United Nations
Educational, Scientific and
Cultural Organization

Ad Hoc Expert Group (AHEG) for the preparation of a draft text of a recommendation on the ethics of artificial intelligence

Distribution: limited

SHS/BIO/AHEG-AI/2020/4 REV.
Paris, 15 May 2020
English only

OUTCOME DOCUMENT:

FIRST VERSION OF A DRAFT TEXT OF A RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

In line with the decision of UNESCO's General Conference at its 40th session (40 C/Resolution 37), the Director-General constituted the Ad Hoc Expert Group (AHEG) for the preparation of a draft text of a recommendation on the ethics of artificial intelligence in March 2020.

Adapting to the challenging situation posed by the COVID-19 pandemic, the AHEG worked virtually from the end of March until beginning of May 2020, and produced the first version of a draft text of the Recommendation on the Ethics of Artificial Intelligence contained in this document.

It is underlined that this first version of a draft text will continue to be revised by the AHEG until beginning of September 2020, taking into account the feedback received during the multi-stakeholder consultation process to be held from June to July 2020.

This document does not claim to be exhaustive and does not necessarily represent the views of the Member States of UNESCO.

FIRST VERSION OF A DRAFT TEXT OF A RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

PREAMBLE

The General Conference of the United Nations Educational, Scientific and Cultural Organization (UNESCO), meeting in Paris from xx to xx, at its xx session,

Recalling that, by the terms of its Constitution, UNESCO seeks to construct the defences of peace in the minds of human beings and aims to promote cooperation among the nations through education, science, culture, and communication and information, in order to further universal respect for justice, for the rule of law and for the human rights and fundamental freedoms which are affirmed for the peoples of the world,

Reflecting on the profound influence that Artificial Intelligence (AI) may have on societies, ecosystems, and human lives, including the human mind, in part because of the new ways in which it influences human thinking and decision-making, and affects education, science, culture, and communication and information,

Considering that AI systems can be of great service to humanity but also raise fundamental ethical concerns, for instance regarding the biases they can embed and exacerbate, potentially resulting in inequality, exclusion and a threat to cultural and social diversity and gender equality; the need for transparency and understandability of the workings of algorithms and the data with which they have been trained; and their potential impact on privacy, freedom of speech, social, economic and political processes, and the environment,

Recognizing that the development of AI can deepen existing divides and inequalities in the world, and that no one should be left behind who does not want to, either in enjoying the benefits of AI or in the protection against its negative implications, while recognizing the different circumstances of different countries,

Conscious of the fact that low and middle income countries (LMICs), including but not limited to those in Africa, Latin America and the Caribbean, and Central Asia, as well as Small Island Developing States, are facing an acceleration of the use of information technologies and AI and that the digital economy presents important societal challenges and opportunities for creative societies, requiring the recognition of endogenous cultures, values and knowledge in order to develop economies,

Recognizing that AI has the potential to be beneficial to the environment, via its roles in ecological and climate research, disaster risk management, and agriculture, but that for those benefits to be realized, fair access to the technology is required and the potential benefits need to be balanced against the environmental impact of the entire AI and information technology production cycle,

Noting that addressing risks and ethical concerns should not hamper innovation but rather stimulate new practices of responsible research and innovation in which the research, design, development, deployment, and use of AI is anchored in moral values and ethical reflection,

Recalling that in November 2019, the General Conference of UNESCO, at its 40th session, adopted 40 C/Resolution 37, by which it mandated the Director-General “to prepare an international standard-setting instrument on the ethics of artificial intelligence (AI) in the form of a recommendation”, which is to be submitted to the General Conference at its 41st session in 2021,

Convinced that the standard-setting instrument presented here, based on a global normative approach, and focusing on human dignity and human rights, including diversity, interconnectedness, inclusiveness and fairness, can guide the research, design, development, deployment, and use of AI in a responsible direction,

Observing that a normative framework for AI and its social implications finds itself at the intersection of ethics, human rights, international and national legal frameworks, and the freedom of research and innovation, and human well-being,

Recognizing that ethical values and principles are not necessarily legal norms in and of themselves, but can powerfully shape the development and implementation of policy measures and legal norms, by providing guidance where the ambit of norms is unclear or where such norms are not yet in place due to the fast pace of technological development combined with the relatively slower pace of policy responses,

Convinced that globally accepted ethical standards can play a helpful role in harmonizing AI-related legal norms across the globe, and responsible application of existing international law, if this application is in line with ethical frameworks and does not cause harm locally,

Recognizing the Universal Declaration of Human Rights (1948), including Article 27 emphasizing the right to share in scientific advancement and its benefits; the instruments of the international human rights framework, including the United Nations Convention on the Elimination of All Forms of Discrimination against Women (1979), the United Nations Convention on the Rights of the Child (1989), and the United Nations Convention on the Rights of Persons with Disabilities (2006); the UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions (2005),

Noting the UNESCO Declaration on the Responsibilities of the Present Generations Towards Future Generations (1997); the United Nations Declaration on the Rights of Indigenous Peoples (2007); the Report of the United Nations Secretary-General on the Follow-up to the Second World Assembly on Ageing (A/66/173) of 2011, focusing on the situation of the human rights of older persons; the Report of the Special Representative of the United Nations Secretary-General on the issue of human rights and transnational corporations and other business enterprises (A/HRC/17/31) of 2011, outlining the 'Guiding Principles on Business and Human Rights: Implementing United Nations "Protect, Respect and Remedy" Framework'; the Human Rights Council's resolution on 'The right to privacy in the digital age' (A/HRC/RES/42/15) adopted on 26 September 2019; the UNESCO Recommendation on Science and Scientific Researchers (2017); the UNESCO Internet Universality Indicators (2019), including the R.O.A.M. principles; the Report of the United Nations Secretary-General's High-level Panel on Digital Cooperation on 'The Age of Digital Interdependence' (2019); and the outcomes and reports of the ITU's AI for Good Global Summits,

Noting also existing frameworks related to the ethics of AI of other intergovernmental organizations, such as the relevant human rights and other legal instruments adopted by the Council of Europe, and the work of its Ad Hoc Committee on AI (CAHAI); the work of the European Union related to AI, and of the European Commission's High-Level Expert Group on AI, including the Ethical Guidelines for Trustworthy AI; the work of the OECD Expert Group on AI (AIGO), and the OECD's Recommendation of the Council on AI; the G20 AI Principles, drawn therefrom, and outlined in the G20 Ministerial Statement on Trade and Digital Economy; the G7's Charlevoix Common Vision for the Future of AI; the work of the African Union's Working Group on AI; and the work of the Arab League's Working Group on AI,

Emphasizing that specific attention must be paid to LMICs, including but not limited to those in Africa, Latin America and the Caribbean, and Central Asia, as well as Small Island Developing States, as they have been underrepresented in the AI ethics debate, which raises concerns about neglecting local knowledge, cultural and ethical pluralism, value systems and the demands of global fairness,

Conscious of the many national frameworks related to the ethics and regulation of AI,

Conscious as well of the many initiatives and frameworks related to the ethics of AI developed by the private sector, professional organizations, and non-governmental organizations, such as the IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems and its work on Ethically Aligned Design; the World Economic Forum's 'Global Technology Governance: A Multistakeholder Approach'; the UNI Global Union's 'Top 10 Principles for Ethical Artificial Intelligence'; the Montreal Declaration for a Responsible Development of AI; the Harmonious Artificial Intelligence Principles (HAIP); and the Tenets of the Partnership on AI,

Convinced that AI can bring important benefits, but that achieving them can also be under tension of innovation debt, asymmetric access to knowledge, barriers of rights to information and gaps in capacity of creativity in developing cycles, human and institutional capacities, barriers to access technological innovation, and a lack of adequate infrastructure and regulations regarding data,

Recognising that economic competition is taking place within and between states and also between multinational companies, potentially causing AI strategies and regulatory frameworks to be focused on national and commercial interests, while global cooperation is needed to address the challenges that AI brings in a diversity and interconnectivity of cultures and ethical systems, and to mitigate potential misuse,

Taking fully into account that the rapid development of AI systems encounters barriers to understand and implement AI, because of the diversity of ethical orientations and cultures around the World, the lack of agility of the law in relation to technology and the information society, and the risk that local and regional ethical standards and values be disrupted by AI,

1. ***Adopts*** the present Recommendation on the Ethics of Artificial Intelligence;

2. ***Recommends*** that Member States apply the provisions of this Recommendation by taking appropriate steps, including whatever legislative or other measures may be required, in conformity with the constitutional practice and governing structures of each State, to give effect within their jurisdictions to the principles and norms of the Recommendation;

3. ***Also recommends*** that Member States bring the Recommendation to the attention of the authorities, bodies, institutions and organizations in public, commercial and non-commercial sectors involved in the research, design, development, deployment, and use of AI systems.

I. SCOPE OF APPLICATION

1. This Recommendation addresses ethical issues related to AI. It approaches AI ethics as a holistic framework of interdependent values, principles and actions that can guide societies in the AI system lifecycle, referring to human dignity and well-being as a compass to deal responsibly with the known and unknown impacts of AI systems in their interactions with human beings and their environment. The AI system lifecycle refers to the research, design, development, deployment, and use of AI systems, and the use of AI systems can be understood to include the maintenance, operation, end-of-use, and disassembly of AI

systems. It is not within the ambition of this instrument to provide one single definition of AI, since such a definition would need to change over time, in accordance with technological developments. Rather, its ambition is to address those features of AI systems that are of central ethical relevance and on which there is large international consensus. For the purposes of this Recommendation, AI systems can be approached as technological systems which have the capacity to process information in a way that resembles intelligent behaviour, and typically includes aspects of learning, perception, prediction, planning or control. This Recommendation approaches AI systems along the following lines:

a. First of all, AI systems embody models and algorithms that produce a capacity to learn and to perform cognitive tasks, like making recommendations and decisions in real and virtual environments. AI systems are designed to operate with varying levels of autonomy by means of knowledge modeling and representation and by exploiting data and calculating correlations. AI systems may include several approaches and technologies, such as but not limited to:
    i. machine learning, including deep learning and reinforcement learning,
    ii. machine reasoning, including planning, scheduling, knowledge representation, search, and optimization, and
    iii. cyber-physical systems, including internet-of-things and robotics, which involve control, perception, the processing of data collected by sensors, and the operation of actuators in the environment in which AI systems work.

b. Second, besides raising ethical issues similar to the ones raised by any technology, AI systems also raise new types of issues. Some of these issues are related to the fact that AI systems are capable of doing things which previously only living beings could do, and which were in some cases even limited to human beings only. These characteristics give AI systems a profound, new role in human practices and society. Going even further, in the long term, AI systems could challenge human's special sense of experience and consciousness, raising additional concerns about human autonomy, worth and dignity, but this is not yet the case.

c. Third, even though ethical questions regarding AI are generally related to the concrete impact of AI systems on human beings and societies, another set of ethical issues is directed at the interactions between AI systems and human beings and its implications for our understanding of both human beings and technologies. This Recommendation acknowledges that both types of questions are closely related and are necessary elements of an ethical approach to AI.

2. This Recommendation pays specific attention to the broader ethical implications of AI in relation to the central domains of UNESCO: education, science, culture, and communication and information, as explored in the 2019 Preliminary Study on the Ethics of Artificial Intelligence by the UNESCO World Commission on Ethics of Scientific Knowledge and Technology (COMEST):

a. AI systems are connected to education in many ways: they challenge the societal role of education because of their implications for the labour market and employability; they might have impact on educational practices; and they require that education of AI engineers and computer scientists creates awareness of the societal and ethical implications of AI.

b. In all fields of the sciences, social sciences and humanities, AI has implications for our concepts of scientific understanding and explanation, and for the ways in which scientific knowledge can be applied as a basis for decision-making.

c. AI has implications for cultural identity and diversity. It has the potential to positively impact the cultural and creative industries, but it may also lead to an increased concentration of supply of cultural content, data and income in the hands of only a few actors, with potential negative implications for the diversity of cultural expressions and equality.

d. In the field of communication and information, machine-powered translation of languages is likely to play an increasingly important role. This might have a substantial impact on language and human expression, in all dimensions of life, bringing a responsibility to deal carefully with human languages and their diversity. Moreover, AI is challenging practices of journalism, and the social role of journalists, media workers, and social media producers who are engaged in journalistic activities, and is connected to both the spreading and the detection of disinformation or misunderstanding.

3. This Recommendation is addressed to States. As appropriate and relevant, it also provides guidance to decisions or practices of individuals, groups, communities, institutions and corporations, public and private, particularly AI actors, understood as those who play an active role in the AI system lifecycle, including organizations and individuals that research, design, develop, deploy, or use AI.

II. AIMS AND OBJECTIVES

4. This Recommendation aims for the formulation of ethical values, principles and policy recommendations for the research, design, development, deployment and usage of AI, to make AI systems work for the good of humanity, individuals, societies, and the environment.

5. The complexity of the ethical issues surrounding AI requires equally complex responses that necessitate the cooperation of multiple stakeholders across the various levels and sectors of the international, regional and national communities.

6. Even though this Recommendation is addressed primarily to policy-makers in and outside UNESCO Member States, it also aims to provide a framework for international organizations, national and transnational corporations, NGO's, engineers and scientists, including representatives of humanities, natural and social sciences, non-governmental organizations, religious organizations, and civil society, stimulating a multi-stakeholder approach, grounded in a globally accepted ethical framework that enables stakeholders to collaborate and take common responsibility based on a global, intercultural dialogue.

III. VALUES AND PRINCIPLES

7. Values and principles are not necessarily legal norms in and of themselves, as stated in the preamble to this Recommendation. They play a powerful role in shaping policy measures and legal norms, because values encompass internationally agreed expectations of what is good and what is to be preserved. As such, values underpin principles.

8. Values thus inspire good moral behaviour in line with the international community's understanding of such behaviour and they are the foundations of principles, while principles unpack the values underlying them more concretely so that values can be more easily actualised in policy statements and actions.

III.1. VALUES

Human dignity

9. The research, design, development, deployment, and use of AI systems should respect and preserve human dignity. The dignity of every human person is a value that constitutes a foundation for all human rights and fundamental freedoms and is essential when developing and adapting AI systems. Human dignity relates to the recognition of the intrinsic worth of each individual human being and thus dignity is not tied to national origin, legal status, socio-economic position, gender and sexual orientation, religion, language, ethnic origin, political ideology or other opinion.

10. This value should be respected by all actors involved in the research, design, development, deployment, and use of AI systems in the first place; and in the second place, be promoted through new legislation, through governance initiatives, through good exemplars of collaborative AI development and use, or through government-issued national and international technical and methodological guidelines as AI technologies advance.

Human rights and fundamental freedoms

11. The value of the respect for, and protection and promotion of, human rights and fundamental freedoms in the AI context means that the research, design, development, deployment, and use of AI systems should be consistent and compliant with international human rights law, principles and standards.

Leaving no one behind

12. It is vital to ensure that AI systems are researched, designed, developed, deployed, and used in a way that respects all groupings of humanity and fosters creativity in all its diversity. Discrimination and bias, digital and knowledge divides and global inequalities need to be addressed throughout an AI system lifecycle.

13. Thus, the research, design, development, deployment, and use of AI systems must be compatible with empowering all humans, taking into consideration the specific needs of different age groups, cultural systems, persons with disabilities, women and girls, disadvantaged, marginalized and vulnerable populations; and should not be used to restrict the scope of lifestyle choices or personal experiences, including the optional use of AI-systems. Furthermore, efforts should be made to overcome the lack of necessary technological infrastructure, education and skills, as well as legal frameworks, particularly in low- and middle-income countries.

Living in harmony

14. The value of living in harmony points to the research, design, development, deployment, and use of AI systems recognising the interconnectedness of all humans. The notion of being interconnected is based on the knowledge that every human belongs to a greater whole, which is diminished when others are diminished in any way.

15. This value demands that the research, design, development, deployment, and use of AI systems should avoid conflict and violence, and should not segregate, objectify, or undermine the safety of human beings, divide and turn individuals and groups against each other, or threaten the harmonious coexistence between humans and the natural environment, as this would negatively impact on humankind as a collective. The purpose of this value is to recognise the enabling role that AI actors should play in achieving the goal of living in harmony, which is to ensure a future for common good.

Trustworthiness

16. AI systems should be trustworthy. Trustworthiness is a socio-technical concept implying that the research, design, development, deployment, and use of AI systems should inspire, instead of infringing on, trust among people and in AI systems.

17. Trust has to be earned in each use context and more broadly is a benchmark for the social acceptance of AI systems. Therefore people should have good reason to trust that AI technology brings benefits while adequate measures are taken to mitigate risks.

Protection of the Environment

18. The aim of this value is to ensure that the research, design, development, deployment, and use of AI systems recognise the promotion of environmental well-being. All actors involved during the lifecycle of AI systems should follow relevant international and domestic laws in the field of environmental protection and sustainable development to ensure the minimisation of climate change risk factors, including carbon emission of AI systems, and prevent the exploitation and depletion of natural resources contributing to the deterioration of the environment.

19. At the same time, AI systems should be used to provide solutions to protect the environment and preserve the planet by supporting circular economy type approaches.

III.2. PRINCIPLES

20. Bearing in mind that any AI system has a number of essential evolving human and technology dependent situational characteristics, principles are presented in two groups.

21. The first group consists of principles reflecting characteristics that are associated with the human-technology interface, i.e. human-AI systems interaction. Note that the research, design, development, deployment, and use of AI systems influence human agency in two ways: First, in terms of expanding the scope for machine autonomy and decision-making, and second, by influencing the quality of human agency in both positive and negative ways.

22. The second group of principles consists of principles reflecting characteristics associated with the properties of AI systems themselves that are pertinent to ensuring the research, design, development, deployment, and use of AI systems happen in accordance with internationally accepted expectations of ethical behaviour.

GROUP 1

For human and flourishing

23. AI systems should be researched, designed, developed, deployed, and used to let humans and the environment in which they live, flourish. Throughout the lifecycle of AI systems the quality of life of every human being should be enhanced and the enjoyment of all human rights for every human being should be promoted, while the definition of ‘quality of life’ should be left open to individuals or groups, as long as no human being is harmed physically or mentally, or their dignity diminished in terms of this definition.

24. AI systems may be researched, designed, developed, deployed or used to assist in interactions involving vulnerable people, including, but not limited to children, the elderly or the ill, but should never objectify humans or undermine human dignity, or violate or abuse human rights.

Proportionality

25. The research, design, development, deployment, and use of AI systems may not exceed what is necessary to achieve legitimate aims or objectives and should be appropriate to the context.

26. The choice of an AI method should be justified in the following ways: (a) The AI method chosen should be desirable and proportional to achieve a given aim; (b) The AI method chosen should not have an excessive negative infringement on the foundational values captured in this document; (c) The AI method should be appropriate to the context.

Human oversight and determination

27. It should always be possible to attribute both ethical and legal responsibility for the research, design, development, deployment, and use of AI systems to a physical person or to an existing legal entity. Human oversight refers thus not only to individual human oversight, but to public oversight.

28. It may be the case that sometimes humans would have to share control with AI systems for reasons of efficacy, but this decision to cede control in limited contexts remains that of humans, as AI systems should be researched, designed, developed, deployed, and used to assist humans in decision-making and acting, but never to replace ultimate human responsibility.

Sustainability

29. In the context of promoting the development of sustainable societies, AI actors should respect the social, economic and environmental dimensions of sustainable development of all of humanity and the environment. AI systems should be researched, designed, developed, deployed, and used to promote the achievement of sustainability related to globally accepted frameworks such as the sustainable development goals.

Diversity and inclusiveness

30. The research, design, development, deployment, and use of AI systems should respect and foster diversity and inclusiveness at a minimum consistent with international human rights law, standards and principles, including demographic, cultural and social diversity and inclusiveness.

Privacy

31. The research, design, development, deployment, and use of AI systems should respect, protect and promote privacy, a right essential to the protection of human dignity and human agency. Adequate data governance mechanisms should be ensured throughout the lifecycle of AI systems including as concerning the collection of data, control over the use of data through informed consent and permissions and disclosures of the application and use of data, and ensuring personal rights over and access to data.

Awareness and literacy

32. Public awareness and understanding of AI technologies and the value of data should be promoted through education, public campaigns and training to ensure effective public participation so that citizens can take informed decisions about their use of AI systems. Children should be protected from reasonably foreseeable harms arising from AI systems, should have access to such systems through education and training, and children should not be disempowered by their interaction with AI systems.

Multi-stakeholder and adaptive governance

33. Governance of AI should be responsive to shifts in technology and associated business models, inclusive (with the participation of multiple stakeholders), potentially distributed across different levels, and ensure through a cross-domain systems approach, fit-for-purpose governance responses.

34. Governance should consider a range of responses from soft governance through self-regulation and certification processes to hard governance with national laws and, where possible and necessary, international instruments. In order to avoid negative consequences and unintended harms, governance should include aspects of anticipation, protection, monitoring of impact, enforcement and redressal.

GROUP 2

Fairness

35. AI actors should respect fairness, equity and inclusiveness, as well as make all efforts to minimize and avoid reinforcing or perpetuating socio-technical biases including racial, ethnic, gender, age, and cultural biases, throughout the full lifecycle of the AI system.

Transparency and explainability

36. While, in principle, all efforts need to be made to increase transparency and explainability of AI systems to ensure trust from humans, the level of transparency and explainability should always be appropriate to the use context, as many trade-offs exist between transparency and explainability and other principles such as safety and security.

37. Transparency means allowing people to understand how AI systems are researched, designed, developed, deployed, and used, appropriate to the use context and sensitivity of the AI system. It may also include insight into factors that impact a specific prediction or decision, but it does not usually include sharing specific code or datasets. In this sense, transparency is a socio-technical issue, with the aim of gaining trust from humans for AI systems.

38. Explainability refers to making intelligible and providing insight into the outcome of AI systems. The explainability of AI models also refers to the understandability of the input, output and behaviour of each algorithmic building block and how it contributes to the outcome of the models. Thus, explainability is closely related to transparency, as outcomes and sub processes leading to outcomes should be understandable and traceable, appropriate to the use context.

Safety and security

39. The research, design, development, deployment, and use of AI systems should avoid unintended harms (safety risks) and vulnerabilities to attacks (security tasks), so as to ensure safety and security throughout the lifecycle of the AI system.

40. Governments should play a leading role in ensuring safety and security of AI systems, including through establishing national and international standards and norms in line with applicable international human rights law, standards and principles. Strategic research on potential safety and security risks associated with different approaches to realize long-term AI should be continuously supported to avoid catastrophic harms.

Responsibility and accountability

41. AI actors should assume moral and legal responsibility in accordance with extant international human rights law and ethical guidance throughout the lifecycle of AI systems. The responsibility and liability for the decisions and actions based in anyway on an AI system should always ultimately be attributable to AI actors.

42. Appropriate mechanisms should be developed to ensure accountability for AI systems and their outcome. Both technical and institutional designs should be considered to ensure auditability and traceability of (the working of) AI systems.

IV. AREAS OF POLICY ACTION

ACTION GOAL I: ETHICAL STEWARDSHIP

43. Ensure alignment of AI research, design, development, deployment, and use with foundational ethical values such as human rights, diversity and inclusiveness, etc.

Policy Action 1: Promoting Diversity & Inclusiveness

44. Member States should work with international organizations to ensure the active participation of all Member States, especially LMICs in international discussions concerning AI. This can be through the provision of funds, ensuring equal regional participation, or any other mechanisms.

45. Member States should require AI actors to disclose and combat any cultural and social stereotyping in the workings of AI systems whether by design or by negligence, and ensure that training data sets for AI systems should not foster cultural and social inequalities. Mechanisms should be adopted to allow end users to report such inequalities, biases and stereotypes.

46. Member States should ensure that AI actors demonstrate awareness and respect for the current cultural and social diversities including local customs and religious traditions, in the research, design, development, deployment, and use of AI systems while being consistent with international human rights standard and norms.

47. Member States should work to address the diversity gaps currently seen in the development of AI systems, including diversity in training datasets and in AI actors themselves. Member States should work with all sectors, international and regional organizations and other entities to empower women and girls to participate in all stages of an AI system lifecycle by offering incentives, access to mentors and role models, and protection from harassment. They should also work to make the domain of AI more accessible to people from diverse ethnic backgrounds as well as people with disabilities. Moreover, equal access to AI system benefits should be promoted, particularly for marginalized groups.

48. Member States should work with international organizations to mainstream AI ethics by including discussions of AI-related ethical issues into relevant international, intergovernmental and multi-stakeholder fora.

ACTION GOAL II: IMPACT ASSESSMENT

49. Build observatory and anticipatory capacities to respond in time to negative or other unintended consequences arising from AI systems.

Policy Action 2: Addressing Labour Market Changes

50. Member States should work to assess and address the impact of AI on labour markets and its implications for education requirements. This can include the introduction of a wider range of 'core skills' at all education levels to give new generations a fair chance of finding jobs in a rapidly changing market and to ensure their awareness of the ethical aspects of AI. Skills such as 'learning how to learn', communication, teamwork, empathy, and the ability to transfer one's knowledge across domains, should be taught alongside specialist, technical skills. Being transparent about what skills are in demand and updating school curricula around these is key.

51. Member States should work with private entities, NGOs and other stakeholders to ensure a fair transition for at-risk employees. This includes putting in place upskilling and reskilling programs, finding creative ways of retaining employees during those transition periods, and exploring 'safety net' programs for those who cannot be retrained.

52. Member States should encourage researchers to analyze the impact of AI on the local labour market in order to anticipate future trends and challenges. These studies should shed light on which economic, social and geographic sectors will be most affected by the massive incorporation of AI.

53. Member States should develop labour force policies targeted at supporting women and underrepresented populations to make sure no one is left out of the digital economy powered by AI. Special investment in providing targeted programs to increase the preparedness, employability, career development and professional growth of women and underrepresented populations should be considered, and implemented if feasible.

Policy Action 3: Addressing the social and economic impact of AI

54. Member States should devise mechanisms to prevent the monopolization of AI and the resulting inequalities, whether these are data, research, technology, market or other monopolies.

55. Member States should work with international organizations, private and non-governmental entities to provide adequate AI literacy education to the public especially in LMICs in order to reduce the digital divide and digital access inequalities resulting from the wide adoption of AI systems.

56. Member States should establish monitoring and evaluation mechanisms for initiatives and policies related to AI ethics. Possible mechanisms include: a repository covering ethical compliance initiatives across UNESCO's areas of competence, an experience sharing mechanism for Member States to seek feedback from other Member States on their policies and initiatives, and a guide for developers of AI systems to assess their adherence to policy recommendations mentioned in this document.

57. Member States are encouraged to consider a certification mechanism for AI systems similar to the ones used for medical devices. This can include different classes of certification according to the sensitivity of the application domain and expected impact on human lives, the environment, ethical considerations such as equality, diversity and cultural values, among others. Such a mechanism might include different levels of audit of systems, data, and ethical compliance. At the same time, such a mechanism must not hinder innovation or disadvantage small enterprises or startups by requiring large amounts of paperwork. These mechanisms would also include a regular monitoring component to ensure system robustness and continued integrity and compliance over the entire lifetime of the AI system, requiring re-certification if necessary.

58. Member States should encourage private companies to involve different stakeholders in their AI governance and to consider adding the role of an AI Ethics Officer or some other mechanism to oversee impact assessment, auditing and continuous monitoring efforts and ensure ethical compliance of AI systems.

59. Member States should work to develop data governance strategies that ensure the continuous evaluation of the quality of training data for AI systems including the adequacy of the data collection and selection processes, proper security and data protection measures, as well as feedback mechanisms to learn from mistakes and share best practices among all AI actors. Striking a balance between metadata and users' privacy should be an upfront concern for such a strategy.

Policy Action 4: Impact on Culture and on the Environment

60. Member States are encouraged to incorporate AI systems where appropriate in the preservation, enrichment and understanding of cultural heritage, both material and intangible, including rare languages, for example by introducing or updating educational programs related to the application of AI systems in these areas, targeted at institutions and the public.

61. Member States are encouraged to examine and address the impact of AI systems, especially Natural Language Processing applications such as automated translation and voice assistants on the nuances of human language. Such an assessment can include maximizing the benefits from these systems by bridging cultural gaps and increasing human understanding, as well as negative implications such as the reduced pervasiveness of rare languages, local dialects, and the tonal and cultural variations associated with human language and speech.

62. Member States should encourage and promote collaborative research into the effects of long-term interaction of people with AI systems. This should be done using multiple norms, principles, protocols, disciplinary approaches, and assessment of the modification of habits, as well as careful evaluation of the downstream cultural and societal impacts.

63. Member States should promote AI education for artists and creative professionals to assess the suitability of AI for use in their profession as AI is being used to create, produce, distribute and broadcast a huge variety of cultural goods and services, bearing in mind the importance of preserving cultural heritage and diversity.

64. Member States should promote awareness and evaluation of AI tools among local cultural industries and startups working in the field of culture, to avoid the risk of greater concentration in the cultural market.

65. Member States should work to assess and reduce the environmental impact of AI systems, including but not limited to, its carbon footprint. They should also introduce incentives to advance ethical AI-powered environmental solutions and facilitate their adoption in different contexts. Some examples include using AI to:

a. Accelerate the protection, monitoring and management of natural resources.

b. Support the prevention, control and management of climate-related problems.

c. Support a more efficient and sustainable food ecosystem.

d. Accelerate the access to and mass adoption of green energy.

ACTION GOAL III: CAPACITY BUILDING FOR AI ETHICS

66. Develop human and institutional capacity to enable ethical impact assessment, oversight and governance.

Policy Action 5: Promoting AI Ethics Education & Awareness

67. Member States should encourage in accordance with their national education programmes and traditions the embedding of AI ethics into the school and university curricula for all levels and promote cross collaboration between technical skills and social sciences and humanities. Online courses and digital resources should be developed in local languages and in accessible formats for people with disabilities.

68. Member States should promote the acquisition of 'prerequisite skills' for AI education, such as basic literacy, numeracy, and coding skills, especially in countries where there are notable gaps in the education of these skills.

69. Member States should introduce flexibility into university curricula and increase ease of updating them, given the accelerated pace of innovations in AI systems. Moreover, the integration of online and continuing education and the stacking of credentials should be explored to allow for agile and updated curricula.

70. Member States should promote general awareness programs of AI and the inclusive access to knowledge on the opportunities and challenges brought about by AI. This knowledge should be accessible to technical and non-technical groups with a special focus on underrepresented populations.

71. Member States should encourage research initiatives on the use of AI in teaching, teacher training and e-learning, among other topics, in a way that enhances opportunities and mitigates the challenges and risks associated with these technologies. This should always be accompanied by an adequate impact assessment of the quality of education and impact on students and teachers of the use of AI and ensure that AI empowers and enhances the experience for both groups.

72. Member States should support collaboration agreements between academic institutions and the industry to bridge the gap of skillset requirements and promote collaborations between industry sectors, academia, civil society, and the government to align training programs and strategies provided by educational institutions, with the needs of the industry. Project-based learning approaches for AI should be promoted, allowing for partnerships between companies, universities and research centers.

73. Member States should particularly promote the participation of women, diverse races and cultures, and people with disabilities, in AI education programs from basic school to higher education, as well as promote the monitoring and sharing of best practices with other Member States.

Policy Action 6: Promoting AI Ethics Research

74. Member States should promote AI ethics research either through direct investments or by creating incentives for the public and private sectors to invest in this area.

75. Member States should ensure that AI researchers are trained in research ethics and require them to include ethical considerations in their research design and end products, particularly analyses of the datasets they use, how they are annotated and the quality and the scope of the results.

76. Member States and private companies should facilitate access to data for research for the scientific community at the national level where possible to promote the capacity of the scientific community, particularly in developing countries. This access should not be at the expense of citizens' privacy.

77. Member States should promote gender diversity in AI research in academia and industry by offering incentives to women to enter the field, put in place mechanisms to fight gender stereotyping and harassment within the AI research community, and encouraging academic and private entities to share best practices on how to promote diversity.

78. Member States and funding bodies should promote interdisciplinary AI research by including disciplines other than science, technology, engineering, and mathematics (STEM), e.g. law, international relations, political sciences, education, philosophy, culture, and linguistic studies to ensure a critical approach to AI research and proper monitoring of possible misuses or adverse effects.

ACTION GOAL IV: DEVELOPMENT AND INTERNATIONAL COOPERATION

79. Ensure a cooperative and ethical approach to using AI in development applications, given the great opportunity this technology affords towards the acceleration of development efforts.

Policy Action 7: Promoting Ethical Use of AI in Development

80. Member States should encourage the ethical use of AI in areas of development such as healthcare, agriculture/food supply, education, culture, environment, water management, infrastructure management, economic planning and growth, and others.

81. Member States and international organizations should strive to provide platforms for international cooperation on AI for development, including by contributing expertise, funding, data, domain knowledge, infrastructure, and facilitating workshops between technical and business experts to tackle challenging development problems, especially for LMICs and LDCs.

82. Member States should work to promote international collaborations on AI research, including research centers and networks that promote greater participation of researchers from LMICs and other emerging geographies.

Policy Action 8: Promoting International Cooperation on AI Ethics

83. Member States should work through international organizations and research institutions to conduct AI ethics research. Both public and private entities should ensure that algorithms and data used in a wide array of AI areas – from policing and criminal justice to employment, health and education – are applied equally and fairly, including investigations into what sorts of equality and fairness are appropriate in different cultures and contexts, and exploring how to match those to technically feasible solutions.

84. Member States should encourage international cooperation in AI development and deployment to bridge geo-technological lines. This necessitates a multi-stakeholder effort at the national, regional and international levels. Technological exchanges/ consultations should take place between Member States and their populations, between the public and private sectors, and between and among Member States.

ACTION GOAL V: GOVERNANCE FOR AI ETHICS

85. Promote and guide the inclusion of ethical considerations in the governance of AI systems.

Policy Action 9: Establishing Governance Mechanisms for AI Ethics

86. Member States should ensure that any AI governance mechanism is:

a. Inclusive: invites and encourages participation of representatives of indigenous communities, women, young and elderly people, people with disabilities, and other minority and underrepresented groups.

b. Transparent: accepts oversight from relevant national structures or trusted third-parties. For the media, this could be a cross-sectoral taskforce that fact-checks sources; for technology companies, this could be external audits of design, deployment and internal audit processes; for Member States, this could be reviews by human rights forums.

c. Multidisciplinary: any issue should be viewed in a holistic way and not only from the technological point of view.

d. Multilateral: international agreements should be established to mitigate and redress any harm that can appear in a country caused by a company or user based in another. This does not negate different countries and regions developing their own rules as appropriate to their cultures.

87. Member States should foster the development of, and access to, a digital ecosystem for ethical AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, Member States should consider reviewing their policies and regulatory frameworks, including on access to information and open government to reflect AI-specific requirements and promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data, among others.

88. Member States should encourage development and use of comparable AI guidelines, including ethical aspects at global and regional levels, and gather the required evidence to evaluate, monitor and control the progression in the ethical implementation of AI systems.

89. Member States should consider the development and implementation of an international legal framework to encourage international cooperation between States and other stakeholders.

Policy Action 10: Ensuring Trustworthiness of AI Systems

90. Member States and private companies should implement proper measures to monitor all phases of an AI system lifecycle, including the behaviour of algorithms in charge of decision making, the data, as well as AI actors involved in the process, especially in public services and where direct end-user interaction is needed.

91. Member States should work on setting clear requirements for AI system transparency and explainability based on:

a. Application domain: some sectors such as law enforcement, security, education and healthcare, are likely to have a higher need for transparency and explainability than others.

b. Target audience: the level of information about an AI system's algorithms and outcome and the form of explanation required may vary depending on who are requesting the explanation, for example: users, domain experts, developers, etc.

c. Feasibility: many AI algorithms are still not explainable; for others, explainability adds a significant implementation overhead. Until full explainability is technically possible with minimal impact on functionality, there will be a trade-off between the accuracy/quality of a system and its level of explainability.

92. Member States should encourage research into transparency and explainability by putting additional funding into those areas for different domains and at different levels (technical, natural language, etc.).

93. Member States and international organizations should consider developing international standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined.

Policy Action 11: Ensuring Responsibility, Accountability and Privacy

94. Member States should review and adapt, as appropriate, regulatory and legal frameworks to achieve accountability and responsibility for the content and outcomes of AI systems at the different phases of their lifecycle. Governments should introduce liability frameworks or clarify the interpretation of existing frameworks to make it possible to attribute accountability for the decisions and behaviour of AI systems. When developing regulatory frameworks governments should, in particular, take into account that responsibility and accountability must always lie with a natural or legal person; responsibility should not be delegated to an AI system, nor should a legal personality be given to an AI system.

95. Member States are encouraged to introduce impact assessments to identify and assess benefits and risks of AI systems, as well as risk prevention, mitigation and monitoring measures. The risk assessment should identify impacts on human rights, the environment, and ethical and social implications in line with the principles set forth in this Recommendation. Governments should adopt a regulatory framework that sets out a procedure for public authorities to carry out impact assessments on AI systems acquired, developed and/or deployed by those authorities to predict consequences, mitigate risks, avoid harmful consequences, facilitate citizen participation and address societal challenges. As part of impact assessment, the public authorities should be required to carry out self-assessment of existing and proposed AI systems, which in particular, should include the assessment whether the use of AI systems within a particular area of the public sector is appropriate and what the appropriate method is. The assessment should also establish appropriate oversight mechanisms, including auditability, traceability and explainability which enables the assessment of algorithms, data and design processes, as well as include external review of AI systems. Such an assessment should also be multidisciplinary, multi-stakeholder, multicultural, pluralistic and inclusive.

96. Member States should involve all actors of the AI ecosystem (including, but not limited to, representatives of civil society, law enforcement, insurers, investors, manufacturers, engineers, lawyers, and users) in a process to establish norms where these do not exist. The norms can mature into best practices and laws. Member States are further encouraged to use mechanisms such as regulatory sandboxes to accelerate the development of laws and policies in line with the rapid development of new technologies and ensure that laws can be tested in a safe environment before being officially adopted.

97. Member States should ensure that harms caused to users through AI systems can be investigated, punished, and redressed, including by encouraging private sector companies to

provide remediation mechanisms. The auditability and traceability of AI systems, especially autonomous ones, should be promoted to this end.

98. Member States should apply appropriate safeguards of individuals' fundamental right to privacy, including through the adoption or the enforcement of legislative frameworks that provide appropriate protection, compliant with international law. In the absence of such legislation, Member States should strongly encourage all AI actors, including private companies, developing and operating AI systems to apply privacy by design in their systems.

99. Member States should ensure that individuals can oversee the use of their private information/data, in particular that they retain the right to access their own data, and "the right to be forgotten".

100. Member States should ensure increased security for personally identifiable data or data, which if disclosed, may cause exceptional damage, injury or hardship to a person. Examples include data relating to offences, criminal proceedings and convictions, and related security measures; biometric data; personal data relating to "racial" or ethnic origin, political opinions, trade-union membership, religious or other beliefs, health or sexual life.

101. Member States should work to adopt a Commons approach to data to promote interoperability of datasets while ensuring their robustness and exercising extreme vigilance in overseeing their collection and utilization. This might, where possible and feasible, include investing in the creation of gold standard datasets, including open and trustworthy datasets, that are diverse, constructed with the consent of data subjects, when consent is required by law, and encourage ethical practices in the technology, supported by sharing quality data in a common trusted and secured data space.

V. MONITORING AND EVALUATION

102. Member States should, according to their specific conditions, governing structures and constitutional provisions, monitor and evaluate policies, programmes and mechanisms related to ethics of AI using a combination of quantitative and qualitative approaches, as appropriate. Member States are encouraged to consider the following:

a. deploying appropriate research mechanisms to measure the effectiveness and efficiency of ethics of AI policies and incentives against defined objectives;

b. collecting and disseminating progress, good practices, innovations and research reports on ethics of AI and its implications with the support of UNESCO and international ethics of AI communities.

103. The possible mechanisms for monitoring and evaluation may include an AI observatory covering ethical compliance across UNESCO's areas of competence, an experience sharing mechanism for Member States to provide feedback on each other's initiatives, and a 'compliance meter' for developers of AI systems to measure their adherence to policy recommendations mentioned in this document.

104. Appropriate tools and indicators should be developed for measuring the effectiveness and efficiency of polices related to ethics of AI against agreed standards, priorities and targets, including specific targets for disadvantaged and vulnerable groups. This could involve evaluations of public and private institutions, providers and programmes, including self-evaluations, as well as tracer studies and the development of sets of indicators. Data collection and processing should be conducted in accordance with legislation on data protection.

105. Processes for monitoring and evaluating should ensure broad participation of relevant stakeholders, including, but not limited to, people of different age groups, persons with disabilities, women and girls, disadvantaged, marginalized and vulnerable populations, and respecting social and cultural diversity, with a view to improving learning processes and strengthening the connections between findings, decision-making, transparency and accountability for results.

VI. UTILIZATION AND EXPLOITATION OF THE PRESENT RECOMMENDATION

106. Member States should strive to extend and complement their own action in respect of this Recommendation, by cooperating with all national and international governmental and non-governmental organizations whose activities fall within the scope and objectives of this Recommendation.

107. Member States and stakeholders as identified in this Recommendation should take all feasible steps to apply the provisions spelled out above to give effect to the foundational values, principles and actions set forth in this Recommendation.

VII. PROMOTION OF THE PRESENT RECOMMENDATION

108. UNESCO has the vocation to be the principal United Nations agency to promote and disseminate this Recommendation, and accordingly shall work in collaboration with other United Nations entities, including but not limited to the United Nations Secretary-General's High-level Panel on Digital Cooperation, COMEST, the International Bioethics Committee (IBC), the Intergovernmental Bioethics Committee (IGBC), the International Telecommunication Union (ITU), and other relevant United Nations entities concerned with the ethics of AI.

109. UNESCO shall also work in collaboration with other international organizations, including but not limited to the African Union (AU), the Association of Southeast Asian Nations (ASEAN), the Council of Europe (CoE), the Eurasian Economic Union (EAEU), the European Union (EU), the Organisation for Economic Co-operation and Development (OECD) and the Organization for Security and Co-operation in Europe (OSCE), as well as the Institute of Electrical and Electronic Engineers (IEEE) and the International Organization for Standardization (ISO).

VIII. FINAL PROVISIONS

110. The Recommendation needs to be understood as a whole, and the foundational values and principles are to be understood as complementary and interrelated. Each principle is to be considered in the context of the foundational values.

111. Nothing in this Recommendation may be interpreted as approval for any State, other social actor, group, or person to engage in any activity or perform any act contrary to human rights, fundamental freedoms, human dignity and concern for life on Earth and beyond.