THE UNITED REPUBLIC OF TANZANIA

MINISTRY OF COMMUNICATION AND INFORMATION TECHNOLOGY

GUIDELINES FOR THE SECURE AND ETHICAL USE OF ARTIFICAL INTELLIGENCE IN TANZANIA

(Draft Version One)

GUIDELINES FOR THE SECURE AND ETHICAL USE OF ARTIFICAL INTELLIGENCE IN TANZANIA

Version:	2.0
Date Release:	June, 2024
Contact:	ps@mawasiliano.go.tz
Ministry:	Ministry of Information, Communication & Information Technology

Foreword by the Permanent Secretary

In recent years, the world has witnessed high growth and use of 4th industrial revolution digital technologies in all social and economic sectors. The technologies include Artificial Intelligence (AI), Distributed Ledger Technologies or Block chain, cloud computing, Big Data Analytics, Internet of Things, virtual and augmented reality, etc. Not to be left behind, the Government of Tanzania has declared that the ICT driven fourth industrial revolution is inevitable and has increased pace in adoption of these digital technologies. This is in line with the development of the ten (10) years Tanzania Digital Economy Strategic Framework (TzDESF) which acknowledges that ICTs are central to a competitive social and economic transformation of the country. Furthermore, both the Tanzania Development Vision 2025 and 3rd National Five Year Development Plan 2021/22- 2025/26 acknowledge that ICTs are central to a competitive social and economic transformation and also they are a major driving force for the realization of the Vision.

Nevertheless, the GoT understands that the 4th industrial revolution technologies, especially Artificial Intelligence (AI), brings both positive and negative impact to an economy of a country. On the one hand, AI offers an exciting extension of many human capabilities such as observation, processing, and decision-making. The output and outcomes of AI systems are nearly instantaneous, offering humans powerful efficiencies that did not exist just a few years ago. The computing power and systems used for AI technologies far exceed human cognitive capabilities, allow for constant “machine learning” without human supervision, and include consideration of patterns that are typically impossible for humans to discern (e.g., the ability to identify an individual based on their gait without ever seeing their face). On the other hand, AI has inherent ethical risks. Ethical problems in AI can lead to a variety of consequences with different levels of severity; these consequences include the profiling and biases of algorithms against a particular race, gender, or specific category of people which can affect how education, healthcare, financial, and privacy systems work. AI can be also used maliciously, in any of these fields to fake data, steal passwords, and interfere with the work of other software and machines, thus undermining public trust in technology even more. These digital crimes put core human values such as personal privacy, data protection, fairness, and autonomy at

risk. Lack of AI policies and regulatory AI framework on the emerging technologies is of concern to the GoT.

As a starting point, the Got through the Ministry of Communication and Information Technology (MICIT) is developing guidelines for the development and use of ethical AI in the country. The purpose of this AI Guideline is to provide guidance which can assist all stakeholders, both public and private, in the development, deployment, and adoption of AI-based solutions for various social and economic activities. The guidelines have been formulated after extensive stakeholders in the AI eco-system.

Mohammed Khamis Abdulla
Permanent Secretary
Ministry of Communication and Information Technology

INTRODUCTION

Artificial intelligence is one of the pillars of the Fourth Industrial Revolution. There is a sudden development and spread of the use of artificial intelligence in areas such as healthcare, finance, education, energy, natural language processing, speech technologies, computer vision, etc. The field of AI has brought about extensive social and economic benefits to the world and competitive advantages to various economic sectors. Artificial Intelligence has become increasingly popular in different industries, from producing different goods in factories to healthcare and finance. AI has been making remarkable changes to how we live and work, transforming industries and creating new possibilities. Machine Learning, a subset of AI, is becoming increasingly popular for its ability to analyze and process large data sets. This technology is used in systems like Cybersecurity, Cloud Computing, Full-Stack Development making it essential for businesses to adopt AI to stay competitive.

The Government of Tanzania (GoT) is spearheading technological innovation and digital transformation and it has recognised the significant role digital technologies play in in the economic development of the country. Artificial intelligence (AI) has the potential to further improve the resilience in the social and economic activities of the nation though the use of AI in Tanzania is still in a nascent stage. Tanzania's push towards AI technology is gaining momentum and aims to benefit the country in the areas of development of AI infrastructures, healthcare, digital economy, environmental conservation, education, justice, agriculture, among others. This is evidenced by the establishment of a Multidisciplinary AI4D Lab at the University of Dodoma, the use of machine learning and mid infrared spectroscopy for rapid assessment of blood-feeding histories and parasite infection rates in field-collected malaria mosquitoes and also the enhanced microbiological testing capacity for COVID-19 and other infectious diseases in Tanzania and Zanzibar, both by Ifakara Health Institute. Also there are a number of Tanzanian startups that use AI in their applications, like Keystride, Elsa Health, Afya Intelligence, Mipango, Agripoa, AfyaTest, Tanzania AI Lab, Shule Direct, to name a few. Recently, the Tanzania judiciary has incorporated

AI in its new transcriptions and translations system that aims to improve court efficiency.

These developments have generated much discussion on the role of AI in various sectors of the economy. Many questions have been raised about risks related to security, privacy, and ethical considerations. As this field is quickly evolving, guidance is needed to help understand and evaluate these risks. At a recent stakeholder AI stakeholder's meeting hosted by the Ministry of Health (MoH), the main recommendation was for the GoT to lead the development of a shared vision for AI in Tanzania. This would establish the government as a leader and ensure the commitment of other stakeholders to the shared vision. The shared vision would also provide opportunities to orient sector leaders and technical teams, development and implementing partners, training institutions, local firms, and other stakeholders to be critical in promoting the longstanding, deeply rooted, and well-developed professional ethics in the community.

Despite Tanzania not having specific legislation for Artificial Intelligence yet, its misuse and development can lead to penalties, such as data breaches based on the Data Protection Act of 2022. Nevertheless, there still is a growing concern about the ethical implications, as AI plays an increasingly decisive role in people's daily lives. In light of this, the Ministry of Communication and Information Technology (MICIT) is developing this guideline document for AI ethical development and uses. The aim is to ensure AI solutions align with all relevant ethical obligations, during its design, development, and use.

IMPORTANCE OF AI ETHICS

Ethics is the science of proper behaviour. It can be claimed that ethics is the basis for creating an ideal model of human interrelations, ensuring optimal communication between people and a reference point for creating a structure of moral consciousness. The practice of AI ethics is the consideration of moral problems related to the interaction of technology, humans, and society. Just as AI is evolving rapidly, so too are AI ethics considerations. The goal of AI ethics is to ensure that AI is developed and used in ways that are fair, transparent, accountable, and beneficial

for society. AI ethics addresses a wide range of issues, including data privacy, bias, discrimination, transparency, accountability and social impact.

While AI can bring many benefits for human beings, it also comes with ethical considerations that cannot be ignored. As these systems are increasingly being used across sectors, AI can have significant effects on credit, employment, education, competition, and more. However, without ethics embedded in AI algorithms, it is hardly possible to guarantee that AI will not enable certain persons to cause more harm than good. With the wider adoption of AI in recent years, it can be used to sow distrust in public information and perpetuate discrimination in the delivery of services and unfavourably profiling segments of the population, raising many other moral concerns.

RATIONALE FOR ADOPTING THE ETHICAL AI GUIDELINES

Ethical AI is crucial as it ensures that the technology used aligns with national values and ethical standards. Implementing ethical AI also helps to ensure that AI is used for the benefit of society. In the meeting of stakeholders utilising AI in Health hosted by the MoH[1], there were challenges that were identified in the use of AI, which included:

- Poorly defined governance structures that bring AI stakeholders together for knowledge sharing and discussion of AI implementation in Tanzania;
- Lack of specifications and an approach to implementing AI in the health sector;
- Limited capabilities and skills to develop appropriate solutions for health sector needs;
- Limited capacity to implement AI tools;
- Lack of understanding among local technology firms on the market opportunities;
- Limited research and evidence generation for AI in Tanzania's health sector.
- Availability of relevant, reliable, and quality data;
- Usability of available data (how clean/accurate is the data collected);
- Limited opportunity and local environment for knowledge and skills development;
- Misconceptions and lack of awareness about AI technology in general.

[1] *Policy Framework for Artificial Intelligence in Tanzania Health Sector, Tanzania Ministry of Health, February 2022*

The use of generative AIs such as ChatGPT, Copilot, Midjourney, and Stability AI, to name a few, are mired in controversy. This controversy revolves around the issue of the images and texts that the algorithms are trained on, as many artists and content creators have filed cases in different courts. This may lead to algorithm disgorgement.

In the realm of education, the use of AI systems can potentially enhance teaching, learning and assessment, provide better learning outcomes and help schools to operate more efficiently. However, if those same AI applications are not properly designed or used carelessly, this could lead to harmful consequences. Educators need to be aware and ask questions whether AI systems they are using are reliable, fair, safe and trustworthy and that the management of educational data is secure, protects the privacy of individuals and is used for the common good. “Ethical AI” is used to indicate the development, deployment and use of AI that ensures compliance with ethical norms, ethical principles and related core values.

Lastly, many Tanzanian entities/institutions are not cognizant of the following issues:

- Standard security, privacy, and compliance provisions are not in place when using this technology.
- AI tools can generate incomplete, incorrect, or biased responses, so any output should be closely reviewed and verified by a human.
- AI-generated code should not be used for institutional ICT systems and services unless it is reviewed by a human.

This guideline document intends to outline key aspects that are to be considered, including stakeholders, principles, and recommendations to guide the implementation and use of AI in the country. This document lays the foundation for the mechanisms to develop and put in place AI systems that ensure their responsible development meeting the highest ethical and security standards. Lastly, the development of artificial intelligence systems requires that they comply with the relevant standards throughout their lifecycle, and on the basis of which it will be possible to characterise these systems as reliable and accountable.

PURPOSE AND OBJECTIVES

The main purpose of adopting these guidelines is to prevent processes involving artificial intelligence systems from endangering or marginalising humans. This is primarily about creating ecosystems that will use artificial intelligence to increase human productivity, optimise the use of resources for work and the functioning of society in general, and improve the quality of human life in addition to business productivity. The strategy is an approach for the GoT integrating Artificial Intelligence across economic sectors, enhancing societal welfare, and fostering innovation. The aim is to build human capital, digital infrastructure, and a supportive ecosystem for AI, ensuring ethical governance and promoting partnerships for growth. The goal is to propel Tanzania into a leadership role in AI within the context of a broader continental ambition for digital transformation and economic resilience. The following objectives are to be considered:

i) **Develop Human Capital:** Elevate AI literacy and specialized skills, preparing the workforce for the future digital economy. Focus on inclusive education and training initiatives to ensure widespread AI competency. This involves establishing national skills-building programs focusing on AI and data literacy, complemented by young professionals/apprenticeship programs to nurture local AI talent.

ii) **Strengthen Digital Infrastructure:** Prioritize the development of digital infrastructure as the backbone for AI applications. Enhance data ecosystems to support AI research and innovation. This includes the establishment of Data & AI Sandboxes for testing and development. Positioning Tanzania as a regional hub for cloud infrastructure with AI-ready storage and compute capacity

iii) **Foster Innovation and R&D:** Encourage innovation in AI through supportive policies, incentives, and investment in research and development. Collaborate with private sector and academia to drive technological advancements. Set the foundations for world-class AI university education and applied research by initiating public-private funded programs

iv) **Ensure Ethical AI Use:** Implement national guidelines for ethical AI use, emphasizing transparency, accountability, and public trust. Safeguard data privacy and promote responsible AI practices.

v) **Cultivate Partnerships:** Leverage national and international collaborations to advance AI knowledge, share technology, and attract investment, enhancing Tanzania's position in the global AI landscape.

THE AI PRINCIPLES

The potential benefits associated with responsible use of AI are significant, any decision on the use of AI must consider both the potential positive and negative impacts. In light of that, the MICIT is proposing the following principles in the development and use of AI:

SNo.	Principle	Description
1	Valid and Reliable	The use of AI systems, including the specific AI method(s) employed, should be justified, appropriate in the context and not exceed what is necessary and proportionate to achieve legitimate aims of producing accurate results within expected timeframes. AI systems should perform reliably and demonstrate that systems are designed to operate within a clear set of parameters and that can be verified they are behaving as intended under actual operating conditions.
2	Safe	AI should produce results that conform to safety expectations for the environment in which the AI is used (e.g., healthcare, justice, education, transportation, etc.) AI systems should not be used in ways that cause or exacerbate harm, whether individual or collective, and including harm to social, cultural, economic, natural, and political environments.
3	Managed bias	Bias can manifest in many ways; standards and expectations for bias minimization should be defined prior to using AI. AI systems should not lead to individuals being deceived or unjustifiably impaired in their human rights and fundamental freedoms. Only people can see the blind spots and biases in AI systems, so they must be taught how to spot and correct any unintended behaviors that may surface.
4	Resilient	Resilience is the degree to which the AI can withstand and recover from cyber-attacks. AI technologies should continuously be assessed and appropriate mitigation and/or prevention measures should be taken to address adverse impacts, including on future generations.
5	Transparency	Users should know where and when AI systems are being used and understand what they do and how they do it. When AI systems are used to (help) make decisions impacting people's careers and lives, those affected (including those making the decisions) should understand how those decisions are made and exactly how AI influences them. The information and reasons for a

		decision should be presented in a manner that is understandable to the user.
6	Accountability	Those who design and deploy AI systems must be accountable for how those systems operate. Clear owners should be identified for all AI instantiations, the processes they support, the results they produce, and the impact of those processes and results on employees. It is the shared responsibility of the developers and their associates of the AI, as well as those who have chosen to implement AI for a particular purpose.
7	Explainable and Interpretable.	The ability to explain how an output was generated, and how to understand the meaning of the output. Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Individuals should be meaningfully informed when a decision which may or will impact their rights, fundamental freedoms, entitlements, services or benefits, is informed by or made based on AI algorithms and have access to the reasons for a decision and the logic involved. The information and reasons for a decision should be presented in a manner that is understandable to them.
8	Privacy-enhanced.	Privacy of individuals and their rights as data subjects must be respected, protected and promoted throughout the lifecycle of AI systems. When considering the use of AI systems, adequate data protection frameworks and data governance mechanisms should be established or enhanced in line with the Data Protection Act of 2022 to ensure the integrity of the data used.

AI CODE OF ETHICS

It is recommended that an entity/institution should have a code of ethics when developing AI. The Code of ethics that is recommended is for the AI system to:

a) Always be aligned with the entity/institution mission and values;
b) Strive for excellence and quality;
c) Adhere to ethical principles;
d) Assist entity/institution in unleashing human potential easily and affordably;
e) Assist entity/institution in unleashing human potential by generating value and positive impact;
f) Be open and transparent;
g) Be fair, inclusive, and diverse;
h) Be reproducible;
i) Have a positive impact on people;
j) Not discriminate in any way against individuals and/or entity/institution;
k) Be in line with the use of AI models that allow commercial use and free use for research purposes;
l) Promote ethical conduct in data pipelines and the use of models, from:

m) data collection for model training;
n) training, evaluation, and inspection of results; and
o) deployment for use in development and production environments.

THE AI GUIDELINES

Following the above principles and code of ethics, the MICIT is proposing the following AI guidelines to be used in Tanzania. The goal is to assist in managing risks for a trustworthy ethical AI. The guidelines are follows:

SNo	Guideline	Explanation
1	Develop a code of ethics	Create a code of ethics is the first step in developing ethical AI. This code should outline the values and principles that the AI system should follow. The code should be created in collaboration with relevant key stakeholders, such as employees, customers, and industry experts. This will ensure that the code reflects the values and needs of all parties involved.
2	Ensure diversity and inclusion	Ensure that the data used to train the AI system is diverse and inclusive, which is crucial to avoiding perpetuating biases. This will avoid discriminatory outcomes that can harm individuals or groups. Therefore, it is essential to ensure that the data used is representative of different genders, races, ethnicities, and other diverse factors.
3	Monitor the AI system	Regularly monitor the AI system to ensure that it is performing as intended and not causing harm. This includes regular testing, auditing, and analysis of the system. Monitoring also involves identifying and addressing any errors or issues that may arise. This will help ensure that the AI system continues to function ethically. It is through this monitoring that responsible parties identify improvements and potential errors, such as biases in responses, and ensure that ethical principles are up-to-date.
4	Partner with ethical providers	Partner with ethical providers who share the values of the entity/institution and can help develop and implement ethical AI. Look for providers who prioritize diversity and inclusion, transparency when developing and using AI systems.
5	Transparency	AI developers must adhere to the principles of explainability and transparency. In other words, the criteria and models should be accessible and clear to users. It is crucial to be transparent about how the AI system works and what data it uses. Transparency helps to build trust with stakeholders. It also helps to ensure that the AI system is not used to exploit individuals or

		groups. Therefore, it is essential to be transparent about the data used to train the AI system, the algorithms used, and how decisions are made.
6	Bias in learning and data	AI learns from training, which is done through data input into its systems. This can be done by the development team or through user interactions, creating data histories. It is essential to be cautious about bias in information that could lead to discrimination. It is important to have careful analysis and filters for the training of each AI system.
7	Address privacy and security concerns	Privacy and security concerns arise when personal data is collected, processed, or stored. It is essential to ensure that the AI system is compliant with data protection regulations, especially when used for customer support, sales, and marketing. It is crucial to ensure that this personal information is protected and compliant with the Data Protection Act of 2022.
8	Educate entity/institution employees	Educating entity/institution employees on the ethical implications of AI and providing them with training on the use of ethical AI is essential. This will help ensure that all employees involved in developing or using AI systems understand the importance of ethical AI. Providing training will also help employees understand how to identify and mitigate potential ethical issues.
9	Anticipate risks	Risks can arise from the data used to train the AI system, the algorithms used, and how the AI system is used. Therefore, it is essential to anticipate potential risks and develop strategies to mitigate them. This will help ensure that the AI system functions ethically and avoids causing harm.
10	Conduct ethical reviews	Conduct regular ethical reviews of the AI system to ensure that it is aligned with the expected standards. Ethical reviews should involve evaluating the AI system's performance, identifying any ethical issues, and taking steps to address these issues.

RECOMMENDATIONS ON IMPLEMETING THE AI GUIDELINES

Proposal 1: Human agency and control

An artificial intelligence system should provide reliable support for decision-making while being subject to constant human monitoring and control. In the development of the AI system, it is necessary to document the following:

- possibilities and functionalities of the AI system;
- scenarios of use;

- operational structure and configurations that contribute to robust and accountable use of the system;
- limitations of the AI system;
- segments in which the AI system is not designed to be used;
- overview of the accuracy and regularity the AI system's work and description of the extent to which such results can be expected for generalised use in scenarios not originally considered;
- limits to the expected further development of the AI system without direct human intervention.

If the persons responsible for the monitoring and control of the system detect some anomalies in the behaviour of the system which, over time, may lead to the undesirable situation described in the point above, they shall have the authority, at their discretion, to temporarily shut down the system for a limited period of time, providing a detailed justification for this action. This decision shall be subject to review by the larger team that took part in designing the system or by the relevant authority.

Recommendation 1

i) Provide technical documentation that clearly explains the design of the AI system, its subsystems and components, including the mechanisms for monitoring and control of the operation of the system.

ii) Design a system that allows monitoring and control of its operation and ex-post analysis of the processing results against the input data.

iii) Allow the option to choose between different processing results produced by the system when the system involves interaction with users, and when it can produce more than one processing result with different likelihoods.

iv) Include early detection of harmful impacts or side effects that affect equality and human rights in the process of planning, designing and developing the system, and allow for monitoring and ex-post analysis of the work of the system.

v) Identify and document a range of models to assess whether the system is functioning properly. Different assessment methods help to identify anomalies in the work of the system more efficiently.

vi) For system designers – analyse the source data used to train the system algorithm and get a comprehensive picture of whether the source data actually represents variability for all users or only for a small group of users. The dataset used to train the system model should be representative - it should be an accurate picture of the real system being modelled.

vii) Identify the persons responsible for responding to problems in the work of the system, who monitor the system and control its work at all stages, from development and testing, through learning and training, to full exploitation, i.e. the day-to-day operation of the system. It is necessary to map these persons, define their exact tasks, and determine the way in which they will be selected, trained, monitored and evaluated in terms of their capacity over time, as it is possible that the AI system will be constantly learning and evolving and, over time, will outgrow the capacity of the individuals responsible for its control.

viii) Identify the elements of the system, user tools and reporting tools including the ability to understand the output results of the system, based on which certain actions can be taken (e.g. shutting down the system);

ix) Establish and document criteria for the use of the system, the criteria must include the metrics and thresholds. If the assessment report shows values that exceed the thresholds, it is necessary to define and document the approach and plan to address the identified issues.

Proposal 2: Technical reliability and security

The key requirement to build a robust artificial intelligence system is its technical reliability and security. Technical reliability is ensured when the general recommendations for the development of software systems is followed. In addition, specific methods should be introduced for systems based on machine learning or other AI development methods. Security ensures that the system functions as intended, regardless of possible attacks. It is essential to assess the security of a system before using it in areas where security is a critical parameter. It should be noted that, it is difficult to predict all scenarios in advance and to develop systems that provide both security constraints and flexibility in generating creative solutions adapted to different input data. As AI technology evolves, there are also new attack

risks that should be anticipated, such as: training data poisoning, avoidance attacks, access to sensitive training data, model theft and adversarial attacks. Before developing a system, all risks and consequences of attacks should be considered in order to make the right decisions for system development.

Recommendation 2

i) Identify other metrics for training and monitoring assessment - Using multiple metrics instead of just one helps to understand the relationship between different types of errors and user experiences. Consider metrics such as collecting user feedback through surveys, values that measure system performance at the level of the whole system, and short- and long-term validity, such as click-through rate or customer lifetime value, as well as the rate of false-positives and false-negatives, disaggregated by subgroups (categories). Also make sure the metrics are relevant.

ii) Whenever possible, verify input data - It is necessary to continuously analyse the input data in order to ensure that it is sufficiently understood. If this is not possible (for sensitive data), the input data should be analysed by calculating aggregated, anonymised group values and statistics.

iii) Understand the limitations of datasets and models - a model that is trained to detect correlations should not be used to make decisions about causality and/or assumption of causation. Modern machine learning models largely reflect the regularities in the data used to train them. In order to recognise the capabilities and limitations of the model, the training procedure must define the scope and coverage of the different usage scenarios. Whenever possible, explain the limitations to users in a transparent way.

iv) Rigorous testing – the testing should be performed in the following manner:
- rigorous modular testing to test each system component individually (components include the code, the data and the model itself);
- integration tests to understand how individual components interact with other parts of the system;

- proactively detect the input data drift by testing the statistical values of data entered into the system to ensure that they do not change in unforeseen ways;
- usage of quality datasets to ensure that the system works as intended. Update this dataset regularly according to changes in users and usage scenarios in order to reduce the risk of training on the test dataset;
- iterative user testing to include the different needs of users in different development cycles; and,
- build quality assurance into the system to avoid unintended errors or prevent them from causing an immediate response.

v) Monitor the system during use - Continuous monitoring ensures that the system works as intended, taking user feedback into account. It is best to plan time intervals for fixing any problems in the system. Also consider both short-term and long-term solutions to problems. Balance the short-term and long-term solutions. Before updating the model used, analyse the differences between the model used and the proposed changed model, and the impact of the new model on the overall quality of the system and the user experience

vi) Identify potential threats - identify the undesirable consequences of system errors and assess the likelihood and severity of these consequences. Develop a rigorous threat model that will predict as many attacks as possible. A system that allows the attacker to change machine learning input data is more vulnerable than a system that processes metadata collected by servers because it is more difficult to change such input characteristics without direct access to the servers.

vii) Define the procedure for removing threats - set up an internal “red team” to try to attack the system or organise an external challenge with prizes that will put the system to the test. Also, develop the procedure for removing different types of threats.

viii) Ongoing education - provide education for the team about the latest types of threats and attacks that are emerging in the field.

Proposal 3: Privacy, personal data protection and data management

The concepts of privacy and protection of personal data are closely linked to the principle of not causing damage. In order to prevent the violation of privacy and the right to the protection of personal data, appropriate data management is required. Such data management includes the quality and integrity of the data used, its relevance to the area of life in which the system is developed and used, data access protocols, and the ability of the system to process data in a way that protects privacy and the right to the protection of personal data. Data management ensures the accuracy, security and availability of data with the aim of maintaining data quality and data protection. A data holder is obliged to protect the data in an AI system. It is necessary to ensure lawful access to data while respecting the privacy of the individual, in accordance with data protection regulations, in particular the protection of personal data.

Recommendation 3

i) Define data elements and data entry to create a common business vocabulary in a glossary;
ii) Identify data attributes (metadata) and data entry;
iii) Define user roles and procedures for authentication and authorisation of data access;
iv) Develop policies and rules that define how specific data should be managed during its lifecycle;
v) Manage data codebooks so that all operational and analytical systems use the same classifications (master data management);
vi) Develop data stores that support data encryption, anonymisation and pseudonymisation, and integration with the data catalogue. A data catalogue includes:
    a) a business glossary;
    b) automated data discovery, profiling, tagging, cataloguing and glossary mapping;
    c) automated detection of sensitive data and management classification; and,

d) interoperability with other catalogues, tools and applications for sharing metadata.

Proposal 4: Transparency and Explainability and Traceability

4.1 Transparency

Transparency is important for at least three reasons:

a) Autonomous and Intelligent Systems (AIS) can make mistakes or cause damage, and transparency is necessary to detect how and why;
b) AIS should be understandable to users; and,
c) accountability is not possible without adequate transparency.

One of the characteristics of intelligent systems is their autonomous nature. The future development of artificial intelligence systems in some fields such as healthcare, pharmaceuticals or law will depend on the approach and ability to trace the decision-making process, on the interpretation techniques and explainability of the results, on the way users interact with intelligent systems, and on the presentation of the results to end users. Transparency is the key component that contributes to the development of reliable and trustworthy artificial intelligence, and it consists of three elements:

d) traceability of the AI system,
e) explainability of the AI system and the system model in particular, and
f) communication - dialogue with all stakeholders about the limitations of the AI system.

4.2 Explainability

Decisions resulting from the use of AI need to be explained and understood by those who are directly or indirectly affected by them, so that the decisions can be challenged. However, it is not always possible to explain why the model suggested a particular decision or outcome (i.e. what combination of input factors led to such an outcome). These are the so-called “black box” models, and they require additional attention, i.e. a different set of measures to achieve explainability (e.g. traceability, external evaluation and transparent communication about the scope and capacities of the AI system).

4.3 Traceability

Traceability is a crucial requirement for creating robust and accountable AI systems. New information and communication technologies, such as the Internet of Things (IoT), Cloud Computing and Mobile Computing, have enabled the further development of approaches to Big Data processing and artificial intelligence algorithms. In order to understand and interpret the information contained in datasets, the most important facts must be filtered out and conclusions must be based on knowledge.

Recommendation 4

i) Distinguish Transparency and Explainability and Traceability at the level of data origin, access and extraction,
ii) Distinguish Transparency and Explainability and Traceability at the level of machine learning algorithms and models,
iii) Distinguish Transparency and Explainability and Traceability processes for automated data preparation and processing, and processes for generating conclusions from identified input factors and output recommendations relevant to problem solving.

Proposal 5: Diversity, non-discrimination and equality

For an AI system to be reliable and accountable, it must allow for inclusion and diversity in its lifecycle. AI systems need to be user-centred and designed so that anyone can use the AI products or services, regardless of age, gender, ability and other characteristics. It is particularly important to make these technologies accessible to persons with disabilities, who can be found in all social groups. An important requirement for robust and accountable artificial intelligence is its non-discriminatory behaviour that respects diversity and contributes to fairness.

Recommendation 5

i) Analyse the system in real time to detect both intentional and unintentional biases and discriminatory patterns. When biases (discriminatory patterns) in

data become apparent, the team needs to analyse and understand where they are coming from and how they can be mitigated (preferably completely eliminated).

ii) Design and develop the system without intentional bias and review the system regularly to avoid it. Unintentional bias also includes stereotypes.

iii) Check data and data sources before starting to train the algorithm.

iv) Develop and integrate mechanisms to ensure user feedback in order to raise awareness of biases and issues that users identify.

v) Establish multidisciplinary teams to evaluate the relevant parameters. Diverse teams help present a wider range of experiences in order to minimise bias and discrimination.

vi) Ensure objectivity and set up a mechanism to eliminate bias.

vii) If the system proves inadequate, if it is biased, if it makes discriminatory decisions or is generally unsuccessful and it is not possible to improve it, withdraw it from circulation. Evaluate the damage that such a system can do to society and individuals in relation to the damage that will be done if you withdraw it.

viii) Include members of different ages, background, genders, qualifications and cultural perspectives in the team. This kind of diversity provides access to a range of experiences in order to minimise bias.

ix) Test the system starting from the early design phase, and test it often.

Proposal 6: Risk management and Accountability

Risk management requires timely identification, assessment, documentation and minimisation of the potentially harmful effects of the AI system. This means that the relevant interests and values represented by an AI system must be identified so that, in the event of conflict, any compromise made can be explicitly recognized and assessed in terms of the risk to safety and ethical principles, including fundamental rights. Any decision to compromise must be well justified and properly documented. The issue of accountability is closely linked to adequate planning for the development and monitoring of the system during its production phase, as well as to risk management and procedures for establishing accountability and remedying

damage caused by the use of the system. All people involved in the development of the system are accountable for considering the impact of the system on the environment in which it is deployed, as are the companies that have invested in its development.

Recommendation 6

i) Establish clear and understandable rules and policies for designers and development teams to avoid disputes over tasks and responsibilities.
ii) Specify where the accountability of those who developed the system ends. This is particularly important because those who developed the system have no control over the way in which the system is used.
iii) Keep records of the design process, functionality development and decision-making method of the system. This procedure should be regulated in a separate document.
iv) When creating the system, align the use and outputs of the system with regulations and national standards.

CONCLUSION

AI systems must be recognisable as such. Users must be informed that they are interacting with a system, and they must be able to request communication with a human being if necessary. The utility, effectiveness, efficiency and usability of an AI system are ensured by involving end users in the design, evaluation and implementation of the graphical user interface.

These Guidelines have been created with the aim of providing guidance to all stakeholders in the artificial intelligence ecosystem. In the absence of a firmer legal framework, which is being worked on, these Guidelines ensure further progress in this area, which will continue to expand in the future.