Digital
Cooperation
Organization

DCO PRINCIPLES FOR ETHICAL AI

2025

DOCUMENT DISCLAIMER

The following legal disclaimer (“Disclaimer”) applies to this document (“Document”) and by accessing or using the Document, you (“User” or “Reader”) acknowledge and agree to be bound by this Disclaimer. If you do not agree to this Disclaimer, please refrain from using the Document.

This Document, prepared by the Digital Cooperation Organization (DCO). While reasonable efforts have been made to ensure accuracy and relevance of the information provided, the DCO makes no representation or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained in this Document.

The information provided in this Document is intended for general informational purposes only and should not be considered as professional advice. The DCO disclaims any liability for any actions taken or not taken based on the information provided in this Document.

The DCO reserves the right to update, modify or remove content from this Document without prior notice. The publication of this Document does not create a consultant-client relationship between the DCO and the User.

The designations employed in this Document of the material on any map do not imply the expression of any opinion whatsoever on the part of the DCO concerning the legal status of any country, territory, city, or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The use of this Document is solely at the User’s own risk. Under no circumstances shall the DCO be liable for any loss, damage, including but not limited to, direct or indirect or consequential loss or damage, or any loss whatsoever arising from the use of this Document.

Unless expressly stated otherwise, the findings, interpretations and conclusions expressed in this Document do not necessarily represent the views of the DCO. The User shall not reproduce any content of this Document without obtaining the DCO’s consent or shall provide a reference to the DCO’s information in all cases.

By accessing and using this Document, the Reader acknowledges and agrees to the terms of this Disclaimer, which is subject to change without notice, and any updates will be effective upon posting.

TABLE OF CONTENTS

01

INTRODUCTION

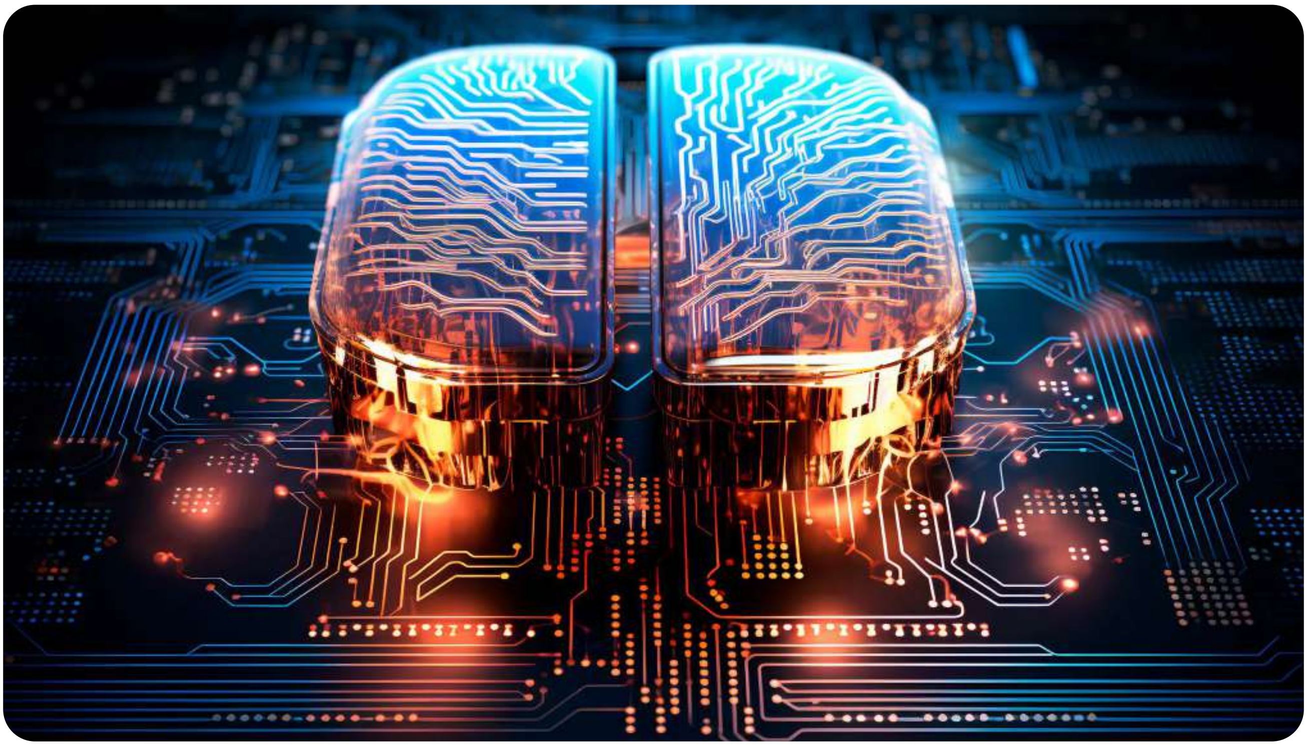
The global AI landscape continues to evolve as countries advance their technological capabilities and governance frameworks. Nations and regions are adopting varied models for AI integration, encompassing regulatory frameworks, industry standards, and implementation strategies. This range of approaches offers valuable insights into AI deployment across different socioeconomic contexts while underscoring the need for adaptable governance frameworks grounded in commonly agreed-upon principles that can bridge regional differences effectively.

Policymakers are increasingly emphasizing the need for clear guidelines governing AI use to harness its potential for social and economic development while mitigating the associated risks. The latter span multiple dimensions, from privacy and data protection concerns to the potential for bias and discrimination in AI systems, from cybersecurity vulnerabilities to challenges with AI transparency and explainability.

The implications also extend to broader societal concerns, including workforce disruption, the safety and reliability of critical AI systems, and the ethical challenges of autonomous decision-making. Understanding and addressing these risks is crucial for developing effective governance frameworks that can maximize AI's benefits while protecting individual and collective interests.

Diverse global approaches to AI governance have emerged, aiming to balance robust societal safeguards with the pursuit of economic growth and technological advancement. Some countries and regions have adopted *prescriptive regulatory frameworks* specifying explicit prohibitions and consequences, particularly for high-risk AI systems, while others have pursued soft-governance models that focus on guidelines and self-regulation. Both approaches are informed by the emerging global best practices and grounded in widely accepted principles, such as transparency, accountability, fairness, and respect for human rights. Additionally, each national approach reflects its distinct blend of legal traditions, political priorities, economic demands, and available resources.

The DCO is committed to fostering an environment within its membership where AI technologies advance responsibly, driving both social and economic progress. This is pursued by the establishment of commonly agreed principles, the development of tools to support national frameworks, the facilitation of stakeholder engagement, and the promotion of international collaboration. At the same time, the DCO places a strong emphasis on adopting a human rights-based and ethical approach to AI governance across its Member States.



In this context, the DCO General Secretariat has developed the **DCO Principles for Ethical AI** (the Principles) to provide Member States with a shared foundation for AI governance. These principles aim to offer clear policy guidance while respecting the diversity of national contexts, with a strong emphasis on human rights protection. They empower Member States to create governance frameworks that align with their specific technological capabilities, regulatory landscapes, and developmental priorities. This approach ensures consistent ethical standards across the DCO ecosystem while allowing flexibility for localized implementation.

The Principles also serve as the foundation for the **DCO AI Ethics Evaluator** (the Evaluator), which is the DCO's policy tool to guide the integration of ethical considerations and human rights perspectives into the design, development, and deployment of AI systems.

The **Principles** and the **Evaluator** (grounded in a risk assessment framework), alongside the in-depth analyses presented in the DCO reports *"Rights by Design: Embedding Human Rights Principles in AI Systems" and "Responsible AI Governance: Global Lessons and International Best Practices for DCO Member States,"* collectively form the **DCO's Ethical AI Governance Toolbox** (the Toolbox). This toolbox is aligned with global standards and incorporates the latest advancements in ethical AI practices.

The Toolbox is designed to assist the DCO Member State governments, developers, and deployers in navigating the intricate and dynamic task of implementing an ethical and human rights-based approach to AI governance. Rooted in the Principles, it provides practical guidance to ensure AI applications adhere to ethical standards, uphold human rights, and deliver meaningful benefits to society as a whole.

Additionally, the Toolbox offers practical and actionable recommendations that bridge the gap between theory and practice. These recommendations serve as a guide for policymakers, developers, and decision-makers in the AI ecosystem to translate ethical AI governance principles into tangible policies and practices.

02

ESTABLISHING THE DCO **PRINCIPLES FOR ETHICAL AI**

2.1 THE IMPORTANCE OF PRINCIPLES

Best practice principles are foundational to national and regional technology governance, providing a consistent, ethical foundation that ensures responsible technology use. They help align diverse stakeholders and regulatory approaches, promoting trust, transparency, and accountability while addressing emerging challenges and risks.

While principles are broadly defined as fundamental laws or moral standards that guide behavior,[1,2] their interpretation can vary across cultural, religious, and historical contexts.

Developing the national ethical AI frameworks begins with identifying the principles to be included. This process requires a clear vision and understanding of national and regional priorities and a deep comprehension of the principles themselves.

Global multilateral organizations are actively developing ethical AI principles to ensure the technology is deployed responsibly across borders. Core principles such as human-centricity, transparency, accountability, fairness, and respect for human rights are consistently emphasized in multilateral AI frameworks, as outlined in Annex C. These international initiatives, together with the frameworks of DCO Member States, form the basis for the DCO Ethical AI Principles, aligning with global standards while accommodating regional and national variations to create a cohesive and inclusive approach to AI governance.

2.2 RIYADH AI CALL FOR ACTION DECLARATION (RAICA)

The Riyadh AI Call for Action Declaration (RAICA),[3] adopted at the Global AI Summit in September 2022, serves as the DCO's reference for the responsible development and deployment of AI. The RAICA identifies and addresses present, emerging, and future humanitarian challenges in the field of AI while emphasizing its potential to improve lives globally, enhance the quality of work, inform better public policies, and drive greater efficiency within ecosystems.

The RAICA is structured around seven pillars that outline shared human centric commitments of the Member States toward:

1. Bridging the digital divide
2. Empowering underprivileged communities
3. Promoting digital development
4. Ensuring fairness and non-discrimination
5. Driving innovation in AI
6. Combatting climate change through AI
7. Enhancing international collaboration and cooperation in AI

Each pillar contains principles to maximize the benefits of AI while mitigating its potential risks. Among the principles included in this declaration are:

1 Human-Centricity & Human Agency

Pillar 1, 2 & 3:
Links to the declaration's focus on capacity building and reskilling programs

Pillar 5:
Reflects the emphasis on empowering individuals to drive digital development

2 Transparency & Explainability

Pillar 1 & 4:
Should be incorporated into the data diversity and algorithmic safeguards

3 Fairness, Non-discrimination & Equity

Pillar 1 & 4:
Safeguards against algorithmic discrimination

Pillar 1, 3, 4 & 7:
Diverse data requirements for different contexts and cultures

Pillar 1, 2, 3, 4, 5, 7:
Universal access to AI benefits

Pillar 6:
Protection of vulnerable communities in climate change context

4 Privacy & Data Governance

Pillar 1 & 4:
Partially implied through data handling requirements

Pillar 1 & 4:
Critical intersection with diverse data collection goals

5 Accountability & Technical Robustness

Pillar 4:
Implicit in monitoring and safeguard requirements

Pillar 6:
Important for environmental decision-making applications

2.3 DCO PRINCIPLES FOR ETHICAL AI

Acknowledging the complexity and rapid evolution of the global AI landscape, and building on the international agreements on the key principles and the Riyadh AI Call for Action Declaration, the DCO presents the **DCO Principles for Ethical AI (the Principles). The Principles** offer comprehensive guidance that respects human rights, fosters technological advancement, and addresses the multifaceted challenges posed by AI technologies.

The Principles emerge from extensive international dialogue and the DCO's deep commitment to ensuring that AI serves humanity's best interests. They reflect a balanced approach that acknowledges the diverse contexts of the DCO Member States while establishing a consistent ethical foundation.

The Principles, detailed below, are not intended to be restrictive but rather serve as a constructive roadmap for the development and governance of AI centered on human rights.

1 Accountability

2 Transparency and Explainability

3 Fairness and Non-discrimination

4 Privacy

5 Sustainability and environmental impact

6 Human-centred development and social benefit

7 Human-autonomy and oversight

1. Accountability

Ensure accountability by establishing clear responsibilities for AI outcomes and guaranteeing reliable, transparent performance throughout the AI system's lifecycle.

Accountability is a fundamental ethical principle that establishes clear responsibility for the development, deployment, and consequences of AI systems. It requires that AI technologies and their creators are transparently answerable for their performance, impacts, and potential risks to individuals and society.

This principle encompasses a comprehensive approach to responsible AI governance, mandating that AI-deploying organizations, developers, and deployers (collectively referred to herein as **AI Entities**), establish clear mechanisms for ownership, system reliability, and ethical oversight. Establishing responsibility means explicitly identifying the individuals, teams, and organizations accountable for AI system design, implementation, and outcomes. This involves **creating transparent ownership frameworks** that assign responsibility at every stage of AI system development and deployment and **developing mechanisms for tracking and addressing system performance** and potential negative impacts.

Reliability is intrinsically embedded within accountability, requiring AI systems to demonstrate consistent and predictable performance under diverse operational conditions.

To meet the **accountability principle** by ensuring **reliability and robustness in AI systems**, the AI Entities from both the public and private sectors must develop, implement, and adhere to continuous and **rigorous testing protocols, fault-tolerance mechanisms**, and **adaptive learning capabilities** that maintain system integrity, while also conducting **comprehensive risk assessments** to identify and mitigate potential vulnerabilities.

These processes to achieve this involve (i) rigorous and ongoing monitoring to identify vulnerabilities and performance issues; (ii) building redundancy into the system, allowing it to continue operating correctly even when parts fail; and (iii) evaluating the likelihood and impact of various risks, establishing mitigation strategies, and implementing safeguards to minimize adverse outcomes.

There must be clear identification of the individuals, teams, or organizations responsible for the outcomes of AI systems. This involves assigning **ownership** for not only the results these systems produce but also for ongoing performance management to maintain safe, ethical, and effective operations.

2. Transparency and Explainability

Promote transparency and explainability by ensuring that the processes, decisions, and underlying logic of AI systems are clearly communicated and accessible to relevant stakeholders, enabling informed understanding and fostering trust.

Transparency refers to providing clear and comprehensive disclosure about AI system usage. This principle requires AI Entities to openly communicate the nature of AI interactions, including the **types of data processed**, the system's **operational mechanisms**, and its **intended purpose**.

Relevant authorities should help ensure an appropriate balance between the need for transparency and commercial considerations, including trade secrets, intellectual property, and data subject rights.

Ultimately, the goal is to ensure that users are fully aware of when they are engaging with an AI system, empowering them to make conscious and informed decisions about their interactions.

Explainability complements transparency by focusing on the ability to communicate the reasoning behind AI-driven decisions in accessible and understandable terms. This principle acknowledges that not all users possess technical expertise, therefore demanding that AI systems articulate their decision-making processes in clear, straightforward language. The objective is to demystify complex technological processes, allowing users to comprehend how and why specific outcomes are reached, regardless of their technical background.

Transparency and explainability are foundational ethical principles that ensure AI systems are open, understandable, and accountable to those they affect.

- **Transparency** addresses the **'when'** (e.g., identifying instances where AI is being used) and the **'what'** (e.g., understanding what data is being used or processed) in AI applications and systems.
- **Explainability** focuses on the **'how'** (e.g., understanding how the system is designed to use the data and make decisions).

These principles demand that AI technologies provide clear, comprehensible information about their operation, purpose, and decision-making processes, enabling stakeholders to make informed choices and maintain trust in technological systems.

Under these principles, AI Entities must **clearly inform users when they are interacting with an AI system**. This disclosure should be proportional to the significance of the interaction, ensuring that AI systems are **clearly identified as non-human entities.** Where appropriate, AI Entities should provide options for human interaction as an alternative.

System transparency provides **meaningful information about the data types utilized, the system's operational principles, and its decision-making processes.** The AI Entities should communicate the intended purpose, application domain, and system limitations clearly and effectively. They must balance the need for transparency with the protection of intellectual property and trade secrets. The focus should remain on **providing information that meaningfully aids understanding, rather than sharing technically complex details** that may not serve this purpose.

Among the most debated risks presented by this technology are the AI-enabled spread of misinformation and the distortion of public discourse. If not properly designed and deployed with robust safeguards, AI systems could be used to generate or amplify the spread of false or misleading information, undermining the integrity of public debate and decision-making processes. AI Entities must implement **measures to ensure the transparency and veracity of the information** being generated or disseminated by AI their systems and proactively address the potential for AI-powered information manipulation.

AI systems also should provide explanations adapted to the expertise level of different stakeholders **using clear, simple terms to describe decision-making factors.** These explanations must enable affected individuals to understand and, when necessary, challenge outcomes. For **"black box" systems,** where direct technical explanations may be challenging, organizations should **implement alternative measures,** such as outcome-based explanations and robust quality assurance documentation.

3. Fairness and Non-discrimination

Uphold fairness and non-discrimination by designing, developing, and deploying AI systems that actively prevent bias, advance equity, and foster inclusive outcomes for all individuals and groups.

Fairness in AI refers to the equitable treatment of all individuals and groups in AI system outcomes, ensuring that benefits, risks, and costs are justly distributed across societies and cultures. It requires that **AI systems do not perpetuate or amplify existing biases** and that their outcomes are consistent across different demographic groups while **respecting cultural diversity, regional differences, and local values – ensuring equal access to AI benefits.** Furthermore, it ensures equal and inclusive access to AI benefits globally, fostering collaboration among nations to address systemic inequities and promote a fairer, more balanced distribution of AI's transformative potential.

Non-discrimination means that **AI systems must not create or contribute to unjust impacts on individuals or groups based on protected attributes,** such as gender, nationality, race, age, disability, ethnic origin, or cultural background. This entails taking proactive measures to prevent, identify, and address both direct and indirect forms of discrimination while promoting inclusive access to AI technologies.

Fairness and non-discrimination are fundamental principles in the development and deployment of AI systems. This is essential for protecting human rights and promoting social justice. These principles ensure that AI technologies serve all members of society equitably while **actively preventing the marginalization of vulnerable groups and the amplification of existing prejudices.** Equal respect for the moral worth and dignity of all human beings must be ensured, going beyond mere non-discrimination to **actively promote equality, inclusion, and cultural diversity.**

AI systems must be designed and deployed to ensure equitable treatment for all individuals and groups, regardless of their background, identity, gender, nationality, or circumstances. This commitment goes beyond technical fairness to **guarantee equal access to the benefits of AI across diverse demographics.** Actors must actively work to **eliminate barriers that might prevent certain groups from accessing or benefiting from AI technologies.** This includes considering economic, educational, linguistic, and cultural factors that could affect access and usage.

AI Entities must take proactive steps to **identify and mitigate bias in both data and algorithms.** This requires a **systematic assessment of training data, algorithm design, and system outputs for potential biases.** User-producer interaction becomes a critical mechanism for identifying and addressing potential biases. By involving diverse stakeholders throughout the AI system development process, AI Entities can gather insights from a range of perspectives, experiences, and potential systemic inequities. This collaborative approach involves soliciting feedback, conducting thorough impact assessments, and establishing channels for continuous dialogue, helping to uncover hidden biases before they are embedded into AI systems.

Proactive bias mitigation is essential, necessitating a systematic evaluation of training data, algorithm design, and system outputs. AI Entities must **implement robust methodologies to identify potential discriminatory patterns,** including both direct discrimination (where systems explicitly treat groups differently based on protected characteristics) and indirect discrimination (where seemingly neutral practices result in disadvantageous outcomes for certain groups). This requires continuous monitoring, regular audits, and a commitment to modifying systems when potential biases are detected.

This principle also directly applies to cultural life and values, which must be protected and promoted in AI system development and deployment. AI systems should be **designed with a deep awareness of different cultural contexts and values,** ensuring that they **enhance, rather than diminish, cultural richness and diversity.**

AI Entities must implement comprehensive **safeguards to prevent their systems from creating or exacerbating discriminatory outcomes.** This encompasses both direct and indirect discrimination. They must also maintain **the quality and integrity of data** throughout its lifecycle by regularly verifying and validating data accuracy, implementing processes to identify and address biases in data sets, establishing secure data access protocols, and documenting data handling procedures at every stage.

Furthermore, equal, inclusive, and non-discriminatory access to the benefits of AI must be ensured globally by fostering international collaboration to address systemic inequities, eliminate discriminatory practices, mitigate disparities, and promote a more equitable and balanced distribution of AI's transformative potential across regions and societies.

4. Privacy

AI systems must be designed and deployed to proactively safeguard individuals' privacy, encompassing not only data protection but also the broader aspects of personal autonomy, consent, and the right to control one's own information. This includes ensuring transparency in data usage, minimizing intrusive practices, and respecting individuals' right to privacy in both digital and physical spaces.

Privacy refers to the protection of individuals' physical, decisional, mental, and associational privacy in the face of increasingly sophisticated AI technologies. This principle recognizes that AI systems, through their extensive data collection and analysis capabilities, have the potential to significantly affect multiple aspects of personal privacy. It recognizes the potential for AI to intrude on physical movements, influence decision-making, analyze mental states, and map social connections, often without explicit consent or awareness of individuals.

At its core, this privacy principle calls for the responsible development and deployment of AI technologies that **respect personal boundaries and autonomy.** It emphasizes the need for transparency in AI operations, particularly in surveillance and decision-making systems, and advocates for safeguards against unwarranted intrusion into personal spaces, choices, thoughts, and relationships.

Cybersecurity is a significant concern closely linked to privacy. If not properly secured, AI systems can be vulnerable to hacking, data breaches, and other cyber threats, potentially exposing sensitive information or enabling malicious use of AI's capabilities. **Data privacy and protection are fundamental elements of ethical AI development and deployment,** requiring AI Entities to implement robust safeguards throughout the entire system lifecycle. This principle encompasses several key dimensions, built upon the established privacy frameworks, which must be carefully considered and addressed at every stage of AI implementation.

AI Entities must obtain **explicit, informed, and freely given consent** from individuals prior to collecting, using, or disclosing certain personal data for AI development and deployment. This is particularly true of sensitive data, such as that related to health, personal beliefs, or political affiliations.

"Personal data" refers to any information about an identified or identifiable living individual. The processing of personal data is generally governed by national laws and policies. When data is categorized as personal, obtaining explicit consent from individuals is usually required, in accordance with these legal frameworks. Anonymous data, which cannot be linked to an individual, falls outside the scope of this principle. However, pseudonymized data, which can potentially lead to the identification of a person when combined with additional information, remains within its scope.

Transparency in data practices requires clear communication about data collection methods, usage purposes, and access permissions. This transparency extends to making individuals aware of how their data contributes to AI system operations and decision-making processes. For example, AI Entities must respect individuals' decisions in opting out of certain communication and, to the greatest extent possible, from data sets that contain their personal information.

Data minimization requires AI Entities to **only collect data that is strictly necessary** for the intended purpose of the AI system. This approach not only respects individual privacy but also reduces potential risks associated with data breaches and misuse. AI Entities must clearly define and limit the scope of data collection to prevent unnecessary accumulation of personal information. Any change in the purpose of the processing requires a new assessment of whether the processing for the new purpose is compatible.

AI Entities must establish **robust data protection and governance frameworks** that implement privacy-by-design principles, ensuring privacy is integrated at every stage of the AI system development lifecycle. This includes conducting **regular Data Protection Impact Assessments (DPIAs)** to evaluate and mitigate privacy risks, which should be part of a broader AI impact assessment strategy.

The AI Entities should also implement **privacy-enhancing technologies (PETs),** such as differential privacy and zero-knowledge proofs, to preserve data utility while safeguarding individual privacy. This includes setting clear data access protocols, defining roles and responsibilities for data handling, implementing security measures to prevent unauthorized access, and conducting regular security audits and updates.

Cybersecurity policies are intrinsically connected to this principle, as they often encompass the necessary privacy protections. The AI Entities must protect the personal data they hold with appropriate safeguards against risks, such as loss, unauthorized access or other misuses. Such safeguards shall be proportional to the likelihood and severity of the potential harm and sensitivity of the data and the context in which it is held. These safeguards should also be subject to periodic review and re-assessment, accompanied by ongoing threat protection monitoring. In the case of a significant security breach affecting personal data, notifying authorities and/or affected individuals may help mitigate potential harm.

5. Sustainability and Environmental Impact

AI Entities must actively design, develop, and deploy AI systems with consideration for their environmental impact, ensuring that AI technologies contribute to sustainability by minimizing energy consumption, reducing carbon footprints, and promoting eco-friendly practices throughout their lifecycle. Additionally, AI Entities should leverage the technology to advance climate action, supporting initiatives aimed at addressing environmental challenges and fostering long-term ecological balance.

Sustainability and environmental impact represent a critical ethical dimension of AI development, recognizing both the technological challenges and transformative potential of AI in addressing global environmental concerns. This principle demands a holistic approach that balances the environmental costs of AI technologies with their capacity to drive climate action and sustainable development.

This principle aims to ensure that AI systems minimize their environmental footprint and promote sustainability. Acknowledging that AI systems can be used to provide solutions to optimize resource use, operating the systems requires significant computing power and energy consumption. Overall, AI solutions should be energy-efficient and environmentally and socially responsible.

AI's potential for **climate action** is profound and multidimensional. Advanced AI systems can provide unprecedented capabilities in climate modeling, environmental monitoring, resource optimization, and sustainable innovation. These technologies enable researchers and policymakers to develop precise climate prediction models, optimize renewable energy grid management, enhance environmental conservation strategies, improve resource allocation and consumption efficiency, support sustainable agricultural practices, and accelerate climate change mitigation research.

To operationalize environmental responsibility, AI Entities must implement concrete strategies that prioritize energy efficiency and sustainable computing. This involves selecting low-power hardware architectures, utilizing cloud computing services with demonstrated environmental credentials, transitioning to renewable energy sources for data centers and computational infrastructure, developing energy-efficient algorithms and computational methods, implementing advanced cooling technologies that reduce energy consumption, and conducting regular environmental impact assessments of AI systems.

Measuring and transparently reporting the environmental performance of AI technologies becomes crucial. AI Entities should establish clear metrics to monitor energy consumption, carbon emissions, and overall environmental impact while also setting progressive targets for reducing the ecological footprint of AI technologies over time.

Collaborative approaches are essential in addressing the complex environmental challenges associated with AI. This requires engaging with environmental scientists and sustainability experts, participating in industry-wide initiatives for green computing, as well as sharing best practices and environmental performance data and investing in breakthrough technologies that can reduce computational energy requirements.

By adhering to this principle, AI Entities can minimize AI's negative environmental impact while harnessing its potential to support global sustainability efforts. This approach ensures that the advancement of AI technology aligns with broader environmental and sustainable development goals.

6. Human-centered Development and Social Benefit

AI systems must be designed and deployed with a strong focus on human well-being, ensuring they contribute positively to social progress and provide tangible benefits to individuals and communities. This involves prioritizing the creation of AI technologies that safeguard humans from harm while addressing societal needs and fostering meaningful improvements in quality of life.

Human-centered development and social benefit refers to the work conducted to prioritize human well-being and societal benefits by aligning AI innovations with human rights, ethical standards, and social values. The principle also encompasses the **establishment of mechanisms to override, repair, or decommission AI systems that cause harm or exhibit undesired behavior.**

This principle underscores the necessity of aligning AI innovations with societal good and human values. AI Entities must **establish governance frameworks** to ensure AI systems adhere to this principle, with **clear oversight mechanisms in place.** To ensure AI systems are designed with human rights at the forefront, AI Entities should **conduct impact assessments** to (i) assess compliance with ethical guidelines, (ii) evaluate the system's potential positive and negative impacts on individuals and communities, and (iii) analyze the effectiveness of governance structures in maximizing societal benefits while minimizing risks.

Moreover, this principle **focuses on augmenting human capabilities** through AI development, rather than replacing human jobs. AI should be used to **empower human productivity, creativity, and decision-making,** with an emphasis on continuous human learning and development. While AI may automate certain tasks, its goal should be to complement and enhance human potential, rather than displace workers. AI Entities must assess the workforce impact, implement retraining initiatives, and ensure that AI benefits are distributed equitably so that the principle of human-centered development and social benefit is fully realized.

Considering the potential psychological impacts of AI, such as reduced social engagement or increased dependence on AI-driven services, it is crucial to proactively address these concerns. These impacts must be carefully considered, and appropriate safeguards should be implemented to mitigate them, ensuring that the increased use of AI enhances, rather than hinders, human flourishing.

AI systems must be developed and **deployed in alignment with fundamental human values, including human dignity, individual freedoms, and social justice.** This alignment requires embedding these values throughout the entire AI lifecycle, from initial design to ongoing operation. **AI Entities must ensure their AI systems respect and promote human rights while maintaining the capacity for meaningful human oversight and intervention.**

7. Human Autonomy and Oversight

AI systems must be designed and deployed in a way that preserves human autonomy and ensures robust oversight, enabling individuals to make informed decisions and intervene in AI-driven processes when necessary.

The concept of human autonomy emphasizes the importance of maintaining human control and decision-making authority over AI systems.

This principle aims to ensure that AI technologies support and enhance human capabilities, rather than replace or undermine human agency. Humans must retain the ability to manage the overall behavior of AI systems, decide when and how to use them, and intervene, adjust, or override AI decisions as needed. The concept of "**human autonomy by design**" recognizes the complex psychological dynamics, such as cognitive biases and automation bias, that can impair human decision-making. Addressing this requires intentional design strategies that actively promote critical thinking, skepticism, and conscious human intervention, ensuring that human oversight remains central to AI-assisted decision-making.

Humans shall always be able to control and supersede AI decisions. This is implemented via **supervisory human control,** which allows humans to monitor and take control when necessary to prevent errors or unintended consequences, and by **human-machine teaming,** which involves AI systems adapting to dynamic conditions using human inputs. Nevertheless, humans must ultimately reserve **the ability to correct, suspend, or shut down faulty AI systems, if needed.** This goes beyond simple override mechanisms, demanding comprehensive frameworks that provide transparent insights into AI decision-making processes, enable meaningful human review of automated decisions, establish clear thresholds for human intervention, and create accessible mechanisms for challenging or rejecting AI-generated recommendations.

The principle of human autonomy is especially critical in high-stakes domains with profound ethical implications, such as healthcare, judicial systems, and military applications. In these areas, where AI systems may influence life-and-death decisions, ensuring that human judgment remains central is crucial for safeguarding against the removal of human oversight. Special care must be taken to maintain human agency and accountability in these contexts to prevent unintended consequences and preserve ethical standards.

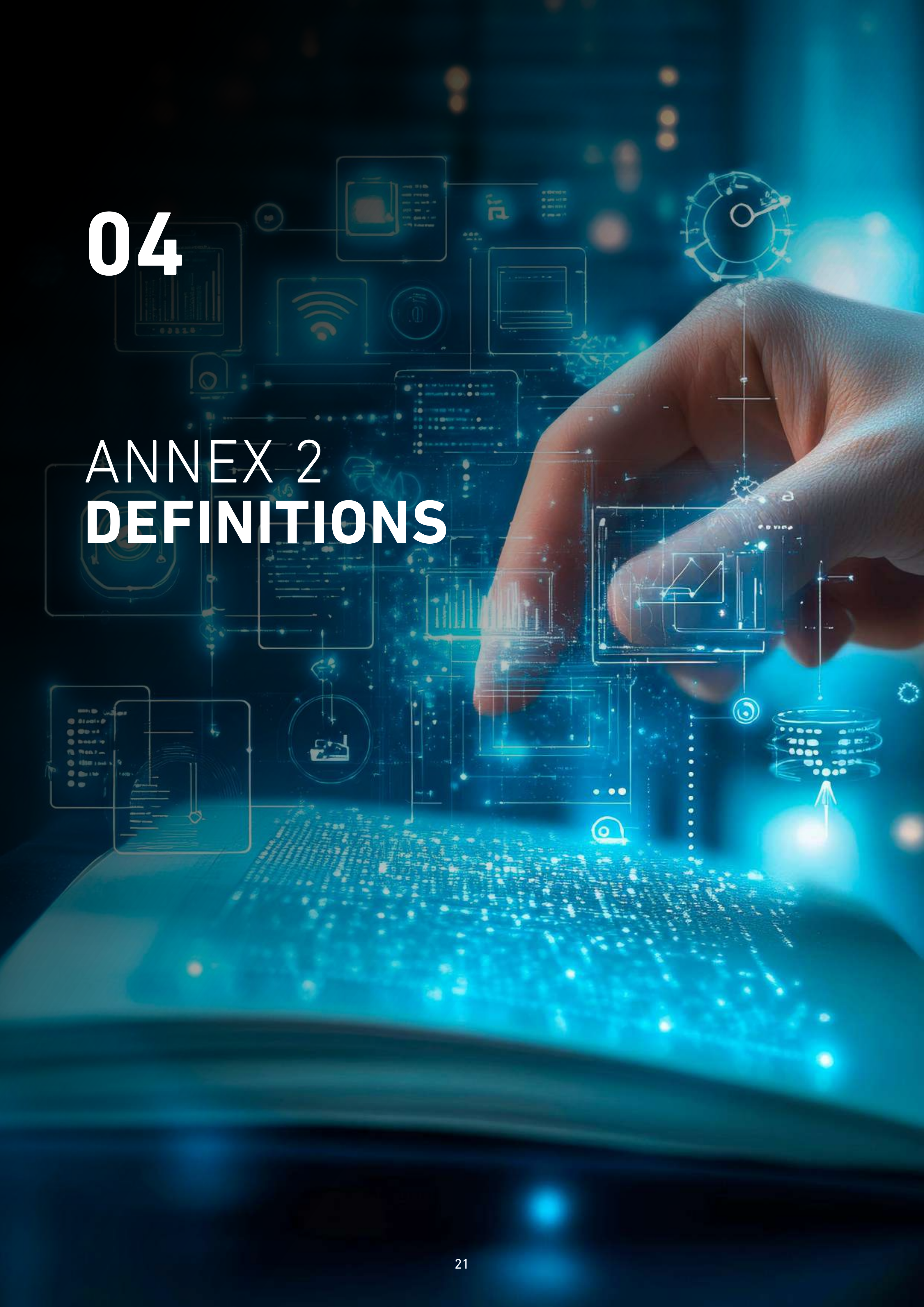
Potential risks associated with this principle include behavioral manipulation, where AI systems could be used to subtly influence or even control human decision-making and behavior, thereby undermining individual autonomy and free will. By exploiting their capacity to analyze and predict human behavior, AI systems could nudge or coerce individuals into making choices that may not be in their best interests or preferences. To mitigate these risks, safeguards must be put in place to prevent manipulative uses of AI, ensuring that humans retain the ultimate authority over the decisions that affect their lives and well-being.

Overall, the principle of human autonomy requires a careful balance between the capabilities of AI systems and the preservation of human agency. It requires the establishment of clear boundaries and control mechanisms to prevent AI from displacing or unduly influencing human decision-making, ensuring that humans retain ultimate control and authority over their actions and choices.

03

ANNEX 1 **ACRONYMS**

Acronym	Meaning
AI	Artificial Intelligence
ASEAN	Association of Southeast Asian Nations
AU	African Union
APEC-ABAC	Asia-Pacific Economic Cooperation - APEC Business Advisory Council
COE	Council of Europe
DCO	Digital Cooperation Organization
DPIAs	Data Protection Impact Assessments
EC	European Commission
G7	Group of Seven
G20	Group of 20
GPA	Global Privacy Assembly
ISO	International Organization for Standardization
ITU	International Telecommunication Union
OECD	Organisation for Economic Co-operation and Development
PETs	Privacy Enhancing Technologies
UN	United Nations
UNESCO	United Nations Educational, Scientific, and Cultural Organization



04

ANNEX 2
DEFINITIONS

1

1. AI Entities:

Organizations and individuals involved in the development and deployment of AI systems, including developers, deploying organizations, and deployers. These entities are responsible for ensuring AI systems align with ethical standards, maintain transparency and explainability, protect privacy, promote fairness and non-discrimination, consider environmental impact, focus on human-centered development, and preserve human autonomy and oversight. They must implement robust governance frameworks, conduct impact assessments, and maintain accountability for system reliability and societal impact.

2

2. Black Box Systems

AI systems whose internal workings and decision-making processes are not easily interpretable or explainable to humans. These systems require alternative measures for transparency, such as outcome-based explanations and robust quality assurance documentation, to maintain accountability and trust while protecting intellectual property rights.

3

3. Direct Discrimination

The explicit treatment of individuals or groups differently based on protected characteristics, such as gender, nationality, race, age, disability, ethnic origin, or cultural background, in AI systems. This form of discrimination is directly observable in system outputs or decisions.

4

4. Indirect Discrimination:

A form of bias where seemingly neutral practices or criteria in AI systems result in disadvantageous outcomes for certain groups or individuals based on protected characteristics. This occurs when system design or implementation leads to unintended discriminatory effects despite appearing impartial.

5

5. Personal Data:

Any information relating to an identified or identifiable living individual; therefore, this concept does not include data about natural persons who are deceased or data about legal persons like corporates. When data is categorized as personal, explicit consent from individuals is usually required for its collection and processing, in accordance with applicable legal frameworks.

6. Anonymous Data:

Information that cannot be linked to an identified or identifiable individual, either directly or indirectly. This type of data falls outside the scope of the Principles and personal data protection requirements as it cannot be used to identify specific individuals.

7. Pseudonymized Data:

Data that has been processed so that it can no longer be attributed to a specific individual without the use of additional information. However, it remains within the scope of data protection requirements as it can potentially lead to the identification of individuals when combined with additional information.

8. Data Protection Impact Assessments (DPIAs):

Systematic evaluations conducted to assess and mitigate privacy risks associated with AI systems. These assessments are part of a broader AI impact assessment strategy and help ensure compliance with data protection requirements throughout the AI system lifecycle.

9. Privacy-Enhancing Technologies (PETs):

Technical solutions and tools, such as differential privacy and zero-knowledge proofs, that preserve data utility while safeguarding individual privacy. These technologies help maintain data protection while allowing AI systems to process and analyze information effectively.

10. Human-Machine Teaming:

A collaborative processing approach requiring human oversight integration into decision-making when involving high risks for individuals’ rights, particularly with complex algorithms having direct strong impacts on the individuals’ sphere.

05

ANNEX 3
AI DEFINITIONS ACROSS MULTILATERAL ORGANIZATIONS

AI DEFINITIONS ACROSS MULTILATERAL ORGANIZATIONS

International and multilateral organizations demonstrate distinct approaches to defining AI, reflecting their institutional perspectives and priorities. Analysis of these definitions reveals three key patterns:

1. Task-Oriented Definitions:

Organizations like the Asia-Pacific Economic Cooperation – APEC Business Advisory Council (APEC-ABAC), the International Organization for Standardization (ISO), and the Organisation for Economic Co-operation and Development (OECD) focus on AI's functional capabilities, defining it through its ability to perform specific tasks, make predictions, or generate outputs based on human-defined objectives. These definitions emphasize practical applications and measurable outcomes.

2. Process-Based Definitions:

The United Nations Educational, Scientific, and Cultural Organization (UNESCO) and the European Commission frame AI in terms of how systems process information and interact with their environment. Their definitions highlight the autonomous nature of AI systems and their ability to analyze and respond to environmental inputs.

3. Human-Comparative Frameworks:

The International Telecommunication Union (ITU) offers a comprehensive definition, bridging both functional and conceptual aspects. It explicitly draws parallels between AI systems and human cognitive processes, emphasizing capabilities like learning, decision-making, and problem-solving while acknowledging AI's role in tackling complex challenges.

This diversity in definitions reflects the multifaceted nature of AI technology and the different regulatory and developmental priorities of these organizations. While some focus on technical capabilities and practical applications, others emphasize the relationship between AI systems and human intelligence, suggesting different approaches to governance and implementation.

Table 1. AI Definitions in International and Multilateral Organizations

Organization	AI Definition
Asia-Pacific Economic Cooperation (APEC) Business Advisory Council (ABAC)	Systems and models that can perform tasks requiring human intelligence. What distinguishes AI is its capacity for autonomous learning. It could take in the data fed to it and teach itself to, for example, solve mathematical conjectures or understand native human speech.[4]
European Commission (EC)	Systems that display intelligent behavior by analyzing their environment and taking actions, with some degree of autonomy, to achieve specific goals.[5]
International Organization for Standardization (ISO)	Engineered system that generates outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives.[6]
International Telecommunication Union (ITU)	Computerized system that uses cognition to understand information and solve problems.[7] The ability of a computer or a computer-enabled robotic system to process information and produce outcomes in a manner similar to the thought process of humans in learning, decision-making, and problem-solving. In a way, the goal of AI systems is to develop systems capable of tackling complex problems in ways similar to human logic and reasoning.[8]
Organization for Economic Cooperation and Development (OECD)	A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.[9]
United Nations Educational, Scientific, and Cultural Organization (UNESCO)	Systems that have the capacity to process data and information in a way that resembles intelligent behavior, and typically include aspects of reasoning, learning, perception, prediction, planning, or control.[10]

Source: Internal research

In this context, it seems that the more thorough definitions used by international organizations may be better suited to shaping governance frameworks that are not only detailed but also actionable, avoiding the ambiguities that can arise from more simplistic or vague descriptions.[11]

An effective AI definition must balance two key elements: technical precision that remains relevant across technological evolution and ethical principles that guide responsible development and deployment.

06

ANNEX 4 **GLOBAL AI PRINCIPLES**

Table 2. Multilateral AI Frameworks

Organisation	Principles	Approach
African Union (AU) Continental AI Strategy (2024)[12]	Human-centricity, transparency, accountability, fairness, human rights, privacy, equitable access, and minimization of bias, discrimination, and societal harms.	Focuses on AI's potential to boost Africa's socioeconomic development and Agenda 2063, promoting ethical AI adoption, local capacity-building, and African-centric solutions. It emphasizes regional cooperation and positions Africa as a key player in global AI governance.
ASEAN Guide on AI Governance and Ethics (2023)[13]	Transparency and explainability, fairness and equity, security and safety, human-centricity, privacy and data governance, accountability and integrity, and robustness and reliability	Practical advice for organizations in the region interested in designing, developing, and deploying traditional AI technologies for commercial, non-military, or dual-use purposes.
Council of Europe (COE) Convention 108+ (2019)[14]	Human rights, democracy, rule of law, transparency, and data privacy.	The COE promotes AI frameworks that protect human rights and privacy. Convention 108+ extends data protection to AI, while the Ad Hoc Committee on AI (CAHAI) explores legal frameworks for ethical AI use, particularly regarding facial recognition.
Council of Europe (COE) Framework Convention on artificial intelligence and human rights (2024)[15]	Human rights, democracy, rule of law, and transparency.	Global legally binding instrument focused on the protection of human rights, democracy, and the rule of law. Designed on a risk-based approach.
European Commission Ethical Guidelines for Trustworthy AI (2019)[16]	Human agency, technical robustness, transparency, and non-discrimination.	Emphasis on lawful, ethical AI development, providing a foundation for ongoing regulations such as the AI Act. The Guidelines guide both private and public sectors in aligning with fundamental rights.
European Union (EU) AI Act (2024)[17]	Safety, transparency, accountability, and non-discrimination.	A risk-based approach to categorize AI systems, ensuring safety and transparency for high-risk sectors like healthcare. The Act provides robust compliance standards while fostering innovation.
G7 Hiroshima AI Process (2023)[18]	Human-centric AI, transparency, accountability, and security.	Focuses on generative AI governance, emphasizing transparency and accountability in AI systems.
G20 AI Principles (2019)[19]	Human rights protection, transparency, explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, biases, privacy, and data protection.	Encourage international cooperation on human-centric AI, reaffirmed in 2023. The G20 aims to use AI to solve global challenges responsibly while ensuring transparency and innovation.

Organisation	Principles	Approach
Global Privacy Assembly (GPA) Declaration on Ethics and Data Protection in AI (2018)[20]	Privacy, fairness, accountability, transparency, and human rights.	Promotes fairness and accountability, calling for stricter governance to mitigate risks to privacy and fundamental rights.
OECD AI Principles (2019, updated in 2024)[21]	Inclusive growth, sustainable development and well-being; human rights and democratic values, including fairness and privacy; transparency and explainability; robustness, security and safety; and accountability.	A global standard for AI policy, updated in 2023 to include generative AI. The OECD AI Policy Observatory supports analysis and alignment of global AI governance efforts.
United Nations Principles for the Ethical Use of Artificial Intelligence (2022)[22]	Do no harm; defined purpose, necessity, and proportionality; safety and security; fairness and non-discrimination; sustainability; the right to privacy, data protection, and data governance; human autonomy and oversight; transparency and explainability; responsibility and accountability; and inclusion and participation.	Guide the use of AI throughout its lifecycle within United Nations system entities. It should be considered alongside other relevant policies and international laws.
United Nations (UN) Roadmap for Digital Cooperation (2020)[23]	Human rights do no harm, transparency, safety, accountability, and inclusion.	Calls for global AI governance based on building capacity, especially in developing nations.
UNESCO Recommendation on the Ethics of Artificial Intelligence (2021)[24]	Human dignity, inclusion, environmental sustainability, and transparency.	Promotes ethical AI use aligned with human rights and sustainability goals, aiming for inclusive, transparent, and accountable AI development.
United Nations Global Digital Compact (2024)[25]	Digital inclusion, security, transparency, equity, and human-centricity.	Focuses on leveraging AI to support the Sustainable Development Goals (SDGs), promoting inclusive, human-centric AI governance at a global scale.

Source: Internal research

ENDNOTES

[1] https://www.merriam-webster.com/dictionary/principle

[2] https://dictionary.cambridge.org/dictionary/english/principle

[3] Digital Cooperation Organization (2024) Riyadh AI Call for Action (RAICA), https://dco.org/wp-content/uploads/2024/06/Riyadh-AI-Call-for-Action-RAICA-Declaration.pdf

[4] APEC (2022) Artificial Intelligence in Economic Policymaking, www.apec.org/docs/default-source/publications/2022/11/artificial-intelligence-in-economic-policymaking/222_psu_artificial-intelligence-in-economic-policymaking.pdf?sfvrsn=341777ad_2

[5] European Commission (2018) A Definition of AI: Main Capabilities and Scientific Disciplines, https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

[6] International Organization for Standardization (2022) ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en

[7] International Telecommunications Union (2021) Recommendation ITU-T M.3080 (02/2021), https://www.itu.int/rec/T-REC-M.3080-202105-I!Err1

[8] International Telecommunications Union (2018) Policy Considerations for AI Governance, www.itu.int/en/ITU-T/studygroups/2017-2020/03/Documents/Shailendra%20Hajela_Presentation.pdf

[9] Organisation for Economic Co-operation and Development (2019) Artificial intelligence and responsible business conduct, https://mneguidelines.oecd.org/RBC-and-artificial-intelligence.pdf

[10] UNESCO (2021) Recommendation on the Ethics of Artificial Intelligence, www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

[11] Brookings (2023) Interpreting the ambiguities of Section 230, www.brookings.edu/articles/interpreting-the-ambiguities-of-section-230

[12] African Union (AU) Continental AI Strategy, https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy

[13] ASEAN Guide on AI Governance and Ethics (2023), https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf

[14] Council of Europe (COE) Convention 108+ (2019), https://www.coe.int/en/web/data-protection/convention108-and-protocol

[15] Council of Europe (COE) Framework Convention on artificial intelligence and human rights (2024), https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

[16] European Commission Ethical Guidelines for Trustworthy AI (2019), http://European Commission Ethical Guidelines for Trustworthy AI

[17] European Union (EU) AI Act (2024), https://artificialintelligenceact.eu/ai-act-explorer/

[18] G7 Hiroshima AI Process (2023), https://www.soumu.go.jp/hiroshimaaiprocess/en/documents.html

[19] G20 AI Principles (2019), https://www.consilium.europa.eu/media/66739/g20-new-delhi-leaders-declaration.pdf

[20] Global Privacy Assembly (GPA) Declaration on Ethics and Data Protection in AI (2018), https://globalprivacyassembly.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf

[21] OECD AI Principles (2019, updated in 2024), https://oecd.ai/en/ai-principles

[22] United Nations Principles for the Ethical Use of Artificial Intelligence (2022), https://unsceb.org/principles-ethical-use-artificial-intelligence-united-nations-system#:~:text=It%20is%20intended%20to%20be,data%20governance%3B%20human%20autonomy-%20and

[23] United Nations (UN) Roadmap for Digital Cooperation (2020), https://www.un.org/en/content/digital-cooperation-roadmap/

[24] UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), https://www.unesco.org/en/articles/unescos-recommendation-ethics-artificial-intelligence-key-facts?hub=32618

[25] United Nations Global Digital Compact (2024), https://www.un.org/techenvoy/global-digital-compact

Digital
Cooperation
Organization

Follow us on

@dcorg | www.dco.org