unesco

KEY FACTS

RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

Published in 2023 by the United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France

SHS/2023/PI/H/1 Rev.

Cover photo:
© metamorworks / Shutterstock.com*

Graphic Design : Yun Chih (Nicole) LU

Printed by UNESCO

Printed in France

Table of Contents

1. WHY A GLOBAL RECOMMENDATION IS NEEDED?

The rapid rise in artificial intelligence (AI) has created many opportunities globally, from facilitating healthcare diagnoses to enabling human connections through social media and creating labour efficiencies through automated tasks.

However, these rapid changes also raise profound ethical concerns. These arise from the potential AI systems have to embed biases, contribute to climate degradation, threaten human rights and more. Such risks associated with AI have already begun to compound on top of existing inequalities, resulting in further harm to already marginalised groups.

Recognising the urgency of this challenge, UNESCO published the first - ever global standard on AI ethics – the **Recommendation on the Ethics of Artificial Intelligence** (hereafter "the Recommendation"). All 193 Member States adopted this framework in November 2021.

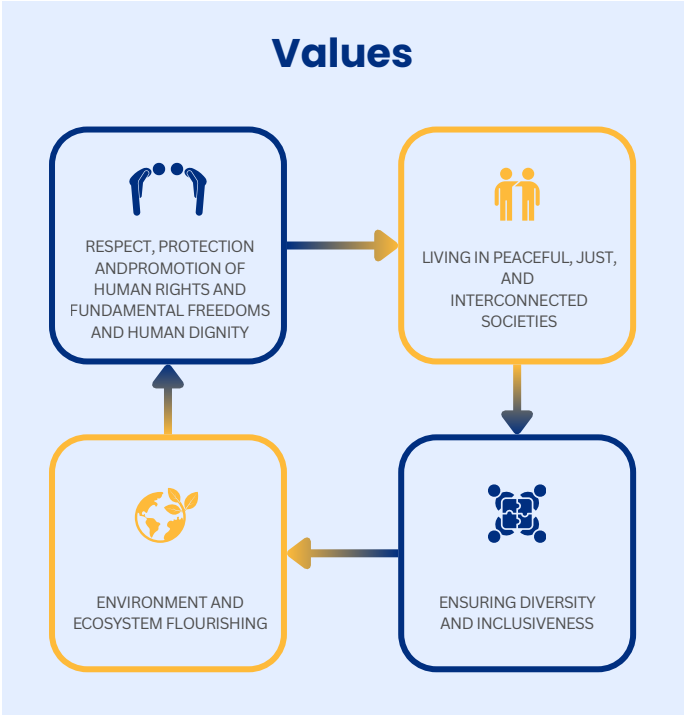
The Recommendation's cornerstone is the protection of human rights and dignity, grounded in principles such as transparency and fairness, while always remembering the importance of human oversight of AI systems.

However, what makes the Recommendation exceptionally applicable are its extensive Policy Action Areas, which allow policymakers to translate the core values and principles into action with respect to data governance, environment and ecosystems, gender, education and research, and health and social well-being, among many other spheres.

BREAKTHROUGH PROVISION

A Dynamic Understanding of AI
The Recommendation interprets AI broadly as systems with the ability to process data in a way that resembles intelligent behaviour. This is crucial as the rapid pace of technological change would quickly render any fixed, narrow definition outdated and make future-proof policies infeasible.

Central to the Recommendation are **four core values** which lay the foundations for AI systems that work for the good of humanity, individuals, societies and the environment:



2. A HUMAN RIGHTS APPROACH TO AI

Ten core principles lay out a human-rights centred approach to the Ethics of AI.

1 PROPORTIONALITY AND DO NO HARM

The use of AI systems must not go beyond what is necessary to achieve a legitimate aim. Risk assessment should be used to prevent harms which may result from such uses.

BREAKTHROUGH PROVISION

No use of AI for social scoring or mass surveillance

The Recommendation is the first international normative instrument that contains a provision against using AI systems for social scoring and mass surveillance purposes.

2 SAFETY AND SECURITY

Unwanted harms (safety risks) as well as vulnerabilities to attack (security risks) should be avoided and addressed by AI actors.

KEY CONCEPT

AI actors and the AI life cycle

AI actors are any actors (natural or legal persons) involved in any stage of the AI life cycle, ranging from research, design, and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, end-of-use, disassembly and termination.

3 RIGHT TO PRIVACY AND DATA PROTECTION

Privacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworksshould also be established.

CASE 1

Up close and personal

The data that we share online can have an impact on our individual privacy, often unbeknownst to us. Individuals' behaviour online, including abstract information such as patterns of social media likes and scrolling speeds, may be modelled and used as a basis for targeted advertising or behavioural manipulation.

4 MULTI-STAKEHOLDER AND ADAPTIVE GOVERNANCE AND COLLABORATION

International law and national sovereignty must be respected in the use of data, meaning States can regulate the data generated within or passing through their territories. Additionally, participation of diverse stakeholders is necessary for inclusive approaches to AI governance.

5 RESPONSIBILITY ANDACCOUNTABILITY

AI systems should be auditable and traceable. There should be oversight, impact assessment, audit and due diligence mechanisms in place to avoid conflicts with human rights norms and threats to environmental well-being.

CASE 2

Automated rejection

When applying for a loan, your bank may use AI to make an automated assessment of your finances and determine if your application will be approved. If these decisions are taken without human oversight and accountability, the consequences can be significant. First, an AI system that is not checked by a human may make a mistake. Second, there is no clear appeals process if there is nobody who can take ultimate responsibility for the decision.

6 TRANSPARENCY AND EXPLAINABILITY

The ethical deployment of AI systems depends on their transparency and **explainability**. For example, people should be made aware when a decision is informed by AI. The level of transparency and explainability should be appropriate to the context, as there may be tensions between transparency and explainability and other principles such as privacy, safety and security.

KEY CONCEPT

Explainability
Explainability refers to the ability to understand and communicate how an AI system arrives at a decision or output. It involves providing meaningful information about the factors, logic or processes that influenced the result, in a way that is understandable to the people affected by it and appropriate to the context.

7 HUMAN OVERSIGHT AND DETERMINATION

Member States should ensure that AI systems do not displace ultimate human responsibility and accountability.

© FOTO Eak / shutterstock.com

8 SUSTAINABILITY

AI technologies should be assessed for their impacts on sustainability, including environmental, social and economic dimensions, and their contribution to achieving the UN Sustainable Development Goals.

9 AWARENESS AND LITERACY

Public understanding of AI and data should be promoted through open and accessible education, civic engagement, digital skills and AI ethics training, media and information literacy

10 FAIRNESS AND NON-DISCRIMINATION

AI actors should promote social justice, fairness, and non-discrimination while taking an inclusive approach to ensure AI's benefits are accessible to all.

KEY CONCEPT

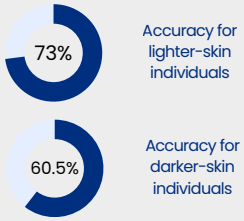
Algorithmic bias

Algorithmic bias refers to systematic and potentially unfair differences in the outputs or performance of an AI system across individuals or groups. It can arise from biased or unrepresentative data, the design of algorithms, or the contexts in which AI systems are developed and deployed.

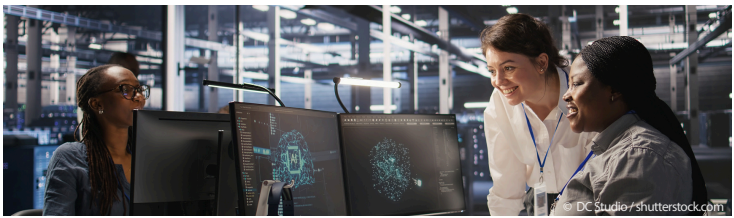
CASE 3

More than meets the eye

An AI system may perform differently across groups when its training data do not adequately represent them. In diabetic retinopathy diagnostics, one study found lower accuracy for individuals with darker skin tones than for those with lighter skin tones.



Modified from Burlina, P., Joshi, N., Paul, W., Pacheco, K. D., & Bressler, N. M. (2021), "Addressing artificial intelligence bias in retinal diagnostics."



3. ACTIONABLE POLICIES

Key policy areas make clear arenas where Member States can make strides towards responsible developments in AI.

While values and principles are crucial to establishing a basis for any ethical AI framework, recent movements in the field have emphasised the need to move beyond high-level principles and toward practical strategies.

The Recommendation does just this by setting out eleven key areas for policy actions, as summarised in the figure below. In this booklet, we explore five of these in greater depth.

Policy Areas

ETHICAL GOVERNANCE AND STEWARDSHIP

AI governance mechanisms should be inclusive, transparent, multidisciplinary, multilateral and multi-stakeholder. In other words, communities impacted by AI must be actively involved in its governance in addition to experts across a range of disciplines. Additionally, governance must extend beyond mere recommendations to include anticipation, enforcement and redress.

BREAKTHROUGH PROVISION

Preventing harm
The Recommendation stresses that AI governance cannot stop once risks and impacts have been identified. Instead, all identified harms must be investigated and addressed so that impacted communities have the right to redress.

ECONOMY AND LABOUR

Member States should consider and attempt to regulate the impact of AI systems on the labour market. AI-related studies should be made a core skill at all educational levels to help close the skill gap. It will boost market competition and ensure consumer protection on a national and international scale.

DATA POLICY

Member States should implement mechanisms for effective data governance strategies to ensure individual privacy while ensuring adequate data collection and means to regulate its use.

BREAKTHROUGH PROVISION

Data for training
The Recommendation prompts and facilitates the use of quality and robust datasets for the training, development, and use of AI. This includes the creation of gold standard datasets or open and trustworthy datasets.

HEALTH AND SOCIAL WELLBEING

Member States should aim to deploy AI to improve health and tackle global health risks. AI in healthcare and mental healthcare should be regulated to be safe, effective, efficient and medically proven. Additionally, Member States should encourage research into the impact of AI on mental health and wellbeing.

BREAKTHROUGH PROVISION

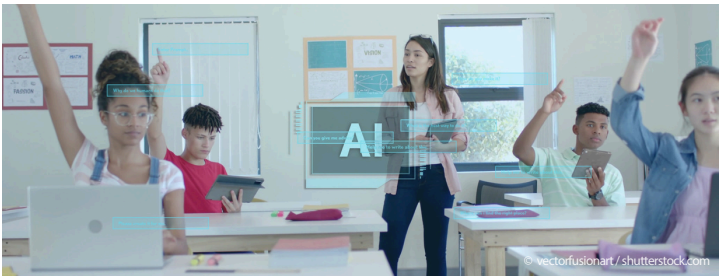
Converging technologies
Emerging technologies are converging with one another, creating distinct ethical concerns. The Recommendation includes provisions for such convergence, including neurotechnology and brain-computer interfaces.

EDUCATION AND RESEARCH

Member States should provide adequate AI literacy education to the public, including awareness programmes on data. In doing so, the participation of marginalised groups should be prioritised. Member States should also encourage research initiatives on ethical AI.

BREAKTHROUGH PROVISION

Ethics skills
In the Recommendation, AI ethics, communication, and teamwork skills are recognised as priorities alongside other widely recognised skills such as basic literacy, numeracy, coding and digital skills.



GENDER

Member States should maximise the potential AI has to contribute to gender equality while preventing any potential for AI to exacerbate gender gaps. There should be dedicated funds for policies which support women and girls to make sure they are not left out. For example, investment in women in STEM careers.

CASE 4

Invisible biases, visible inequalities

AI systems can reproduce or amplify existing gender biases when the data, design choices or objectives underlying them reflect unequal patterns in society. Even systems designed to be gender-neutral can produce unequal outcomes.

20% more men saw a gender-neutral STEM career ad than women

Lambrecht, A., & Tucker, C. E. (2019). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. Management Science, 65(7), pp. 2966-81.

ENVIRONMENT AND ECOSYSTEMS

Member States and businesses should assess direct and indirect environmental impacts of AI systems, including their carbon footprint, energy consumption and raw material extraction. Where necessary, Member States should also introduce incentives to ensure AI solutions are used to support the prediction, prevention, control and mitigation of climate-related problems.

CASE 5

Power-hungry algorithms

Many computations are needed to train large AI models based on big data sets. For instance, a common AI training model can emit more than 626,000 pounds of CO2 equivalent - about five times the lifetime emissions of the average car - including the manufacture of the car itself.

Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for modern deep learning research.

4. IMPLEMENTING THE RECOMMENDATION

Part of effectively implementing the Recommendation includes providing Member States with actionable resources, including two practical tools.

1 READINESS ASSESSMENT METHODOLOGY (RAM)

Description:

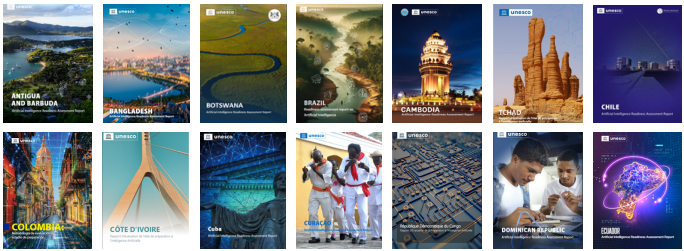
The RAM is a comprehensive diagnostic tool that helps assess a country's level of AI Readiness across the Recommendations' 10 Policy Areas and an Infrastructure Snapshot. This inclusive process has proved to stress-test existing AI laws, policies and strategies, help reduce policy blind spots, and act as a coordination mechanism across national AI ecosystems.

Purpose:

This methodology provides Member States with incentives to boost their AI policy in the form of evidence, prompting them to invest in concrete areas which require further development and encouraging them to collect data when it is lacking.

Deployed in 75+ Member States

20,000 stakeholders consulted



© UNESCO

2 ETHICAL IMPACT ASSESSMENT (EIA)

Description:

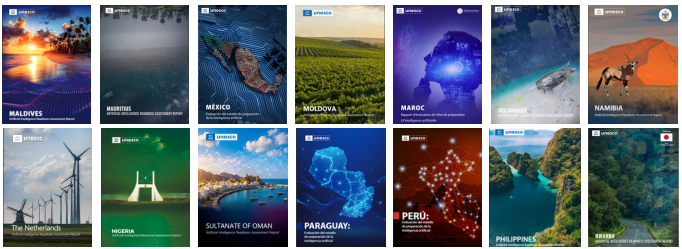
The EIA is an action-oriented, step-by-step process to help government officials, organisations, and communities ensure that their procured and in-house developed AI systems are aligned with the values and principles of the Recommendation.

Purpose:

EIA provides an opportunity to reflect on the potential impacts of an AI project and to identify the needed harm-prevention actions.

The EIA addresses questions such as:

- What limitations have been placed on the scope of this project to ensure it remains proportional to the stated objective?
- What measures were put in place to ensure the safety and security of the data processed by the AI system?
- Who is most likely to be adversely affected by this AI system?



5. JOIN US

Taken together, these values, principles, policy action areas, and practical methodologies of the Recommendation provide guidance to Member States on how to foster responsible AI innovation and equitable distribution of its benefits. UNESCO is working with governments, the private sector, academic institutions, and civil society organisations to translate the Recommendation into policies and actions. The multidimensional implementation strategy includes community engagement initiatives such as:

Global Observatory on the Ethics of AI, an innovative digital platform which enables the sharing of best practices and dialogue between countries, acting as a one-stop shop for the latest analysis on the ethical development and use of AI around the world.

Global Forum on Ethics of AI, a high-level annual event to advance the state-of-the-art knowledge of the challenges raised by AI technologies and to facilitate peer-to-peer learning among governments and other stakeholders.

Partnership and Projects, a series of collaborations with public and private stakeholders, such as the European Commission, the Government of Japan, and the Patrick J. McGovern Foundation, to enhance capacity-building in Member States.

Patrick J McGovern FOUNDATION

From the People of Japan

Funded by the European Union

The implementation strategy also includes such networks and platforms as:

AI Ethics Experts Without Borders (AIEB) Network

A flexible facility of experts for deployment in Member States on a needs basis to assist in the implementation of the Recommendation and the application of the capacity-building tools.

Women4EthicalAI (W4EAI) Network

A platform for leaders in industry, government, and civil society, driving transformations towards gender equality in and through AI.

Business Council for Ethics of AI

A platform for companies to come together, exchange experiences, and promote ethical practices within the AI industry.

Global Network of CSOs and Academic Institutions

A network to enhance the voice of CSOs and academia in the conversation of global governance.

Global Network of AI Supervisory Authorities (GNAIS)

A network to enhance regulatory expertise and cooperation between competent authorities on AI.

Coalition for Linguistic Diversity in AI

A multi-stakeholder coalition developed in partnership with the government of Iceland to safeguard linguistic diversity in Artificial Intelligence.

This brochure is a summary of **"Recommendation on the Ethics of Artificial Intelligence".**

For more information and a more in-depth look at the recommendation can be found by **clicking on the following link** or by **scanning the QR code.**

unesco

United Nations
Educational, Scientific
and Cultural Organization

Sciences Sector

7 Place de Fontenoy 75007 Paris, France

ai-ethics@unesco.org

on.unesco.org/EthicsAI

Follow us

@UNESCO #AI #AIEthics