UNESCO

Key facts
UNESCO's Recommendation on
the Ethics of Artificial Intelligence

Adopted on 23 November 2021

Table of Contents

MESSAGE FROM GABRIELA RAMOS
Assistant Director-General, Social and Human Sciences, UNESCO

With its unique mandate, UNESCO's Social and Human Sciences Sector has led the international effort to ensure that science and technology develop with strong ethical guardrails for decades.

Be it on genetic research, climate change, or scientific research, we have delivered global standards to maximize the benefits of the scientific discoveries, while minimizing the downside risks, ensuring they contribute to a more inclusive, sustainable, and peaceful world. We have also identified frontier challenges in areas such as the ethics of neurotechnology, on climate engineering, and the internet of things. This has been done with the unvaluable support of our expert committees, the Committee on the Ethics of Science and Technology and the International Bioethics Committee.

In no other field is the ethical compass is more relevant than in artificial intelligence. These general-purpose technologies are re-shaping the way we work, interact, and live. The world is set to change at a pace not seen since the deployment of the printing press six centuries ago. AI technology brings major benefits in many areas, but without the ethical guardrails, it risks reproducing real world biases and discrimination, fueling divisions and threatening fundamental human rights and freedoms. AI business models are highly concentrated in just few countries and a handful of firms — usually developed

UNESCO / C. Alix

in male-dominated teams, without the cultural diversity that characterizes our world. Contrast this with the fact that half of the world's population still can't count on a stable internet connection.

To correct this, under the leadership of UNESCO's Director-General Audrey Azoulay, and a clear mandate by our Member States, we elaborated the world's most comprehensive international framework to shape the development and use of AI technologies. The Recommendation on the Ethics of Artificial Intelligence was adopted by acclamation by 193 Member States at UNESCO's General Conference in November 2021. This comprehensive instrument was two years in the making

and the product of the broadest global consultation process of expert, developers, and other stakeholders from all around the world. It was an honour to spearhead this effort and support Member States in accomplishing this important task.

The Recommendation emphasizes who should be in control of these technologies. It makes a strong call to governments around the world to establish the necessary institutional and legal frameworks to govern these technologies and ensure they contribute to the public good. It clearly signals the end of the "self-regulatory model" that has prevailed, prioritizing commercial and geopolitical objectives over people for too long.

The Recommendation establishes a set of values in line with the promotion and protection of human rights, human dignity, and environmental sustainability. It advances essential principles such as transparency, accountability, and the rule of law online.
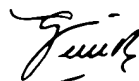
It also includes concrete policy chapters that call for better governance of data, gender equality, and important aspects of AI applications on education, culture, labour markets, the environment, communication and information, health and social well-being, and the economy. Unlike other international instruments, the Recommendation includes monitoring and evaluation chapters and means for implementation in the form of a Readiness Assessment and the Ethical Impact Assessment to ensure real change on the ground.

UNESCO is now focused on implementing the Recommendation, and we are piloting the tools, establishing regional roundtables for peer learning, and development networks of partners around the world, such as the AI Ethics Experts Without Borders network and the Women 4 Ethical AI network. The first Global Forum on the Ethics of Artificial Intelligence took place in the Czech Republic in the context of the country's Presidency of the Council of the European Union, and work is being done to launch the Observatory on the Ethics of AI. Given that the private sector is the major contributor to these developments, engagement with leading firms has also been established, as well as interactions with the civil society. It is clear that ensuring ethical AI is everybody's business.

This level of commitment and engagement from Member States, the private sector, civil society, and academic institutions is a sign that the ethical framework we have developed was long overdue. Now it is in our hands to realize its full potential and ensure that the new era of AI delivers the progress and solutions that we are all hoping for.

Gabriela Ramos
🐦 @gabramosp

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"). This framework was adopted by all 193 Member States in November 2021.

The protection of human rights and dignity is the cornerstone of the Recommendation, based on the advancement of fundamental principles such as transparency and fairness, always remembering the importance of human oversight of AI systems.

BREAKTHROUGH PROVISION

A Dynamic Understanding of AI

The Recommendation interprets AI broadly as systems with the ability to process data in a way which 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.

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 wellbeing, among many other spheres.

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:

VALUES

1
RESPECT, PROTECTION AND PROMOTION OF HUMAN RIGHTS AND FUNDAMENTAL FREEDOMS AND HUMAN DIGNITY

2
LIVING IN PEACEFUL, JUST, AND INTERCONNECTED SOCIETIES

4
ENVIRONMENT AND ECOSYSTEM FLOURISHING

3
ENSURING DIVERSITY AND INCLUSIVENESS

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 frameworks should 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.

SeventyFour / Shutterstock.com

5 RESPONSIBILITY AND ACCOUNTABILITY

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 wellbeing.

CASE 2 Automated rejection

When applying for a loan, it is possible that your bank uses 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

The term 'black box' has been used to describe AI wwsystems which are opaque and difficult to interpret. 'Explainability' requires that the logic behind algorithmic decision-making can be fully interpreted by experts and that this logic can be explained to users in accessible language.

7 HUMAN OVERSIGHT AND DETERMINATION

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

8 SUSTAINABILITY

AI technologies should be assessed against their impacts on 'sustainability', understood as a set of constantly evolving goals including those set out in the UN's 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.

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

KEY CONCEPT Machine learning

Machine learning is an application of AI that enables systems to learn from data and to improve without being explicitly programmed, thus improving their accuracy over time.
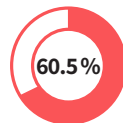
CASE 3

More than meets the eye

Machine learning algorithms can be biased, often performing worse for certain sub-groups defined by protected attributes. This has been shown to be the case in diabetic retinopathy diagnostics:

73%

Accuracy for lighter-skin individuals

60.5%

Accuracy for darker-skin individuals

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 AI ethics 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.
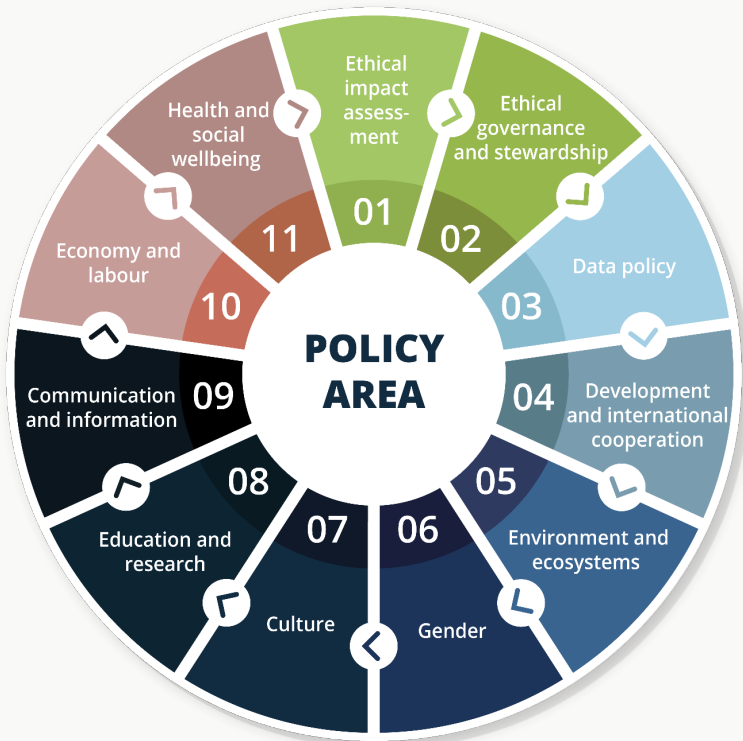
ETHICAL GOVERNANCE AND STEWARDSHIP

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.

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.

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.

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 data sets or open and trust-worthy datasets.

POLICY AREAS

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.

Ruslana Iurchenko / Shutterstock.com

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 for women in STEM careers.

20% more men than women received a Facebook ad for STEM careers

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.

CASE 4 Invisible biases, visible inequalities

AI may be trained on personnel datasets that represent pre-existing human hiring biases, for example featuring a strong male skew. This will result in AI-systems which replicate male biases.

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 CO_2 equivalent - about five times the lifetime emission of the average car - including the manufacture of the car itself (Strubell et al, 2019).

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.

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

4. IMPLEMENTING THE RECOMMENDATION

There is still a long way to go to provide Member States with actionable resources that ensure the effective implementation of the Recommendation.

For this reason, UNESCO will develop two practical methodologies:

1. READINESS ASSESSMENT METHODOLOGY (RAM)

Description:

The RAM is designed to help assess whether Member States are prepared to effectively implement the Recommendation.

The RAM may address questions such as:

- Are laws in place to regulate AI?
- Does the national infrastructure support the accessibility of AI technologies?

Purpose:

This methodology will help Member States identify their status of preparedness and provide a basis for UNESCO to custom-tailor its capacity-building support. It may also be used to help Member States develop a roadmap towards ethical development and use of AI.

2. ETHICAL IMPACT ASSESSMENT (EIA)

Description:

EIA is a structured process which operationalises the Recommendation by helping AI project teams, in collabouration with the affected communities, to identify and assess the impacts an AI system may have.

The EIA may address questions such as:

- Who is most likely to be adversely affected by this AI system?
- What form will these impacts take?
- What can be done to prevent these harms and have resources been allocated to this harm prevention?

Purpose:

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

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 organizations to translate the Recommendation into policies and actions. The multidimensional implementation strategy includes such elements as:

- **Global Observatory on the Ethics of AI**, an innovative digital platform serving as a one-stop-shop for the latest analysis on the ethical development and use of AI around the world, and for the knowledge generated through the implementation of the Recommendation in different countries.

- **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.

- **AI Ethics Experts Without Borders (AIEB) network**, a flexible facility of experts for deployment in Member States on needs basis to assist in the implementation of the Recommendation and the application of the capacity-building tools.

- **Women for AI Ethics (W4AIEthics) network**, a platform for influential women leaders in industry, government, and civil society, driving transformations towards gender equality in and through AI.

Stay tuned for the developments on these initiatives by visiting the website of the Recommendation (www.unesco.org/artificial-intelligence). If you wish to be involved in these initiatives, please contact us at **ai-ethics@ unesco.org**.

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**.