United Nations System **10 October 2024**

CEB/2024/HLCM/28/Add.1/Rev.1

CEB

Chief Executives Board

for Coordination

High-level Committee on Management (HLCM)
48th Session, 3-4 October 2024
UPU Headquarters, Bern

HLCM Task Force on the use of Artificial Intelligence in the UN system

Framework for a Model Policy on the Responsible Use of Artificial Intelligence in UN System Organizations

Table of Contents

Section 1: Preamble

A. The potential of AI in advancing the UN System Organizations' missions and business operations

Artificial Intelligence (AI) has the potential to significantly enhance the United Nations System Organizations' missions and operations by improving efficiency, decision-making, and global reach. By leveraging AI technologies in ethical and properly governed ways, the entities are unlocking their potential to better analyze their data to identify patterns and trends, enabling more informed decision-making to support mandate delivery, protecting human rights, advancing progress towards the Sustainable Development Goals, as well as supporting access to institutional knowledge and the generation of new insights. AI-supported data analytics can help, for example, to monitor and predict local and global crises, optimize resource allocation, ensure targeted and effective assistance to affected populations, and improve the effectiveness of peacekeeping missions. AI can enhance the UN's capacity to address complex global challenges such as climate change, health pandemics, and sustainable development by providing advanced modeling and simulation tools.

Furthermore, AI has the potential to streamline the UN's administrative and operational processes, ensuring consistent levels of process and service quality and processing times, and potentially leading to increased available capacity for innovation and service improvement. Automation of routine tasks within the areas of document processing, data management and translation and editing of documents, can enable personnel to focus on activities of a different nature, in line with current and emerging needs of the organizations. AI-powered chatbots and virtual assistants can improve communication and service delivery to personnel, member states, partners, and the general public. By integrating AI in a transparent, responsible and trusted manner into their operations, the organizations can enhance their responsiveness and adaptability in a rapidly changing world, ultimately advancing their mission to promote peace, security, and development.

B. Challenges and risks

While AI holds tremendous potential for advancing the UN System Organizations' missions and business operations, it is crucial to acknowledge and address the associated risks to UN system organizations, their personnel, stakeholders and affected populations. The integration of AI presents challenges such as ethical and legal considerations, data protection and data privacy concerns, intellectual property issues, information security considerations, environmental and sustainability concerns, and the potential for harm, including discriminatory outcomes, caused by biased data, algorithms and new or deepened injustices from, *inter alia*, the application of automated decision-making tools. AI system testing, development and maintenance also have the potential to require significant organizational and financial resources.

AI systems cannot compensate for inconsistencies, errors, or bias that may be inherent in data sources not under the strict ethical and editorial control of the UN. Entities applying AI systems should ensure transparency, fairness and accountability in AI deployment, which are essential principles to mitigate these risks. They should adopt robust governance frameworks based on human rights and ethical guidelines as outlined in this document to harness the benefits of AI while safeguarding against its potential adverse impacts, thereby promoting a balanced and responsible approach to AI integration.

C. Guiding principles and frameworks

The following principles and frameworks have informed and guided the formulation of this framework for a model policy and should guide its implementation.

a) Principles for the Ethical Use of Artificial Intelligence in the UN System (see Section 3 below),
b) UN Principles on Personal Data Protection and Privacy,
c) International Data Governance – Pathways to Progress,
d) Proposed Normative Foundations for an International Data Governance Framework: Goals and Principles
e) UN 2.0 Quintet of Change,
f) The UN Secretary-General's Guidance on Human Rights Due Diligence for Digital Technology Use,
g) The UN Secretary General's Roadmap for Digital Cooperation: Ensuring the Protection of Human Rights.

Section 2: Purpose, Scope, and Applicability

D. Purpose

The use of this framework is not mandatory. Rather, it serves as a guidance for the development and implementation of policies on the responsible use of artificial intelligence by UN System Organizations. This framework aims to provide a common reference for AI policy development to ensure a level of consistency and harmonization among the policies of UN System Organizations while taking into account their distinct mandates, working modalities and operational realities.

This framework seeks to promote good governance standards in UN System Organizations for the use of AI in furtherance of each organization's mandate, in a manner that is in line with the *Principles for the Ethical Use of Artificial Intelligence in the United Nations System* and that upholds human rights and values expressed in the Charter, Standard of Conduct and treaties of the United Nations.

This framework for a model policy aims to provide guidance to UN System Organizations in developing AI policies that will address the mitigation of AI-related risks, through the introduction of principles, rules and controls while ensuring that the organizations are mindful and compliant with ethical standards, as well as expectations for an international intergovernmental organization.

This framework advocates the adoption by each organization of an institutional accountability structure that (i) facilitates a multidisciplinary approach towards AI impact assessments and risk management, and (ii) ensures that risks associated with AI integrated or deployed by the relevant organization are systematically mapped, measured and managed and regularly re-assessed using, for example, AI impact assessments and tailored procurement guidelines.

E. Scope and Applicability

The UN System Organizations are encouraged to adhere to the elements outlined in this framework including through (i) the issuance of *inter-alia* detailed and underpinning policies and guidelines, (ii) the review and update of related and existing policies, guidelines and procedures, which should, where applicable, be revised to include references to their established AI policy and codify elements that are related but not strictly or solely under the scope of their AI policy, and (iii) capacity-building for staff to ensure the skills and knowledge required to ensure compliance. Organizations should ensure consistency between AI policies that they develop and already-existing internal regulations that may also regulate AI use such as data protection and privacy policies, procurement guidelines, or rules on the approval of IT tools and systems.

Organizations should conduct consultations with internal stakeholders prior to the issuance or material update of AI or related policies, guidelines and procedures.

This framework is applicable to those parts of services or technologies that are using any form of AI directly or indirectly to ingest, transform, and generate content, including data. It strictly applies to AI systems, AI-supported services, and any form of AI augmented solution, used, integrated or deployed by UN System Organizations, whether developed in-house, acquired or consumed from external sources, including free to use AI tools and services.

This framework covers AI systems and AI-enabled solutions of any sort, offered by the organization for internal and external audiences, in line with the relevant data protection and classification standards. It applies to publicly available AI platforms and solutions offered as part of the organization's service catalogues or approved for use by personnel or stakeholders. It also applies to the organization's development of any AI solution for a third party's use and ongoing maintenance and provision of an AI solution as a service (SaaS) to any end user, even those that are external to the UN system.

This framework does not cover a non-UN System third party's use of AI solutions that were developed by a UN System Organization for the third party.

Section 3: Application of the Principles for the Ethical Use of Artificial Intelligence in the UN System

AI systems are systems that have the capacity to automatically process data and information in a way that resembles intelligent human behaviour, and typically include aspects of reasoning, learning, perception, prediction planning or control.[1] AI systems intended for use in the UN System Organizations should be managed in accordance with the *Principles for the Ethical Use of AI in the United Nations System* as listed below, along with any other internal guidance and procedures. Annex N provides the full text of the *Principles for the Ethical Use of AI in the United Nations System*.

- Do no harm
- Defined purpose, necessity, and proportionality
- Safety and security
- Fairness and non-discrimination
- Sustainability
- Right to privacy, data protection and data governance
- Human autonomy and oversight
- Transparency and explainability
- Responsibility and accountability
- Inclusion and participation

These principles should be incorporated into the development, acquisition, and use of AI tools and systems.

Personnel should be trained, guided and reminded about their own obligations towards upholding ethics at the workplace, including the use of AI data and tools.

Section 4: Institutional Structures for Accountability

The UN System Organizations should establish AI accountability structures that facilitate the implementation of their AI policy and align with or embed these structures in existing accountability mechanisms within their organization. To that end, organizations should define in their respective AI policy roles and responsibilities for the implementation and monitoring of the policy based on their already-existing accountability structures and processes. These structures and their designated agents are responsible for ensuring continuous evaluation, transparency, and accountability in AI use, development and deployment across the organization. The following key functions should be taken into consideration.

[1] Principles for the Ethical Use of Artificial Intelligence in the UN System.

- Guidance on the strategic role of AI in the organization,
- Oversight over the implementation of the organization's AI policy and its alignment with relevant international policy mechanisms and inter-agency collaboration,
- Monitoring changes in the AI regulatory landscape including changes related to local jurisdiction that may be of relevance for the organisation (within and outside the realm of privileges and immunities),
- Implementation and integration of the AI strategy within the organization, including prioritization of use cases and resource allocation,
- Use and deployment of AI within the organization as well as the use of external partners, vendors and implementing partners,
- Governing the use of AI solutions not provided as part of organisational service offerings but consumed by personnel,
- Alignment of AI initiatives with ethical principles, guidance on and capacity-building and knowledge management for ethical AI development and use, evaluation and risk mitigation, as well as organisational processes to establish best practices and reflect on lessons learnt,
- Compliance of AI initiatives with human rights (human rights due diligence) and with data protection and privacy policies through, *inter alia*, personal data mapping and data protection and privacy impact assessments,
- Identification, assessment and mitigation of risks associated with AI projects, including receipt and management of complaints,
- Provision of a digital environment to deploy AI solutions that are secure, scalable, and compliant with governance requirements,
- Identification and implementation of measures covering a necessary shutdown of AI solutions as well as business continuity plans in case of AI degradation,
- Development of instruments governing the use of data for the purposes of AI as well as using AI directly at the edges of the organisational boundaries, e.g. governing whether public organisation data is under fair use, restricted or open for being fed into AI models by external parties,
- Governing the terms of use, particularly by external parties, of any data or service that is AI based and offered by the organisation,
- Encouraging transparency and dialogue among staff about the role of AI within the organization and facilitating consultation with internal stakeholders prior to the deployment of AI systems and the issuance of AI policies, guidelines and procedures.

In fulfilling these tasks, cross-functional representation of pertinent entities in the organization is required.

Section 5: Risk Management and Mitigation

AI systems are dynamic and may perform in unexpected ways once deployed or after deployment. The use and impact of AI systems can also vary based on the deployment context. AI systems' performance and impact should be closely monitored across their entire life cycle so that risks to, *inter alia*, UN organizations, their personnel, stakeholders and affected populations that may be affected directly or indirectly by the AI system can be managed and mitigated. Organizations should continuously monitor and assess their AI systems in order to understand the risks, and take measures to avoid, manage, or mitigate risks, leveraging existing risk management processes and AI accountability structures and processes, as well as ensuring compliance with human rights (through human rights due diligence or child rights impact assessments, for example) and data protection and privacy policies (through personal data mapping and data protection impact assessments, for example).

F. Risk Assessments

The organization should ensure that all risks identified in accordance with their AI policy are mapped, measured, and managed. If risks associated with an AI system or application are deemed as "High Risk" or "Very High Risk", as set out below, advice from the established AI accountability body on how to address the identified risks should be sought. It is the responsibility of the responsible official to decide on what course of action to take to mitigate, or where possible avoid identified AI risks.

The organization should develop an AI impact assessment tool, which should include a risk assessment tool, as well as the designated digital environment that should host the inventory of all AI impact assessments, for use across the organization. These tools as well as the digital environment where they will be hosted should be updated, as needed, based on lessons learned and operational necessity.

The organization should conduct AI impact assessments throughout the lifecycle of AI systems they develop, integrate, or deploy. The deployment context, or the way in which the systems operate or are used, may change. When these changes are deemed as significant by those deploying or using the systems, the AI impact assessment should be reviewed. If no changes to the operating environment or intended use of the AI system have taken place, the AI impact assessment, including the risk assessment, should be reviewed periodically, at least once every 24 months. Responsible officials should conduct, maintain, and make available to the organization inventories of all AI impact assessments performed in their entity.

The organization should define linkages to and dependencies on other existing risk assessments, such as those related to data protection and privacy or human rights risks, that should be carried out regularly throughout the lifecycle of the AI system and define how to address overlap and implementation of all necessary assessments. The organizational risk assessment framework

should be sensitive to the potential for multiple and interrelated risks to people, organizations, and ecosystems.

To the extent possible, the organization should evaluate any bias in AI outputs, particularly related to gender and racial discrimination, and information that would contrast UN System ethics.

G. Risk Classification

The risk-based approach distinguishes between four distinct risk levels. The AI risk management process should consist of a continuous iterative process applied throughout the entire AI lifecycle and should be documented in an AI impact assessment. For this framework's purposes, risks associated with AI systems can be classified as follows.

- Low Risk – No specific obligations, beyond existing institutional safeguarding, risk and data protection requirements,
- Medium Risk – The organization should implement specific remedial or monitoring measures, including continuous monitoring of the AI system,
- High Risk – The organization should implement specific remedial or monitoring measures, including consultations and alignment with the accountability mechanisms that facilitate the implementation of the organization's AI policy,
- Very High Risk – The Head of Entity should be required to implement specific remedial or monitoring measures, including consultations and alignment with the accountability mechanisms that facilitate the implementation of the organization's AI policy. If the Very High Risks cannot be properly mitigated, the adoption and use of the AI system may be prohibited ("Unacceptable risk").

Given the evolving nature of the AI field, the applicability of this framework and determination of risk levels should not depend on specific types of AI systems but rather on a series of criteria. The following criteria are among key factors to be examined in determining risk levels.

- Compliance with international law, including human rights law and data protection and privacy policies and regulations, and responsible AI principles,
- Scale, i.e., seriousness of adverse impacts (e.g. potential discrimination or deepening inequity) and likelihood associated with the AI system, throughout the AI lifecycle,
- Scope, i.e., breadth of application of the AI system, such as the number of individuals that are or will be affected,
- Prohibition of use cases representing unacceptable risks to human rights or violation of data privacy requirements.

H. Risk Ownership

AI systems development, deployment and use should be managed through a risk-based approach so that risk can be mitigated or, where possible, avoided. The organization should conduct risk assessments to identify potential risks and their impact on the organization. These assessments should involve key stakeholders from relevant departments, including substantive areas, legal (including data protection), IT, ethics, safeguarding, and staff representation bodies.

In accordance with existing procedures, the organization should establish cross-functional capacity focused on risk identification and mitigation. The AI risk management and oversight functions should integrate within the organization's broader risk management approaches. The organization should also designate specific risk owners for each identified risk. Risk owners should be assigned and are responsible for developing and implementing risk mitigation strategies, monitoring risk indicators, and reporting on risk status to the existing and relevant accountability structures.

The risk owner should coordinate with relevant data and application owners to identify specific risks and mitigation measures.

Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks should be documented and clear to individuals and teams throughout the organization.

As a measure of segregation of duties, control and production positions should be clearly differentiated: Professionals carrying out test and evaluation tasks should be independent from AI system developers, with a reporting line to risk management functions or directly to senior management.

The organization should ensure that response responsibilities and authorities are defined, assigned and regularly tested, to ensure immediate actions can be taken in case of critical incidents related to AI solutions.

Regular training and updates should be provided to AI risk owners to keep them informed of the latest AI developments and risk management techniques.

I. Continuous Monitoring

Establishing continuous monitoring and business case re-evaluation, particularly considering organizational risk appetite and changing priorities, can help organizations to track AI risks and impacts. This involves setting up metrics for secure and trustworthy AI, capacity to identify and mitigate threats, and monitoring AI systems outputs and interactions. The organization's established AI accountability body should ensure that the following risk monitoring and mitigation measures are put in place.

- Establishment of metrics and indicators for secure and trustworthy AI and systematic monitoring of AI systems and activities that are guided by and aligned with existing frameworks within the UN system,
- Development of capacity for identification of threats, vulnerabilities and other cybersecurity risks,
- Creation of an AI risk register across AI systems and data management and sensitivity,
- AI Impact Assessments to be carried out throughout the lifecycle of AI systems,
- Human rights due diligence, personal data mapping, and data protection impact assessments to be carried out regularly, throughout the lifecycle of AI systems,
- Monitoring own use and potential external use of AI solutions, including conducting regular assessments on whether the organization's data is being integrated into external AI solutions without appropriate permissions,
- Establishment of risk reporting mechanisms and incident response plans or expansion of already-existing ones to apply to AI systems,
- Tools to manage the AI systems lifecycle including procedures for the decommissioning of systems,
- Training of the workforce on risk monitoring and management.

Section 6: Data Governance and AI Systems Deployment Governance

J. Data Governance

Data governance is required for the deployment and use of AI systems. The organization should determine the relationship between this framework for a model policy and the existing governance framework (if any) or the need to establish a data governance framework. Elements to be considered are as follows.

- Relevant data classification processes, procedures, and organizational frameworks to identify, classify, and establish control requirements for data and information assets, including definitions,
- Organization's data protection and privacy regulatory instruments and their applicability to the use of AI,
- Standards for the collection, storage, retention, sharing, and access of data used in AI systems,
- Standards for data quality, integrity, and suitability including representativeness and proprietary data considerations in AI applications and systems,
- Accountability structures for value-driven and risk-based implementation of the data governance framework integrated in the organization's existing oversight structures,
- Mechanisms for regular audits, assessments and updates of data governance practices to ensure compliance with evolving standards, good practice, and regulations,

- The ownership and intellectual property rights related to AI-generated content and AI-driven decisions are clearly defined. To the extent possible, ensure that the organization retains the rights to AI-generated outputs used for its mission,
- Respect the intellectual property rights of third parties by ensuring that any data or content used by AI systems is properly licensed or falls under fair use,
- Provide appropriate attribution for any third-party data, content, or algorithms used in AI systems,
- Consider open licensing models for AI-generated content where appropriate to support the organization's mission and promote knowledge sharing.

In scope of the data governance framework are data used to train AI systems ("input data"), AI model parameters/weights ("model data"), data produced by AI systems ("output data"), and data produced by the interaction with AI systems ("user data"). An organization's data governance framework and its policies, procedures, standards, and processes should be in recognition of the organization's privileges and immunities and special attention should be paid when third-party systems are used for data storage or processing including AI model training, re-training, and fine-tuning. All personal data should be processed in accordance with the UN Principles on Personal Data Protection and Privacy.

K. Governance of AI Systems Deployment

The organization should clarify the relationship between this framework for a model policy and the existing governance framework (if any) or the need to establish an application development, service delivery and lifecycle management framework. Elements to be considered are as follows.

- Standards for the development, testing, acquisition, sharing, and interoperability of foundational AI technologies, including open-source technologies,
- Standards for evaluating potential tools and suppliers of AI technologies,
- Incorporation of *Principles for the Ethical Use of AI in the United Nations System* and Secure by Design and Data Protection by Design in the development of technologies, along with respect for intellectual property,
- Processes and standards for applications lifecycle management, from requirements gathering and system design to development, deployment, performance and adverse incident monitoring, and decommissioning,
- Labelling of AI-generated content to distinguish it from human-created content,
- Implementation of quality control measures to ensure that AI-generated content meets the organization's standards and aligns with its mission,
- Ensuring that AI-generated content respects cultural sensitivities and avoids perpetuating harmful stereotypes,
- Mechanisms for regular audits, assessments, including human rights due diligence, personal data mapping and data protection impact assessments,

- Updates of the application development lifecycle management framework to ensure compliance with evolving standards, good practice, and regulations.

The organization should determine the need to provide training and other capacity building initiatives to staff to equip them with the necessary skills and knowledge to manage data governance and AI system deployment governance challenges and opportunities.

Section 7: Monitoring, Accountability and Compliance

L. Compliance

Regular compliance reviews and audits should be conducted to ensure adherence to the established AI policy and identify areas for improvement. To ensure the auditability of AI systems, the organization should establish a body of verifiable information, including:

- Documentation of standards for AI development and deployment processes,
- Records of the AI model development process, including data sources, data preprocessing steps, model selection, training procedures,
- Mechanisms for tracking and explaining AI-driven decisions, including the rationale behind choosing specific algorithms, data sets, and model architectures,
- Data retention and access policies for audit purposes,
- Data lineage.

M. Complaints

The UN system organization should establish or adapt existing mechanisms through which individuals or entities can raise complaints if they have identified cases where the use of AI is non-compliant with the organization's established AI policy or where the use of AI by the organization has had an adverse impact on human rights.

In the case of AI-supported enabled decisions, the system should allow sufficient traceability in order to support a comprehensive review. Organizational policies should clarify what parts of AI solutions including their data can be disclosed as part of forensics exercises related to investigations and critical incidents.

A transparent and accessible, and, as far as possible, independent and impartial, reporting system should be established to handle complaints and investigate non-compliance. Such a system should be linked to existing whistleblower protection policies and procedures. Complaints procedures should include a mechanism for immediate action in cases of imminent threat or severe impact on personnel or stakeholders. Such remedy should include the temporary suspension of the AI solution as a whole in case the issue cannot be resolved at the case level.

Organizations should determine whether AI use may impact on any individual rights established by other internal policies and ensure consistency between the established AI policy and such

policies. For example, an organization’s data protection and privacy policy may establish a right not to be subject to automated decision-making where the decisions are likely to have adverse legal or other significant adverse effects on the individual. Such policies may establish mechanisms and processes through which individuals can lodge complaints and seek redress.

Annex

N. Principles for the ethical use of artificial intelligence in the United Nations system

Do no harm

Artificial intelligence systems should not be used in ways that cause or exacerbate harm, whether individual or collective, including harm to social, cultural, economic, natural or political environments. All stages of an artificial intelligence system's life cycle should operate in accordance with the purposes, principles and commitments of the Charter of the United Nations. All stages of an artificial intelligence system's life cycle should be designed, developed, deployed and operated in ways that respect, protect and promote human rights and fundamental freedoms. The intended and unintended impacts of artificial intelligence systems, at any stage in their life cycle, should be monitored in order to avoid causing or contributing to harm, including violations of human rights and fundamental freedoms.

Defined purpose, necessity and proportionality

The use of artificial intelligence systems, including the specific artificial intelligence method(s) employed, should be justified, appropriate in the context and not exceed what is necessary, and proportionate to achieve legitimate aims that are in accordance with each United Nations system organization's mandate and governing instruments, rules, regulations and procedures.

Safety and security

Safety and security risks should be identified, addressed and mitigated throughout the artificial intelligence system's life cycle to prevent or, at least, limit any potential or actual harm to humans, the environment or ecosystems. Safe and secure artificial intelligence systems should be enabled through robust frameworks.

Fairness and non-discrimination

United Nations system organizations should aim to ensure the equal and just distribution of the benefits, risks and costs associated with artificial intelligence systems and to prevent bias, discrimination and stigmatization of any kind, in compliance with international law. The use of artificial intelligence systems should not lead to individuals being deceived or to unjustifiable restrictions on their human rights and fundamental freedoms.

Sustainability

Artificial intelligence should be aimed at promoting environmental, economic and social sustainability. To this end, the human, social, cultural, political, economic and environmental impacts of such technologies should be continuously assessed, and appropriate mitigation and prevention measures should be taken to address adverse impacts, including on future generations.

Right to privacy, data protection and data governance

Individuals' privacy and rights as data subjects must be respected, protected and promoted throughout the life cycle of artificial intelligence systems. When the use of artificial intelligence systems is considered, adequate data protection frameworks and data governance mechanisms should be established or enhanced, in line with the personal data protection and privacy principles, also to ensure the integrity of the data used.

Human autonomy and oversight

United Nations system organizations should ensure that artificial intelligence systems do not impinge on human beings' freedom and autonomy and should guarantee human oversight. All stages of an artificial intelligence system's life cycle should follow and incorporate human-centric design practices and leave meaningful opportunity for human decision-making. Human oversight includes ensuring that humans have the capability to manage the overall activity of an artificial intelligence system and the ability to decide when and how to use it in specific situations, including whether to use such a system, and the ability to override a decision made by such a system. As a rule, life or death decisions or other decisions affecting fundamental human rights require human intervention and must not be ceded to artificial intelligence systems.

Transparency and explainability

United Nations system organizations should ensure the transparency and explainability of artificial intelligence systems that they use, at all stages of their life cycles, and of decision-making processes involving such systems. Technical explainability requires that the decisions made by an artificial intelligence system can be understood and traced by human beings. Individuals should be fully informed when a decision that may or will affect their rights, fundamental freedoms, entitlements, services or benefits is informed by or made based on artificial intelligence algorithms and should have access to the reasons and logic behind such decisions. The information and reasons for a decision should be presented in a manner that they can understand.

Responsibility and accountability

United Nations system organizations should have appropriate oversight, impact assessment, audit and due diligence mechanisms, including protection for whistle-blowers, to ensure accountability for the impacts of the use of artificial intelligence systems throughout their life cycles. Appropriate governance structures should be established or enhanced to ensure that humans or legal entities are made ethically and legally responsible and accountable for artificial intelligence-based decisions made at any stage of an artificial intelligence system's life cycle. Harm caused by or as a result of the use of artificial intelligence systems should be investigated, and appropriate action taken in response. Information on accountability mechanisms should be communicated widely throughout the United Nations system in order to build shared knowledge, resources and capacities.

Inclusion and participation

When designing, deploying and using artificial intelligence systems, United Nations system organizations should take an inclusive, interdisciplinary and participatory approach, and promote gender equality. They should conduct meaningful consultations with all relevant stakeholders and affected communities as part of the processes of defining the purpose of an artificial intelligence system, identifying the assumptions underpinning its use, identifying the associated benefits, risks, harm and adverse impacts, and adopting prevention and mitigation measures.

O. Risk Categories and Areas

Adherence to an established AI policy is not alone an appropriate risk response mechanism. AI-related controls, risk management and remedial processes should be continuously reviewed and tightened, alongside existing procedures related to information security, data protection, ethics, performance management and misconduct.

Existing Risk-Management Systems should assess and cover the following risk categories, particularly incorporating aspects of AI use and AI content generation:

- Risks to the organization, its mandate, agency and ability to deliver,
- Risks to personnel and anyone in a direct contractual relationship with the organization,
- Risks and potential liabilities as a result of consumption or use by external parties,
- Risks to other individuals and unintended groups, i.e. residual risk and unintended negative consequences.

The use of AI systems by UN System Organizations can create a number of risks, including but not limited to:

- Reputational risks, particularly related to AI hallucinations,
- Traceability and liability stemming from AI-supported administrative decisions,
- Undetected biases in data and models potentially aggravating existing biases in organizational decision making,
- De-professionalization of the workforce and degradation in quality of work output related to over-reliance on AI tools and vendors as well as AI-supported recommendations,
- Discrepancy in the pace of usage and governance of AI, potentially leading to blind spots and unmanaged risks,
- Vendor and data-lock and the lack of viable alternatives including absence of emergency procedures in case of AI-system degradation,
- Unpredictable cost implications related to equipment, licenses, and power consumption.

P. Detailed suggestions for continuous monitoring of AI risks

Establishing continuous monitoring can help organizations to track AI risks and impacts. This involves setting up metrics as provided in the following examples.

- **Metrics:** Guided by existing frameworks within the UN system, such as the *Principles for the Ethical Use of Artificial Intelligence in the United Nations System* and the *Guidance*

on Human Rights Due Diligence for Digital Technology Use, the organization should establish metrics and indicators for secure and trustworthy AI and systematically monitor AI systems and activities.

- **Red teaming:** A diverse team of AI experts or organizational units should be assigned to simulate attacks, identify main risks, and conduct stress testing of AI systems to uncover potential vulnerabilities and weaknesses. By provisioning processes for red teaming, organizations can build resilient AI systems and enhance their overall security posture. This proactive approach can help in detecting incidents and preventing AI abuse, ensuring robust risk management.
- **Logging:** Creating risk inventories classified by levels of risk and data sensitivity is essential for effective AI risk monitoring. This includes logging AI outcomes and interactions to automate alarms and content moderation, minimizing errors, inaccuracies, and biases. Provisions for securing the logged data and storage are critical to ensuring safety of personal data and effective use for the improvement of AI systems.
- **Impact Assessment:** Organizations should conduct AI impact assessments, human rights due diligence, personal data mapping and data protection impact assessments throughout the lifecycle of AI systems they develop, integrate, or deploy. The AI impact assessment should include an analysis of the systems' intended use and efficacy as well as assessment of risks and cost benefits. The deployment context, systems use, and operation may change over time. Significant changes should prompt a review of the AI impact assessment at least once every 24 months to ensure continued relevance and accuracy.
- **Reporting and Incident Response:** Establishing robust risk reporting mechanisms is critical for the observability, management, and refinement of AI systems. Centralized reporting systems should be in place to capture and address risks promptly. Additionally, having a well-defined incident response plan ensures that any issues or incidents are managed effectively, minimizing potential harm.
- **Decommissioning and Training:** AI systems that are no longer in use or have been deemed unsafe should be decommissioned in a controlled manner to prevent any residual risks. Continuous training and capacity strengthening are also essential for maintaining a workforce adept at managing AI risks. Regular assessments and monitoring by third parties can also provide an unbiased evaluation of AI systems, further enhancing their safety and trustworthiness.

Q. Library of reference frameworks and guidance documents within and outside the UN system

- UN Secretary-General's AI Advisory Body Interim Report: Governing AI for Humanity
- OECD Principles on AI (Organization for Economic Cooperation and Development)
- UNESCO Recommendation on the Ethics of Artificial Intelligence
- EU's Ethics Guidelines for Trustworthy AI

- ISO/IEC Standards on AI
- G20 AI Principles
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- WIPO: Generative AI: Navigating Intellectual Property
- ISO/IEC 42001-2023: Information technology — Artificial intelligence — Management System
- EU AI Act
- NIST AI Risk Management Framework
- Institute of Internal Auditors: AI Auditing Framework

R. Definitions

Artificial intelligence:

Artificial Intelligence is the development of computer systems and algorithms that can learn from, and make predictions or decisions, based on data. It involves building models inspired by assumptions of how the human brain is structured and functions. Machine learning, and particularly deep learning, is at the core of AI, enabling computers to perform tasks that traditionally require human intelligence, such as understanding natural language, recognizing patterns, and solving complex problems.

AI degradation:

Synonym for model collapse, a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality.

AI lifecycle:

The AI lifecycle comprises all processes AI systems undergo from their inception to their retirement, which may be iterative and non-sequential.

AI systems:

Systems that have the capacity to automatically process data and information in a way that resembles intelligent human behaviour, and they typically include aspects of reasoning, learning, perception, prediction, planning or control.

Data governance:

Used as umbrella term for responsible use and protection of data, including roles, responsibilities, and processes for ensuring accountability for and ownership of data assets across the organization.

Human rights due diligence:

Identification and addressing of adverse human rights impacts that the organization may cause or contribute to through its own activities, or which may be directly linked to its operations, products or services by its business relationships.

Risk-based AI management:

Risk-based AI management describes organisational practices concerned with anticipating, understanding, and mitigating risks associated with the deployment of AI systems. It consists of mapping, measuring, and managing risk throughout the AI lifecycle, using predefined criteria. The goal of such practices is to increase the robustness of AI systems, account for their limitations, and ensure their alignment with fundamental and organisational values.