencies, such as age, gender, or religion. These records must not be used to support illegal or unfair discrimination. AI systems should ensure privacy and data protection throughout the entire system lifecycle, including both user-submitted information and data generated over time during user interactions with the system.

- **Restricting unrestricted data use:** The document advises following best practices in data handling and applying appropriate encryption methods. It is essential to limit the unrestricted use of data, especially personal data, within AI systems. The purpose of data use, expected outcomes, and risk assessment must be clearly defined for each system.

- **Restricting unrestricted data access:** The document recommends implementing data governance principles to ensure data security and accountability for violations of privacy. Access to personal data used in AI systems should be restricted. Protocols must define who has access to the data and under what conditions. Only qualified and authorized personnel with a legitimate need should be granted access to personal data, with proper safeguards in place. Audit trails for data access and usage must be effectively implemented and monitored.

- **Monitoring data quality and integrity:** The document emphasizes the need to monitor the quality and integrity of data used in AI systems and to conduct adequate testing of data sets. This is critical to system performance. An internal automated system for monitoring and alerts should be developed, incorporating human oversight during data collection. This is necessary because the data may contain social biases, inaccuracies, errors, or harmful content that could affect AI behavior during training or deployment, especially in self-learning systems. Tests must be conducted using clear standards, and documentation must be maintained at every stage including planning, training, testing, and deployment. These standards should also apply to AI systems sourced from external entities. Feedback mechanisms must be in place to ensure that AI technologies cannot be manipulated by external parties.

- **Establishing suitable frameworks and mechanisms for data protection:** The document recommends establishing data protection frameworks and mechanisms based on national principles and standards for handling personal data. This includes procedures for collection, use, processing, disclosure, and the rights of data subjects. These frameworks must ensure that any use of personal data is based on a legitimate purpose and sound legal grounds, including obtaining informed consent from the individuals concerned. These frameworks should be

safeguarded by judicial systems and applied consistently throughout the AI system lifecycle. A practical example is granting users of metaverse technologies, or their representatives, full control and management rights over their data, in accordance with personal data protection laws and policies.

Principle Five: Transparency, explainability, and interpretability

Efforts must be made to enhance the transparency of AI systems and strengthen their explainability throughout their entire lifecycle. However, the level of transparency and explainability should always be appropriate to the context and consequences. Explainability is essential for building and maintaining users' trust in AI systems.

People must be fully informed of any decision made based on information derived from AI systems and algorithms, especially when such decisions affect their safety or human rights. In such cases, individuals have the right to request clear explanations from relevant AI actors or public institutions regarding the reasons behind decisions that affect their rights and freedoms.

This can be supported through mechanisms that allow appeals or objections to be submitted to a designated official responsible for reviewing and, if necessary, correcting the decisions made. Decision-making processes should not function as “black box” algorithms in these circumstances.

AI actors must also clearly and promptly inform users when products or services are provided directly or through the use of AI systems. It should be noted that achieving transparency and explainability may require balancing other principles such as the right to privacy and the need for safety and security.

Recommendations and proposals for implementing this principle

- **Developing traceability of decision-making:** This document recommends enhancing traceability by documenting datasets and the processes that lead to decisions made by AI systems, including data collection, classification, and the algorithms used. This should follow the best available standards to support traceability and increase transparency. This also applies to the decisions made by the AI system. The mechanism should allow for identifying the reasons behind an incorrect AI decision, which can help prevent similar errors in the future.

- **Developing explainable AI systems:** The document calls for developing AI systems that are as explainable as possible. Explainability refers to the ability to explain both the technical processes of the AI system and the associated human decisions, such as the system's application areas. Explanations should be

provided in a timely manner and tailored to the knowledge levels of relevant stakeholders, whether they are general users, regulators, or researchers.

- **Informing users of AI interaction:** The document recommends that users be clearly informed when they are interacting with an AI system, and the system itself should be identifiable as such. In addition, users should be given the option to choose human interaction instead of AI-based interaction when necessary to ensure compliance with fundamental rights. AI practitioners or end users should also be informed about the system's capabilities and limitations in a way that is relevant to the current use case. This could include information on the system's accuracy level and any potential risks associated with its use.

Principle Six: Responsibility, accountability, and awareness

The conditions of responsibility and accountability are closely tied to the principle of justice. This requires establishing mechanisms to ensure accountability for the performance of AI systems and enabling responsibility for their outcomes before, during, and after development, deployment, and use.

It also involves emphasizing individuals' awareness and ethical responsibility for their actions and behaviors within virtual environments, especially given the current lack of clear legal frameworks governing such spaces. Efforts should be made to promote public understanding and awareness of AI technologies and ethics through education, awareness campaigns, and training programs, taking into account the linguistic, social, and cultural diversity of the local community.

Recommendations and proposals for implementing this principle

- **Developing legislation to ensure legal accountability:** This document recommends enacting legislation that clearly establishes that responsibility for AI system performance lies with identifiable entities that can be held legally accountable. Responsibility should not be attributed to the system itself, but rather clearly assigned to the parties involved in its development and throughout its lifecycle.

- **Enacting laws to enable auditing and evaluation:** The document calls for the development of legislation and procedures that ensure technical auditing of system performance, including the ability to assess algorithms, data, and design processes. AI systems should always be accessible for authorized evaluations by both internal and independent external auditors, with assessment reports made available to help improve the credibility of the technology.

- **Developing mechanisms to report ai system outputs:** The document recommends establishing mechanisms for reporting actions or decisions generated by AI systems. There must be guarantees of appropriate responses to these reports in a way that satisfies users. Reports should be documented and analyzed by specialized entities to assess and reduce potential negative impacts of AI systems, while providing proper protection to whistleblowers who report system performance in good faith.

- **Raising awareness and disseminating knowledge:** The document emphasizes the importance of increasing public awareness of AI systems by creating educational materials for all age groups. These materials should be made openly accessible to promote understanding of AI technologies, their impact on society, the pros and cons of their usage, and the importance of data throughout the AI lifecycle. Public awareness campaigns should be expanded with sensitivity to linguistic, social, and cultural diversity to ensure the broadest possible benefit and to help individuals make informed decisions about using AI systems. Joint training sessions among GCC countries can also be organized to promote knowledge-sharing and a deeper understanding of how AI systems operate.

Principle Seven: Integrity and non-falsification

The principle of integrity should be upheld in the use of artificial intelligence. This includes a commitment to avoiding the falsification of facts and refraining from exaggerating, inflating, or distorting the capabilities of AI for secondary purposes such as profit, gaining a competitive edge, or influencing public opinion. Such practices can negatively impact the primary beneficiaries. The core interests of individuals and key users must always be prioritized.

Recommendations and proposals for implementing this principle

- **Establishing mechanisms to verify compliance with defined standards for AI development or deployment:** This document recommends creating clear mechanisms that allow for verifying the application of specific, well-defined standards and principles when making decisions regarding the development or deployment of AI systems. These standards should be clearly documented, transparent, and accessible for any authorized evaluation carried out by internal or external auditors.

Use, benefit, and application of this guide

Member states and all relevant entities should respect, promote, disseminate, and protect the ethical values, principles, and standards outlined in this document concerning artificial

intelligence. They should also take all possible measures to implement the provisions of this guide.

Member states are encouraged to expand their efforts in relation to this guide and complement those efforts through cooperation with all relevant national and international organizations, both governmental and non-governmental, as well as with transnational companies and scientific organizations whose activities fall within the scope and objectives of this guide.

Establishing a methodology for assessing ethical consequences and creating national committees for AI ethics are two important tools for achieving these goals.

Promoting AI ethics

The Gulf Cooperation Council (GCC) will promote this guide by publishing it on the official website of the GCC General Secretariat and the GCC eGovernment portal. It will also be shared through the official eGovernment portals and social media platforms of the member states, as deemed appropriate by each state.

This document should be taken into account when developing new policies or legislation, or when preparing national strategies related to eGovernment, digital transformation, the digital economy, and information security, including associated programs, initiatives, and projects.

Final provisions

This document is a single, indivisible whole, and the core values and principles it contains should be regarded as interconnected.

No provision of this document may be interpreted in a way that replaces or overrides the duties or rights of states, nor in a manner that alters or undermines those duties and rights in any other way.