United Nations
Educational, Scientific and
Cultural Organization

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

April 2020

Version 1.1
Distribution: limited

SHS/BIO/AHEG-AI/2020/3 REV.
Paris, 10 April 2020
Original: English

WORKING DOCUMENT:

TOWARD A DRAFT TEXT OF A RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

BACKGROUND

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

Due to the COVID-19 pandemic the AHEG will convene virtually in April 2020.

The outcome of the AHEG's work will be a first draft text of a recommendation on the ethics of artificial intelligence which should be finalized by the AHEG by the end of April 2020.

The present document aims to provide the experts with background information as a starting point to develop the UNESCO's preparation of such a recommendation and research-based proposals on the potential substance and structure of the outcome document of the virtual discussion of the AHEG.

TABLE OF CONTENTS

United Nations
Educational, Scientific and
Cultural Organization

From
the People
of Japan

ACKNOWLEDGEMENT

UNESCO wishes to acknowledge the Government of Japan for its generous financial support leading to the preparation of this document and the ensuing work towards the draft text of a recommendation on the ethics of artificial intelligence.

LIST OF ABBREVIATIONS

AHEG	Ad Hoc Expert Group
AI	Artificial intelligence
CEB	United Nations System Chief Executives Board for Coordination
CoE	Council of Europe
COMEST	World Commission on the Ethics of Scientific Knowledge and Technology
EIA	Ethical Impact Assessment
HLCP	High-Level Committee on Programmes
IEEE	Institute of Electrical and Electronics Engineers
ILO	International Labour Organization
IOM	International Organization for Migration
IRCAI	International Research Centre on Artificial Intelligence (a UNESCO Category II Centre)
ITU	International Telecommunication Union
OHCHR	Office of the United Nations High Commissioner for Human Rights
R.O.A.M.	Human Rights, Openness, Accessibility and Multi-stakeholder participation
SDGs	Sustainable Development Goals
UN	United Nations
UN DESA	United Nations Department of Economic and Social Affairs
UN System	UN principal organs, specialized agencies and affiliated organizations
UNCTAD	United Nations Conference on Trade and Development
UNEP	United Nations Environment Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNFCCC	United Nations Framework Convention on Climate Change
UNFPA	United Nations Population Fund
UNICRI	United Nations Interregional Crime and Justice Research Institute
UNIDO	United Nations Industrial Development Organization
UNODC	United Nations Office on Drugs and Crime
UNSG	United Nations Secretary-General
UNU	United Nations University
WFP	World Food Programme

WHO	World Health Organization
WIPO	World Intellectual Property Organization

"How will machines know what we value if we don't know it ourselves?"

John C. Havens

I. INTRODUCTION

1. 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.

2. In the context of UNESCO, recommendations are instruments "in which the General Conference [the highest governing body of UNESCO] formulates principles and norms for the international regulation of any particular question and invites Member States to take whatever legislative or other steps may be required – in conformity with the constitutional practice of each State and the nature of the question under consideration – to apply the principles and norms aforesaid within their respective territories".[1] Adopted by the General Conference, the norms contained in a recommendation are not subject to ratification, but Member States are invited to apply them. Recommendations are intended to influence the development of national laws and practices. Therefore, the UNESCO recommendation on the ethics of artificial intelligence, once elaborated and adopted, will outline recommended principles and policy actions addressed primarily to Member States, as well as other stakeholders such as the private sector, civil society, technical community, international organizations. If the recommendation is adopted, Member States will be invited to submit periodic reports (generally every four years) on the measures that they have adopted in relation to the recommendation. This reporting modality will also serve as a monitoring mechanism to identify best practices, gaps, challenges for implementation, emerging risks and new principles that are needed as AI develops. Support will also be provided to assist Member States on the implementation of the recommendation as necessary and appropriate. In this regard, the recommendation will be an opportunity for Member States to discuss and agree upon an initial non-exhaustive set of basic principles and recommended policy actions with human rights guardrails for the ethical design, development and deployment of AI.

3. 40 C/Resolution 37 confirmed the unique perspective of UNESCO in promoting an ethical framework for AI given UNESCO's strong comparative advantage, recognizing its universality in membership and drawing on its multidisciplinary expertise. It is the only UN agency with a specialized mandate in the social and human sciences, in communication and information, as well as in the natural sciences, education and culture, whose constitutional aim is to advance international peace and the common welfare of mankind through strengthening its "intellectual and moral solidarity". Thus, recognizing that AI technologies are not value-neutral, in addition to the many ethical guidelines and frameworks that are currently being developed by governments, companies, technical community, civil society and international organizations, UNESCO brings a multidisciplinary, pluralistic, universal and holistic approach to the development of AI in the service of humanity, sustainable development, and peace. It is also the only UN agency that has embarked upon the process of developing a recommendation on the ethics of AI that is to be negotiated by representatives of Member States.

4. The international community requested UNESCO, at the World Summit on the Information Society in 2003 and 2005, confirmed by the United Nations General Assembly in 2015, to lead and facilitate international work on the "ethical dimension of the information society". UNESCO has a longstanding leading role at the UN level and globally in promoting ethical science, which harnesses technological and scientific advancements for the benefit of

[1] Article 1(b) of UNESCO's Rules of Procedure concerning recommendations to Member States and international conventions covered by the terms of Article IV, paragraph 4, of the Constitution.

all, protects the planet from ecological collapse and constitutes a solid basis for peaceful cooperation among peoples. Through the work of its consultative organs – the International Bioethics Committee (IBC, created in 1993) and the World Commission on the Ethics of Scientific Knowledge and Technology (COMEST, 1998) in coordination with the Intergovernmental Bioethics Committee (IGBC, 1998); – UNESCO has deepened its reflection on the role of sciences, technology and innovation in sustainable development, its interface with societies, on equitable and inclusive social development, including a coherent response to climate change through addressing the ethical principles of climate change adaptation and mitigation.

II. WORK OF COMEST AS THE TECHNICAL BASIS FOR THE DECISION OF UNESCO

5. The decision to proceed with a recommendation followed on from consideration of these issues by COMEST, a multidisciplinary scientific advisory body of UNESCO, made up of independent experts. The work of COMEST has built on and complemented work on AI being done within the United Nations system, other international organizations, nongovernmental organizations, academia and others.

6. Starting from its very first Ordinary Session (Oslo, Norway, 1999), COMEST has devoted significant efforts to the ethics of new technologies. More recently, this work has led to the adoption of two major COMEST documents directly linked to AI: "Report of COMEST on Robotics Ethics" (2017); and "Preliminary Study on the Ethics of Artificial Intelligence" (2019). The latter was produced by the COMEST Extended Working Group on the Ethics of AI and served as the technical basis for discussion at the 40th session of the General Conference.

7. In addressing the need to tackle the substantial societal and cultural implications of advancements in AI, COMEST underlined the ethical aspects thereof. In its 2019 Study, COMEST identified that the most central ethical issues from UNESCO's field of competencies concern "its implications for culture and cultural diversity, education, scientific knowledge, and communication and information." However, it goes beyond this as "given UNESCO's global orientation, the global-ethical themes of peace, sustainability, gender equality, and the specific challenges for Africa also deserve separate attention."

8. 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. Global cooperation on the ethics of AI and global inter-cultural dialogue are therefore indispensable for arriving at complex solutions.

9. Furthermore, COMEST has sought to identify a set of consensual ethical principles that address the ethical dimensions of AI and that could be included in an eventual recommendation on the ethics of AI. These ethical principles have been distilled from existing relevant international conventions and literature, classified and further elaborated in content and relevance. These suggestions also embody the global perspective of UNESCO, as well as UNESCO's specific areas of competence, and are discussed later in the document.

III. THE RELEVANCE OF UNESCO'S WORK ON A DRAFT TEXT OF A RECOMMENDATION ON THE ETHICS OF AI WITHIN THE UNITED NATIONS SYSTEM

10. The UN and the institutions within its system have been active in addressing challenges related to development, implementation and use of AI. As noted by the UN Secretary General, there is a need to ensure that AI becomes a force for good.[2] In May 2019, the UN Chief

[2] Special Address by Antonio Guterres, Secretary-General of the United Nations, World Economic Forum, 23 January 2020, https://www.weforum.org/events/world-economic-forum-annual-meeting-2020/sessions/special-address-by-antonio-guterres-secretary-general-of-the-united-nations-1.

Executives Board for Coordination (CEB) has adopted a United Nations system-wide strategic approach and road map for supporting capacity development on AI.[3] It outlines an internal plan to support capacity development efforts related to AI technologies, especially for developing countries, with a particular emphasis on the bottom billion, in the context of achieving the Sustainable Development Goals (SDGs). CEB members highlighted that while difficult, it was the responsibility of Member States and the UN system to start a global and inclusive conversation on the ethics of AI and contribute to the shaping of human rights-based norms and standards.

11. The UN Secretary-General's High-level Panel on Digital Cooperation produced the 2019 report The Age of Digital Interdependence recommending building an inclusive digital economy and societies; develop human and institutional capacity; protect human rights and human agency; promote digital trust, security and stability; and foster global digital cooperation. It provides recommendations on how the international community could work together to optimize the use of digital technologies and mitigate the risks. Recommendation 3C of the Report is directly relevant for to the ethics of AI. As a follow-up process, series of roundtables were and are being organized in 2019-2020 to provide inputs and advice on the status and feasibility of advancing recommendations with one specifically devoted to AI.

12. The International Telecommunication Union (ITU), the UN specialized agency for information and communication technologies, has been organizing, since 2017, the AI for Good Global Summit as a platform for global and inclusive dialogue on AI. In particular, the 2017 Summit shone a spotlight on the ethical development of AI and the last AI for Good UN Partners meeting decided to create a "working group on AI and Ethics" to be led by the Global Pulse and the World Bank.

13. Many other UN institutions have also engaged into addressing AI challenges. For an overview of the work of UN institutions in the field of AI see the 2019 ITU compendium United Nations Activities on Artificial Intelligence, and for a summary of past, ongoing and future initiatives see Annex 6. In particular, the UN specialized agency World Health Organization (WHO) has established an expert group to prepare a Guidance Document on Ethics and Governance of AI for Health, and is engaged in developing documents aimed at national and sub-national governments to encourage them to have appropriate policy and governance mechanisms to ensure ethical and safe use of AI in healthcare without hindering innovation. The United Nations Conference on Trade and Development (UNCTAD) has recognized the ethical dimension of development of new technologies, and AI, especially, since most developing countries do not have the capacity to make comprehensive risk assessments. Its forthcoming Technology and Innovation Report 2020 will outline the state-of-the-art debate and critically examine the possibility of frontier technologies (including AI) widening existing inequalities and creating new ones. The UN related International Organization for Migration (IOM) has a dedicated work stream focusing on developing ethics and guidance through inter-agency collaborations. It leads an inter-agency group on Data Science, Artificial Intelligence and Ethics, which established inter-agency peer review mechanisms for mathematical AI models and ethics. The World Food Programme has an AI and ethics project aimed at building a framework for AI governance in humanitarian aid. The United Nations Interregional Crime and Justice Research Institute (UNICRI) has established the Centre for Artificial Intelligence and Robotics with the aim to enhance understanding of the risk-benefit duality of AI and Robotics through improved coordination, knowledge collection and dissemination, awareness-raising and outreach activities.

14. UNESCO has specific added value in addressing challenges related to development, implementation and use of AI. While it seeks to counter the risk of a growing digital and knowledge divides that could leave behind those who are relatively disadvantaged, or excluded, such as people in least developed countries, women and girls, youth, people with

[3] A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence, CEB/2019/1/Add.3.

disabilities and marginalized groups in all societies, it aims to foster policies that can maximize the opportunities offered by AI for enhancing democracy, development and human rights. It brings the Global South into the discussion and ensures multidisciplinarity, multiculturalism and pluralism of value systems in the process. UNESCO is therefore a platform that brings different value systems together. It further helps strengthen the capacity of Member States to harness AI for the benefit of humanity and for achieving the SDGs in line with ethical principles particularly in the areas of education, the sciences, culture and communication and information. It creates awareness and information about the participation of scientists and engineers (women and men) in all regions in the world in the development of AI science and technologies, encouraging sound science, technology and innovation ecosystems and building endogenous capacities, particularly for higher education institutions. This commitment to a human-centred approach to digital technologies is reflected in UNESCO's framework of "Internet Universality" and the associated "R.O.A.M. principles" (Human Rights, Openness, Accessibility and Multi-stakeholder participation), which were endorsed by the Organization's Member States in 2015. UNESCO's approach to digital issues within this framework helps towards understanding part of the key ecosystem of AI. An example is a concrete tool for measuring R.O.A.M. at country level in the form of Internet Universality Indicators, agreed by the Member States. The value of using the R.O.A.M. frame for assessing AI in particular is also reflected in the publication entitled "Steering AI and Advanced ICTs for Knowledge Societies: a ROAM perspective", which was launched at the Internet Governance Forum in 2019.

15. UNESCO's AI work includes establishing a network of UNESCO Chairs and Category II Centres, such as the International Research Centre on Artificial Intelligence (IRCAI), providing policy fora and engaging in special partnerships. It works with AI laboratories and the private sector to develop innovative and efficient projects on the ground related to meeting the SDGs and to harnessing AI in UNESCO's fields of competence. The Organization is developing model curricula and training modules for the beneficiaries of UNESCO Major Programmes, as well as reviewing and contributing to the diffusion of good practices in the application of AI technologies to fields such as water resources and ecosystems management, or disaster risk reduction. UNESCO also works with AI designers and professional associations to promote relevant guidelines and ethical processes as well as ethical codes, as well as to ensure an approach of ethics and human rights by design for AI, without stifling innovation. In line with its Operational Strategy on Youth (2014-2021), UNESCO also recognizes young people as knowledge-holders and innovators, and works in partnership with youth and their organizations. The youth of today stand at the vanguard of the digital innovations of tomorrow. At the same time, young people have valid concerns relating to ethical issues of AI. UNESCO works to promote the discussion and consideration of concepts and concerns voiced by youth populations across the world.

IV. TASKS OF THE AD HOC EXPERT GROUP RELATED TO THE PREPARATION OF A FIRST DRAFT TEXT OF A RECOMMENDATION ON THE ETHICS OF AI

16. The preparation of a UNESCO draft text of a recommendation is guided by the Rules of Procedure concerning recommendations to Member States and international conventions covered by the terms of Article IV, paragraph 4, of the Constitution.

17. Pursuant to 40 C/Resolution 37, 206 EX/Decision 42 and 207 EX/Decision 5.I.A, the Ad Hoc Expert Group (AHEG) was constituted by UNESCO's Director-General to advise in the preparation of a draft text of a recommendation on the ethics of AI. The AHEG's main objectives are:

i. As a group, the experts should collectively elaborate the first draft text of a recommendation that fully covers the topic. The text should be of a nature and quality that can be translated and transmitted directly to a wide range of stakeholders, including Member States, for their consideration. This task is to be completed during the virtual discussion of the AHEG in April 2020. The first draft

will then undergo extensive multi-stakeholder consultations at national, regional and global levels, as well as online consultations from May to July 2020. These broad consultations will contribute to ensuring that the elaboration of the draft text of a recommendation is inclusive, multicultural, pluralistic and multi-stakeholder shaped.

ii. The experts should collectively revise the first draft so as to reflect the comments provided by the multi-stakeholder consultations, and thus elaborate a second draft text of a recommendation, reporting to the Director-General no later than mid-September 2020. This task is to be completed at the meeting of the AHEG that could be convened for five working days in late August 2020 or first week of September at the latest. The exact dates of the meeting would be decided on by UNESCO in agreement with the Chairperson of the AHEG and its Bureau. Member States and relevant international and regional organizations could participate as observers at the meeting. The UNESCO Secretariat is to finalize the work on the second draft by mid-September 2020 for its transmission to Member States for comments.

18. UNESCO will provide the Secretariat for the AHEG to assist in the preparation of its reports, to be responsible for preparing, translating and distributing all official documents of the AHEG, and undertaking all practical arrangements for its meetings.

19. AHEG is independent when deciding on the content and structure of the draft text of a recommendation. However, based on the outcome of intergovernmental discussions and negotiations in 2021, the final recommendation will be agreed by representatives of UNESCO's Member States and adopted at its General Conference.

V. SOURCES TO BE CONSIDERED IN ORDER TO FULFILL THE MAIN OBJECTIVES OF THE AHEG

20. In order to meet the objectives described above, each AHEG member will consider a variety of source materials and apply his/her own expertise to arrive at conclusions. The sources which shall be considered by the AHEG members, so that their results meet UNESCO's requirements, include UNESCO and United Nations system sources.

21. In addition to the UNESCO and UN system-wide sources, sources of other international organizations and Member States' policy documents may be included. Thus, the AHEG, in order to fulfil its task, needs to thoroughly elucidate whatever tacit and explicit principles and approaches are propelling the multi-stakeholders' undertakings in national, regional and international settings. This should cover organizations from the Global South,[4] as well as initiatives within the commercial and non-commercial sectors.

22. Although the current endeavour is a recommendation, the AHEG can also refer to international declarations as well as other publications and documents, as deemed appropriate by members of the AHEG.

23. An indicative list of documents is available in Annex 5.

VI. POSSIBLE SUBSTANCE OF THE OUTCOME DOCUMENT OF THE VIRTUAL DISCUSSION OF THE AHEG (APRIL 2020)

[4] In particular, the African Union and the Arab League have established working groups on AI. The former focuses on three main objectives: 1) The creation of a common African stance on AI; 2) The development of an Africa wide capacity building framework; and 3) Establishment of an AI think tank to assess and recommend projects to collaborate on in line with Agenda 2063 and the UN Sustainable Development Goals. See https://au.int/en/pressreleases/20191026/african-digital-transformation-strategy-and-african-union-communication-and.

24. To meet its aims, in the context of this process, the text resulting from the AHEG's virtual discussion should identify and clarify a set of ethical principles, their interlinkages and policy actions of the international community to address issues around development, implementation and application of AI, keeping in mind that AI technologies are not value-neutral.

25. Some of the key questions to be mindful of when preparing the first draft: Who is it intended for, and what is its purpose? Why should it be followed? How to follow or implement it? How should the conflicting interpretations of essentially contested concepts be resolved? How will we know that the recommendations are being implemented? What if recommendations are not applied? How can disagreements or questions for clarification be raised?[5]

26. The recommendation under elaboration needs to remain flexible to accommodate new opportunities, risks and challenges as they emerge due to continued developments in AI and its applications both in the public and private sector. Furthermore, the recommendation must also address the concerns of developing countries and resource-poor settings in developed countries, the good of present and future generations, the 2030 Agenda for Sustainable Development, gender and cultural, including religions and spiritual bias, inequalities between and among countries, and leaving no one behind. The recommendation, while incorporating an inclusive multi-stakeholder approach, must also consider areas related to UNESCO's mandate: education, the sciences, culture, communication and information, with additional focus on the two global priorities of the Organization, namely gender equality and priority Africa, as well as a focus on priority groups.

VI.1 Defining AI

27. It is not the task of the AHEG and, further, UNESCO, to provide for an authoritative definition of AI. In fact, it might be counterproductive, as AI can refer to different things depending on the focus and the context, and there is currently no consensus regarding the various definitions developed throughout decades of AI research and applications. It is not necessary nor possible to settle this debate during the virtual discussion.

28. Nevertheless, for the purposes of the draft text of a recommendation, an option to consider would be to adopt a functional approach covering current and possible future applications of AI technologies, akin to the one adopted by COMEST in its 2019 Study referring to "machines capable of imitating certain functionalities of human intelligence, including such features as perception, learning, reasoning, problem solving, language interaction, and even producing creative work." It is also possible to take a minimalist approach, as e.g. the European Commission did in its 2020 White Paper on Artificial Intelligence, calling it "a collection of technologies that combine data, algorithms and computing power". It must be noted, however, that even the latter minimalist approach is open to criticism as one can imagine AI not requiring the combination of all of the three factors.

VI.2 The approach to identifying principles and policy actions

29. Principles and policy recommendations must be firmly grounded in the international human rights framework. The UN Secretary-General's Strategy on New Technologies underlines that technologies like AI must be aligned with the values enshrined in the UN Charter, the Universal Declaration of Human Rights, and the norms and standards of international law.[6] When considering the wider digital and internet ecosystem required for the design, development and deployment of AI, the Internet Universality framework endorsed by the UNESCO General Conference in 2015 and underlined as an important framework by the Secretary-General's High Level Panel on Digital Cooperation for developing frameworks for

[5] Adapted from B. Mittelstadt, "Principles alone cannot guarantee ethical AI", *Nature Machine Intelligence* 1, 501-507 (2019), p. 504.
[6] UN Secretary-General's Strategy on New Technologies, September 2018, https://www.un.org/en/newtechnologies/.

digital governance going forward, provides a strong foundation for articulating principles that will ensure a universal approach to AI. These principles guide stakeholders in steering AI for knowledge societies from the perspective of Human Rights, Openness, Accessibility and Multi-stakeholder participation (R.O.A.M. principles).

30. Various documents have identified multiple principles for ethical AI, which often coincide. There is an emerging convergence on a number of ethical principles that can be mainstreamed and adopted throughout the whole process of development, deployment and uptake of AI. Nevertheless, there are many gaps and contradictions (which are to be discussed later) including the uncertainty of prioritization of principles and identifying the missing ones. A way to address these problems is to boil down to foundations.[7]

31. With this in mind and drawing on the range of principles identified by the various initiatives both within the UN system and beyond, the proposed framing for discussion is to identify foundational values and closely interlinked principles. Foundational values have a role of necessary preconditions, prerequisites, or otherwise *conditiones sine quibus non* for principles to work and, in effect, to ensure ethical AI. The principles *per se* would then address specific narrower areas of concern. The foundational values also have a role of a starting point allowing for identifying relevant principles and, subsequently, policy actions to implement them separated by the relevant stakeholder groups concerned. The principles, while informed by the foundational values, unpack them in a more detailed manner regarding specific areas of ethical concern, which when implemented as a whole ensure alignment with the foundational values.

VI.3 Foundational values

32. As explained above, it is suggested that respect for human rights and fundamental freedoms should be considered as one of the foundational values as development, deployment and uptake of AI technologies must occur in accordance with international human rights standards. Among others, this would address upfront such issues as ensuring that data mining and analysis is done in a manner that respects human agency and privacy, development of free and independent thought, ensuring no insidious attempts to influence human behaviour, etc. Having such concerns as a foundational value will guide the adoption and mainstreaming of a human rights-based approach to AI in Member States. This foundational value is also consistent with the approach taken across the UN system. An example is the Human Rights Council's resolution on "The right to privacy in the digital age" (A/HRC/RES/42/15) and the Report of the UN High Commissioner for Human Rights on this issue (A/HRC/39/29), which could be expanded to other human rights.

33. In order to ensure an inclusive AI, it is crucial that issues such as discrimination and bias, including on the basis of gender, as well as diversity, digital and knowledge divides are addressed. This is why leaving no one behind could be considered as another foundational value throughout the AI system lifecycle. Thus, the development and use of AI systems must be compatible with maintaining social and cultural diversity, different value systems, take into consideration the specific needs of different age groups, persons with disabilities, women and girls, disadvantaged, marginalized and vulnerable populations and must not restrict the scope of lifestyle choices or personal experiences. This also raises concerns about neglecting local knowledge, cultural pluralism and the need to ensure equality. The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity. Particular attention must be paid to the lack of necessary technological infrastructure and legal frameworks in low-income countries and to ensuring that they benefit from and participate equally in the AI ethics debate.

34. The 2030 Agenda for Sustainable Development acknowledges the social, economic and environmental dimensions of sustainable development, which must be addressed through integrated approaches. Analysis of various documents proposing ethical principles for AI

[7] This does not deny the fact that certain assumptions must be made, which may change with time as our understanding of the world advances.

shows that protection of the environment receives little attention or is overlooked. However, (1) the environment is the existential necessity for the humanity to be able to enjoy the benefits of advances in AI in the first place; (2) certain societies do not see human being as separate from the environment or even provide agency to the latter;[8] (3) there is an unprecedented rising international commitment to preservation of environment, including prominently due to the fight against climate change. The major AI methods currently employed are based on processing of ever-increasing amounts of data, which leads to increasing energy consumption and consequential impact on carbon emission, depletion of resources and deterioration of the environment.[9] This is unsustainable in the long run. Elevating environmental concerns related to AI technologies to the same level as the other two foundational values is therefore a necessity. Simultaneously, this can stimulate the development of AI-based solutions to prevent harm to the environment. Building on ground-breaking international achievements, including the 2016 Paris Agreement and the 2017 UNESCO Declaration of Ethical Principles in relation to Climate Change, elevating this principle can become an unequivocal concrete added value that the whole UN system and UNESCO in particular can provide in forming the global discourse around AI.

35. The basics of the overall approach can be visualized as follows (Figure 1).

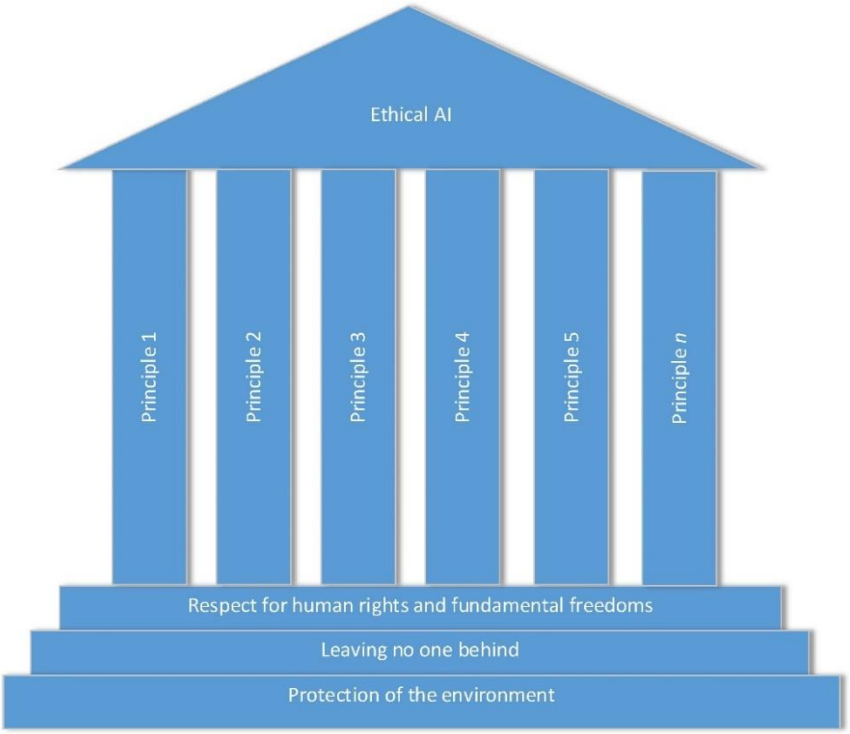

Figure 1. Ethical AI Approach[10]

[8] See e.g. the example of the Whanganui River recognized as a legal entity with corresponding rights, powers and obligations under the New Zealand law *Te Awa Tupua (Whanganui River Claims Settlement) Act* of 2017 : http://www.legislation.govt.nz/act/public/2017/0007/latest/whole.html.

[9] Data extraction consumes nearly 10% of energy globally. See P. Corcoran, A. Andrae, "Emerging Trends in Electricity Consumption for Consumer ICT", 2013. https://aran.library.nuigalway.ie/xmlui/bitstream/handle/10379/3563/CA_MainArticle14_all-v02.pdf?sequence=4.

[10] The visualization resembles the UNESCO emblem, but there are other ways to visualize the approach, such as portraying it through the mythology available in various parts of the world where the Earth stands on three animals depending on the region (turtles, whales, elephants, etc.).

VI.4 Principles

36. There is an emerging convergence around the importance of certain principles that allow to ensure ethical approaches within the whole process of research, development, deployment and uptake of AI. In particular, the follow-up roundtable discussion on Recommendation 3C regarding AI of the UN Secretary-General's High Level Panel on Digital Cooperation has preliminarily identified consensus on the following fifteen principles: accountability, accessibility, diversity, explainability, fairness and non-discrimination, human-centricity, human-control, inclusivity, privacy, reliability, responsibility, safety, security, transparency, and trustworthiness. COMEST has also identified the following generic principles, many of which overlap with the ones above: human rights,[11] inclusiveness,[12] flourishing, autonomy, explainability, transparency, awareness and literacy, responsibility, accountability, democracy, good governance, and sustainability.[13]

37. Nevertheless, the variety of source materials shows that both the notions and substance of principles may vary significantly. The multi-interpretability of certain terms (in fact, "AI" itself) is problematic in many respects. Moreover, conceptual and procedural divergences in the sets of principles reveal uncertainty as to which ethical principles should be prioritized and how conflicts between ethical principles should be resolved (e.g. larger datasets to unbias AI vs privacy, avoiding harm vs accepting some degree of harm (risk-benefit), etc.).

38. In particular, there is often confusion within the following two sets of principles: *responsibility* and *accountability*; *transparency* and *explainability*. The doctrinal understanding of responsibility alone has at least four different interpretations, e.g. role-responsibility, causal responsibility, liability-responsibility and capacity-responsibility.[14] The two understandings important for the purposes of this recommendation are causal responsibility and liability. The former is about a result being attributed to some event or agent, which is held to be the cause of the result. The latter is about bearing the consequences. Therefore, when providing for a principle of responsibility it is important to make it clear whether it is about attributing an act to an agent or attributing the consequences to an agent, as these can differ, and one does not presuppose the other. In order to avoid confusion, different terminology can be considered altogether. COMEST somewhat departs from common use of such terminology and provides the following: responsibility – "developers and companies should take into consideration ethics when developing autonomous intelligent system"; accountability – "arrangements should be developed that will make possible to attribute accountability for AI-driven decisions and the behaviour of AI systems". In addition, the AHEG might wish to be aware of the possibility to consider multi-stakeholder governance, accepted by UNESCO Member States, as a process modality for underpinning the ethical principles of accountability, responsibility, transparency and explainability.[15]

39. Further, several principles are closely interlinked and are dependent on each other. Thus, *human control* (also known as "autonomy") is dependent on attribution of consequences as situations may arise where control at all times is not possible. This in turn requires establishment of necessary and sufficient conditions for such control.

40. Concern for *gender* equality is sometimes bundled as a subset of a general concern for bias in the development of algorithms, in the datasets used for their training, and in their use in decision-making. However, gender is a much larger concern, which includes, among others, women empowerment linked with their representation of women among developers,

[11] Suggested by this background document as a foundational value in a different formulation.
[12] Suggested by this background document as a foundational value in a different formulation.
[13] Suggested by this background document as a foundational value in a different formulation.
[14] HLA Hart, *Punishment and responsibility*, 2nd edn. Oxford University Press, 2008, pp. 210-237.
[15] These issues, including relevance to AI, are assessed in the UNESCO publication "What if we all governed the Internet? Advancing multistakeholder participation in Internet governance" https://unesdoc.unesco.org/ark:/48223/pf0000259717.

researchers and company leaders, as well as access to initial education and training opportunities for women,[16] which requires the consideration of gender equality as a stand-alone principle.

41. *Trustworthiness* is essential for societies around the world to accept the new technologies. Typically, the form of the question that is asked about such technologies is "Can computers do X?" as it was at the inception of AI research.[17] What is need to be done within the draft text of a recommendation is to shift the question to "What does it take for computers to do X?" Essentially, what are the ethical and, broader, social, economic, political and cultural conditions for computers to be considered or accepted as doing X.

42. The idea of *flourishing* follows from the UN SDGs. There are multiple metrics in use that measure well-being through Indicators such as the UN Human Development Index or Genuine Progress Indicator.

43. *Privacy* is usually featured as a concern tied to data usage, particularly big data. Since the major AI methods currently employed are based on processing of ever-increasing amounts of data, privacy becomes a particular concern in terms of both the use of data for the development of AI systems and providing impacted people with agency over their data and decisions made with it.[18]

44. Particular attention must be paid to geographic areas such as Africa, Latin America and the Caribbean, Small Islands Developing States, and Central Asia, as they are underrepresented and are not participating equally in the AI ethics debate.[19] This raises concerns about neglecting local knowledge, cultural and ethical pluralism, value systems and the demands of global fairness.

45. Some principles have not yet achieved a widespread adoption and can be considered as emerging principles. *Solidarity* could be an important example, which is not always found in relevant documents, while, as suggested by the UN Global Pulse Chief Data Scientist, it "should be a core ethical principle of AI".[20] Although solidarity is usually featured in relation to the implications of AI for the labour market, it can have a larger significance in terms of international cooperation, notably supporting AI development and redistribution of the benefits of AI.

46. Last but not least, when considering various principles, it is always about balance of input, output and the process surrounding it. It is not just about how we design AI technologies but how we define their success. Being safe, secure, compliant to legislation, and economically beneficial can still have negative consequences on human rights, identity, autonomy, dignity, mental health, etc. The purpose of ethics is to highlight that just because something is technically feasible does not mean it should be developed.

47. COMEST has also identified some central ethical concerns regarding the specific focus of UNESCO: education, science, culture, communication and information, peace, Africa, gender, environment. Some of these concerns have been featuring as specific principles in various documents. These are available in the table in Annex 2.

48. All of the aforementioned and other issues are to be addressed by the AHEG.

---

[16] See e.g. "I'd blush if I could: closing gender divides in digital skills through education", EQUALS and UNESCO, 2019.
[17] A.M. Turing, "Computing Machinery and Intelligence", *Mind 49*, 433-460.
[18] There are ethical issues tied to intellectual property rights in this regard, among others. However, the protection of intellectual property is the mandate of the World Intellectual Property Organization (WIPO).
[19] A. Jobin, M. Ienca and E. Vayena, "The global landscape of AI ethics guidelines", *Nature Machine Intelligence* 1, 389-399 (2019), p. 396.
[20] M. Luengo-Oroz, "Solidarity should be a core ethical principle of AI", *Nature Machine Intelligence* 1, 494 (2019).

VI.5 Making principles actionable

49. It must be noted that the most challenging task for the AHEG is not identification of principles but clarifying their meaning and making sure that they work in practice. In fact, many available AI ethics guidelines remain vague and hard to implement. With this draft text of a recommendation, the aim is to move on from the high-level statements that have been produced so far, as the recommendation will only work if the principles identified therein are actionable. In order to ensure that, a number of policy actions addressed to Member States and other stakeholders must be proposed. In particular, Member States should be recommended to take certain actions that apply to the private sector. Some policy actions within the draft text of a recommendation should be addressed to the private sector directly, especially in the case of transnational practices. Recommendation as such can guide various entities in their internal policymaking. Some potentially relevant policy actions are listed in Annex 4. Certain policy actions can address several principles at a time.

50. When considering possible policy actions addressed to Member States, the CEB's "UN system-wide strategic approach and roadmap for supporting capacity development on artificial intelligence" could provide an entry point. In this regard, it should be emphasized that ethics should not be seen as a mere list of principles, but also a process of decision-making of what should be considered acceptable or not. Hence, ethics cannot be seen as separate from designing, application and evaluation of AI but should instead inform capacity building work of the UN system. With this in mind, a number of the commitments and measures outlined in the roadmap, although addressed to the UN system, could also be relevant for Member States. These policy actions could be of a more general and overarching nature, relevant not only for UNESCO's areas of work, but also for work across the UN system.

51. The rapid development of AI technologies requires providing support to Member States that struggle to keep up with the tremendous pace of innovation and change. UNESCO must work on providing technical assistance and capacity building to policy makers and governments on implementation the ethical values and principles as identified in the recommendation. Thus, another area of possible policy actions from the roadmap that could be considered is for Member States to engage in capacity building as outlined by the system-wide strategic approach. This could be viewed as both a practical and ethical approach that cuts across all areas of the UN system. It would also require international cooperation among Member States to avoid deepening inequalities as well as technological and knowledge divides, within and between countries.

52. An additional area of possible policy actions that could be considered is the ethical dimension of AI governance at the national level in terms of addressing common concerns and issues for work across the UN system. Governance is understood as making decisions and exercising authority in order to guide the behaviour of individuals and organizations.[21] In the UN perspective, corresponding to the World Summit of the Information Society, it covers shared principles, norms, rules, decision-making processes and programmes that shape the evolution and use of digital systems. Underpinning such governance are institutional arrangements that set standards and create incentives for behaviours corresponding to the identified principles ensuring ethical AI. Governance strategies can be various: active cooperation across disciplines and stakeholders, compliance, oversight processes and practices (tests, monitoring, audits and assessments by internal units, customers, users, independent third parties or governmental entities, often geared towards standards for AI implementation and outcome assessment, etc.).

53. As one of the important elements thereof, a number of initiatives have identified the need for some form of risk-benefit assessment, especially when considering the use of AI technologies. Two consecutive fundamental questions that such an assessment should address are 1) whether the use of AI technologies within a particular area of the public sector

[21] See Report of the Working Group on Internet Governance (WGIG), June 2005, p. 4. See also R. Baldwin, M. Cave, and M. Lodge, Oxford Handbook on Regulation, Oxford University Press, 2010.

is appropriate, and if so 2) what the appropriate AI method is. Answers to these questions have direct effect on the foundational values. For example, data-intensive AI methods have bearing on human rights (e.g. using human beings as a resource), inclusivity (e.g. disproportionate availability of data on different groups), and environment (e.g. increased energy consumption). In certain situations, it could be more beneficial to use AI methods that require only limited amounts of data, do not require data at all, or not to use the technology at all. In this regard, the evaluation methodology should identify and assess benefits and risks, as well as risk mitigation and monitoring measures. At a minimum, risks assessment should identify impacts on human rights, the environment, and related ethical and social implications. The approach for such an assessment should also be multidisciplinary, multi-stakeholder, multicultural, pluralistic and inclusive. The requirement for such an assessment could be framed as a policy action by Member States, and the guidelines with appropriate methodology can be developed in cooperation with the relevant UN entities, depending on which area of work is being considered.

54. There are different types of impact assessment, particularly prominent ones being regulatory impact assessment and environmental impact assessment. The latter is used as a "systemic approach to critically assessing the positive and negative effects of proposed and existing regulations and non-regulatory alternatives."[22] It employs such methods as cost-benefit analysis, cost-effectiveness analysis, multi-criteria analysis and others. Conducting regulatory impact assessments within an appropriate systematic framework can underpin the capacity of governments as well as international organization to ensure that regulations and non-regulatory alternatives are efficient and effective, especially in the times of digital disruption. Environmental impact assessment "evaluates the effects of human intervention on the biophysical environment by considering the intended (products) and unintended (waste) consequences of industry."[23] Both kinds of impact assessments can be conducted *ex ante* or *ex post*.

55. The current recommendation, albeit along the lines of other impact assessments, can benefit from a different kind of impact assessment – an ethical impact assessment (EIA). An ethical, value-based analysis will allow to predict consequences, mitigate risks, avoid harmful consequences, facilitate participation and address societal challenges in line with the principles identified in the recommendation.[24] Therefore, it is not only about assessing the data employed to train an algorithm, but impact on everyone who might be affected by its decisions. Since ethics has different reasoning methods, applying them can lead to different conclusions. EIA can lift a discussion about the design or implementation of an AI system to a higher level and help to make the right choices, including on ethical use of AI. Capacity building is also necessary to support EIA efforts by Member States so that this tool delivers the results. UNESCO can develop a framework and guidelines for EIA that can be used by Member States and other stakeholders to plan, identify, evaluate the impact and mitigate risks where necessary.

56. "For the whole of society to truly be able to benefit from all AI developments, education and an honest and accessible AI narrative are needed. Only then, will everybody be able to understand AI's impact and truly benefit from its results."[25] Therefore, there is a need to include steps to ensure proper and wide value-based education of all stakeholders present and future. "Informed participation of all stakeholders, which means that education plays an important role,

---

[22] Regulatory Impact Analysis, OECD, http://www.oecd.org/gov/regulatory-policy/ria.htm.
[23] R.A. Calvo, D. Peters and S. Cave, "Advancing impact assessment for intelligent systems", *Nature Machine Intelligence* 2, 89-91 (2020), p. 89.
[24] Consider also the "human impact assessment for technology (HIAT)". See R.A. Calvo, D. Peters and S. Cave, "Advancing impact assessment for intelligent systems", *Nature Machine Intelligence* 2, 89-91 (2020).
[25] V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer, 2019, p. 51.

both to ensure that knowledge of the potential impact of AI is widespread, as well as to make people aware that they can participate in shaping the societal development".[26] Thus, education policy addresses both the principles of inclusiveness and trustworthiness.

57. To ensure that values and principles are being followed, AHEG might consider the possibility to recommend organizations, including commercial ones, to invest in establishing advisory panels and hiring ethics officers. It can also be considered that ethics officers and advisers are able to veto any projects or deliverables that do not adhere to the processes for ethical decision-making and the principles thereof.

VII. POSSIBLE FORMAT OF THE OUTCOME DOCUMENT OF THE VIRTUAL DISCUSSION OF THE AHEG

58. The text resulting from the AHEG's virtual discussion should therefore identify and clarify a set of ethical principles and policy actions to implement them with regard to development, implementation and use of AI.

i. Ideally, the form of the text will be short, accessible and easily communicated to a variety of audiences, including after its translation into many languages.
ii. The text should be precise and use common terms that are aligned with technical terms that appear in other international texts.
iii. There is also concern that the first draft be complete and balanced, covering all the essential principles and policy actions of approximately equal importance, without neglecting or misrepresenting any important matters.
iv. The AHEG should consider its work as aiming at clarifying ethical principles and process for ethical decision-making for the international community in such a way as to make it easier to adapt these ethical principles to practical uses through subsequent policy actions, implemented not just by states but also by a variety of actors at various levels.

59. The recommendations adopted by UNESCO follow different formats, as can be seen on the web page of UNESCO that contains the texts of all recommendations. Several recommendations are composed of: a preamble; general provisions, covering the scope and aims; the principles; a description of how these principles will be applied; measures aimed at the promotion of a recommendation and follow-up action by UNESCO; as well as final provisions. Others contain a preamble; main principles; and measures of implementation. The AHEG will have to decide on the format of the outcome document that **it should draft by the end April 2020** (a suggested provisional skeleton of an outcome document is presented in Annex 1).

VIII. TABLES OF EXAMPLES TO GUIDE REFLECTION

60. The tables of examples annexed to this document (Annexes 3 and 4) are conceived as a working tool that may serve the AHEG in its preparations for the virtual discussion. In Annex 3, the ethical principles identified by COMEST in the 2019 study have been entered as a baseline (first rows) to facilitate the work of the AHEG and are followed by similar principles and their interpretations available in other documents. Thus, the AHEG may consider the extent to which COMEST's work presents a starting point for its reflection. In addition, the same table continues with a set of other potentially relevant principles or sub-principled contained in various international documents, and which are not found in the COMEST study. Finally, Annex 4 provides for a table with a number of potentially relevant practical policy actions which could serve as a reference for the policy actions to be elaborated for the purposes of the draft text of a recommendation.

[26] A. Theodorou and V. Dignum, "Towards ethical and socio-legal governance in AI" *Nature Machine Intelligence* 2, 10-12 (2020), p. 11.

61. The AHEG is invited to consider these tables together with this background document as its working tool, modifying it as members deem appropriate. This will entail:

i. reworking the set of principles and policy actions, by identifying others, and re-ordering them. (Until the principles and policy actions are conceptually ordered, it is evident that they will appear in this table with repetitions, reflecting the variety of sources.);

ii. re-formulating key messages so as to develop a complete and well-ordered set containing the responsibilities that have been identified. This can be achieved either by selecting and reworking the early proposals that appear in the current document, or by beginning afresh.

iii. deriving policy actions from broad principles where that is feasible, so as to make the ethical principles easier to adapt to practical uses, not just by states but also by a variety of actors at various levels; and

62. It is the final step of ordering and assembling the key messages that may bring to life the prototype of a first version of a draft text of a recommendation.

United Nations
Educational, Scientific and
Cultural Organization

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

April 2020

ANNEX 1: PROVISIONAL SKELETON OF AN OUTCOME DOCUMENT

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

(to be written by the AHEG by the end of April 2020)

Preamble

The Member States of the United Nations Educational, Scientific and Cultural Organization (UNESCO), meeting in Paris at the forty-first session of the General Conference, from … to … November 2021,

Recalling …….

Considering …….

Recognizing …….

Also recognizing …….

Observing ……

Persuaded ……

Believing ……

Convinced ……

Conscious ……

However, taking fully into account ……

Desiring ……

Having ……

Reflecting on …….

Resolving ……

Noting …….

Aware that ……

Bearing in mind ……

Stressing ……

Having decided ……

Adopts the Recommendation on the Ethics of Artificial Intelligence;

Recommends ……

Also recommends ……

Further recommends ……

I. SCOPE OF APPLICATION

1. For the purposes of this Recommendation:

…

2. This Recommendation applies with respect to:

…

II. AIMS AND OBJECTIVES

3. …

4. …

III. FOUNDATIONAL VALUES FOR ETHICAL DEVELOPMENT, DEPLOYMENT AND UPTAKE OF AI TECHNOLOGIES

Respect of human rights and fundamental freedoms

5. …

6. …

Leaving no one behind

7. …

8. …

Protection of the environment

9. …

10. …

IV. PRINCIPLES

Principle 1

11. …

12. …

Principle 2

13. …

14. …

Principle 3

15. …

16. …

Principle 4

17. …

18. …

Principle 5

19. …

20. …

Principle 6

21. …

22. …

Principle 7

23. …

24. …

Principle 8

25. …

26. …

Principle 9

27. …

28. …

Principle 10

29. …

30. …

V. POLICY ACTIONS

Policy action 1

31. …

32. …

Policy action 2

33. …

34. …

Policy action 3

35. …

36. …

Policy action 4

37. …

38. …

Policy action 5

39. …

40. …

VI. MONITORING AND EVALUATION

41. …

42. …

VII. UTILIZATION AND EXPLOITATION OF THE PRESENT RECOMMENDATION

VIII. PROMOTION OF THE PRESENT RECOMMENDATION

43. …

44. …

IX. FINAL PROVISIONS

45. …

ANNEX 2: UNESCO-SPECIFIC CENTRAL ETHICAL CONCERNS

	Ethical concern	*Key message*	*Source*
1	**Education**	AI requires that education fosters AI literacy, critical thinking, resilience on the labour market, and educating ethics to engineers.	COMEST 2019
2	**Science**	AI requires a responsible introduction in scientific practice, and in decision-making based on AI systems, requiring human evaluation and control, and avoiding the exacerbation of structural inequalities.	COMEST 2019
3	**Culture**	AI should foster cultural diversity, inclusiveness and the flourishing of human experience, avoiding a deepening of the digital divide. A multilingual approach should be promoted.	COMEST 2019
4	**Communication and information**	AI should strengthen freedom of expression, universal access to information, the quality of journalism, and free, independent and pluralistic media, while avoiding the spreading of disinformation. A multi-stakeholder governance should be promoted.	COMEST 2019
5	**Peace**	In order to contribute to peace, AI could be used to obtain insights in the drivers of conflict, and should never operate out of human control.	COMEST 2019
6	**Africa**	AI should be integrated into national development policies and strategies by drawing on endogenous cultures, values and knowledge in order to develop African economies.	COMEST 2019
7	**Gender**	Gender bias should be avoided in the development of algorithms, in the datasets used for their training, and in their use in decision-making.	COMEST 2019
8	**Environment**	AI should be developed in a sustainable manner taking into account the entire AI and IT production cycle. AI can be used for environmental monitoring and risk management, and to prevent and mitigate environmental crises.	COMEST 2019

ANNEX 3: SUMMARY TABLE OF POSSIBLE PRINCIPLES TO GUIDE REFLECTION

	Principle	***Key message***	***Sources***
Potentially relevant principles			
1	**Human rights**[27]	AI should be developed and implemented in accordance with international human rights standards.	COMEST 2019
	Principle of human rights	All artificial intelligence-related capacity-building programming by United Nations entities should respect the principles of human rights, thereby helping to ensure that a human rights-based approach should be mainstreamed into the approach to artificial intelligence adopted by Member States	CEB 2019
	Human dignity	Dignity is inherent to human beings, not to machines or robots. Therefore, robots and humans are not to be confused even if an android robot has the seductive appearance of a human, or if a powerful cognitive robot has learning capacity that exceeds individual human cognition. Robots are not humans – they are the result of human creativity and they still need a technical support system and maintenance in order to be effective and efficient tools or mediators.	COMEST 2017
	Rights-based	[Internet] rooted in the Universal Declaration of Human Rights and its associated Covenants.	UNESCO General Conference 2015 decision on the Internet Universality
	Principle of respect of fundamental rights	Ensuring that the design and implementation of AI tools and services are compatible with fundamental rights.	CoE Ethical Charter 2018
	Human-centered values and fairness	i. AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality,	G20 AI Principles 2019 = OECD AI Principles 2019

[27] Suggested by this background document as a foundational value in a different formulation.

	Principle	*Key message*	*Sources*
		diversity, fairness, social justice, and internationally recognized labor rights. ii. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.	
	Human rights	Ensure autonomous and intelligent systems do not infringe on internationally recognised human rights.	IEEE's Ethically Aligned Design 2019
	Secure a just transition and ensure support for fundamental freedoms and rights	As AI systems develop and augmented realities are formed, workers and work tasks will be displaced. It is vital that policies are put in place that ensure a just transition to the digital reality, including specific governmental measures to help displaced workers find new employment.	UNI Global Union 2017
2	**Inclusiveness**[28]	AI should be inclusive, aiming to avoid bias and allowing for diversity and avoiding a new digital divide.	COMEST 2019
	Accessibility	[Internet] accessible to all, in both infrastructure and content.	UNESCO General Conference 2015 decision on the Internet Universality
	Diverse perspectives on the benefits and risks of AI technologies	Artificial intelligence-related capacity-building programming should gather diverse perspectives on the benefits and risks of artificial intelligence technologies and take into consideration the needs of all people, including those at risk of being left behind, especially those who are marginalized and vulnerable. People and particularly those farthest behind, including women and girls, should be at the centre of all artificial intelligence-related capacity-building programming and decision-making processes.	CEB 2019
	"Whole-of-government" and "whole-of-society" approach	Artificial intelligence-related capacity-building programming should strive to foster a "whole-of-government" and a "whole-of-society" approach, in particular in taking into account the bottom billion.	CEB 2019

[28] Suggested by this background document as a foundational value in a different formulation.

	Principle	***Key message***	***Sources***
	Multi-stakeholder partnerships	Artificial intelligence-related capacity-building programming should make efforts to strengthen multi-stakeholder partnerships, especially between Governments, private sector, international organizations, civil society and academia.	CEB 2019
	Cooperation and synergy	All artificial intelligence-related programming by United Nations entities should actively seek cooperation and synergy with complementary developmental programmes that deliver other key elements in order to reach common goals.	CEB 2019
	Diversity inclusion principle	The development and use of AI systems must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices or personal experiences	Montreal Declaration 2018
	Inclusive growth, sustainable development and well-being	Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.	G20 AI Principles 2019 = OECD AI Principles 2019
	Share the benefits of AI systems	The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity. Global as well as national policies aimed at bridging the economic, technological and social digital divide are therefore necessary.	UNI Global Union 2017
	Fairness and non-discrimination	With concerns about AI bias already impacting individuals globally, Fairness and Non-discrimination principles call for AI systems to be designed and used to maximize fairness and promote inclusivity. Fairness and Non-discrimination principles are present in 100% of documents in the dataset.	Berkman Klein Center 2020
3	**Flourishing**	AI should be developed to enhance the quality of life.	COMEST 2019
	Balancing economic, social and environmental goals	Artificial intelligence-related capacity-building programming should balance economic, social and environmental goals: reducing inequalities and ensuring equal access to opportunities, promoting productive transformation of the economy and	CEB 2019

	Principle	***Key message***	***Sources***
		protecting the natural environment. Such a process generates social justice within and between generations, sustainable development, peace and prosperity.	
	Value of beneficence	Robots are useful for facilitating better safety, efficiency, and performance in many human tasks that are physically hard. Industrial robots, disaster robots, and mining robots can be used to replace human beings in dangerous environments. However, the beneficence of robots is subject to further discussion and reflection when they are designed to interact in a social context, such as in education, health care or surveillance/policing by the State.	COMEST 2017
	Well-being principle	The development and use of AI systems must permit the growth of the well-being of all sentient beings	Montreal Declaration 2018
	Prioritising well-being	Prioritise metrics of well-being in the design and use of AISs because traditional metrics of prosperity do not take into account the full effect of AI systems technologies on human well-being	IEEE's Ethically Aligned Design 2019
	Promotion of human values	Human Values principles state that the ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well-being. Promotion of Human Values principles are present in 69% of documents in the dataset.	Berkman Klein Center 2020
	Beneficence	While promoting good is often mentioned, it is rarely defined, though notable exceptions mention the *augmentation of human senses, the promotion of human well-being and flourishing, peace and happiness*, the creation of socio-economic opportunities, and *economic prosperity*. Similar uncertainty concerns the actors that should benefit from AI: private sector issuers tend to highlight the benefit of AI for customers, though overall many sources require AI to be shared and to benefit everyone "humanity", both of the above, "society", "as many people as possible", "all sentient creatures", the "planet" and the environment.	The global landscape of AI ethics guidelines, Nature 2019
	AI must serve people and planet	Codes of ethics for the development, application and use of AI are needed so that throughout their entire operational process, AI	UNI Global Union 2017

	Principle	*Key message*	*Sources*
		systems remain compatible and increase the principles of human dignity, integrity, freedom, privacy, and cultural and gender diversity, as well as fundamental human rights.	
	Well-being	AIs should be used to support prosperity, health, democratic civic processes, personal freedom, goodwill, environmental sustainability, and the protection of children, people with disabilities, displaced people and other vulnerable populations.	WEF Principles Development Tool 2020
4	**Autonomy**	AI should respect human autonomy by requiring human control at all times. *NB: need to take into account situations where human control could be detrimental.*	COMEST 2019
	Value of autonomy	The recognition of human dignity implies that the value of autonomy does not solely concern the respect of individual autonomy, which can go as far as to refuse to be under the charge of a robot. The value of autonomy also expresses the recognition of the interdependency of relationship between humans, between humans and animals, and between humans and the environment. To what extent social robots will enrich our relationships, or reduce and standardise them? This needs to be scientifically evaluated in medical and educational practices where robots can be used, especially when vulnerable groups such as children and elderly persons are concerned. The extensive use of robots can accentuate in certain societies the rupture of social bonds. Interdependency implies that robots are part of our technical creations (part of the technocosm that we construct) and they also have environmental impacts (e-waste, energy consumption and CO2 emissions, ecological footprint) that must be considered and evaluated in the balance of benefit and risk.	COMEST 2017
	Principle “under user control”	Precluding a prescriptive approach and ensuring that users are informed actors and in control of their choices.	CoE Ethical Charter
	Respect for autonomy principle	AI systems must be developed and used while respecting people’s autonomy, and with the goal of increasing people’s control over their lives and their surroundings.	Montreal Declaration

	Principle	***Key message***	***Sources***
	Adopt a human-in-command approach	The development of AI must be responsible, safe and useful, where machines maintain the legal status of tools, and legal persons retain control over, and responsibility for, these machines at all times.	UNI Global Union 2017
	Human control of technology	The principles under this theme require that important decisions remain subject to human review. Human Control of Technology principles are present in 69% of documents in the dataset.	Berkman Klein Center 2020
5	**Explainability**	AI should be explainable, able to provide insight into its functioning.	COMEST 2019
	Transparency and explainability	AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: i. to foster a general understanding of AI systems; ii. to make stakeholders aware of their interactions with AI systems, including in the workplace; iii. to enable those affected by an AI system to understand the outcome; and, iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.	G20 AI Principles 2019 = OECD AI Principles 2019
	Transparency and explainability	Principles under this theme articulate requirements that AI systems be designed and implemented to allow for oversight, including through translation of their operations into intelligible outputs and the provision of information about where, when, and how they are being used. Transparency and Explainability principles are present in 94% of documents in the dataset.	Berkman Klein Center
	Comprehension	The reasons for any AI decisions and actions should be understood well enough for humans to control AIs for consistency with ethical principles, and to make human accountability possible.	WEF Principles Development Tool 2020

	Principle	***Key message***	***Sources***
6	**Transparency**	The data used to train AI systems should be transparent.	COMEST 2019
	Openness	Open, in the way that Internet protocols are developed, applications are designed, and services are made available to their users.	UNESCO General Conference 2015 decision on the Internet Universality
	Principle of transparency, impartiality and fairness	Making data processing methods accessible and understandable, authorising external audits	CoE Ethical Charter
	Transparency	Ensure autonomous and intelligent systems operate in a transparent manner.	IEEE's Ethically Aligned Design 2019
	AI systems must be transparent	Workers should have the right to demand transparency in the decisions and outcomes of AI systems, as well as their underlying algorithms. They must also be consulted on AI systems implementation, development and deployment.	UNI Global Union 2017
	Transparency	Featured in 73 of our 84 sources, transparency is the most prevalent principle in the current literature. Thematic analysis reveals significant variation in relation to the interpretation, justification, domain of application and mode of achievement. References to transparency comprise efforts to *increase explainability, interpretability or other acts of communication and disclosure*. Principal domains of application include data use, human–AI interaction, automated decisions and the purpose of data use or application of AI systems. Primarily, transparency is presented as a way to minimize harm and improve AI, though some sources underline its benefit for legal reasons or to foster trust. A few sources also link transparency to dialogue, participation and the principles of democracy.	The global landscape of AI ethics guidelines, Nature 2019
7	**Awareness and literacy**	Algorithm awareness and a basic understanding of the workings of AI are needed to empower citizens.	COMEST 2019
	AIS technology misuse and awareness of it	Minimise the risks of misuse of AIS technology	IEEE's Ethically Aligned Design 2019
8	**Responsibility**	Developers and companies should take into consideration ethics when developing autonomous intelligent system.	COMEST 2019

	Principle	*Key message*	*Sources*
	Principle of responsibility	Deterministic robots, and even sophisticated cognitive robots, cannot take any ethical responsibility, which lies with the designer, manufacturer, seller, user, and the State. Therefore, human beings should always be in the loop and find ways to control robots by different means (e.g. traceability, off switch, etc.) in order to maintain human moral and legal responsibility.	COMEST 2017
	Professional responsibility	These principles recognize the vital role that individuals involved in the development and deployment of AI systems play in the systems' impacts, and call on their professionalism and integrity in ensuring that the appropriate stakeholders are consulted and long-term effects are planned for. Professional Responsibility principles are present in 78% of documents in the dataset.	Berkman Klein Center 2020
9	**Accountability** *(often cited in combination with responsibility or used interchangeably)*	Arrangements should be developed that will make possible to attribute accountability for AI-driven decisions and the behaviour of AI systems.	COMEST 2019
	Accountability	AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.	G20 AI Principles 2019 = OECD AI Principles 2019
	Accountability	Ensure that designers and operators of AISs are responsible and accountable.	IEEE's Ethically Aligned Design 2019
	Ban the attribution of responsibility to robots	Robots should be designed and operated as far as is practicable to comply with existing laws, and fundamental rights and freedoms, including privacy.	UNI Global Union 2017
	Accountability	This theme includes principles concerning the importance of mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided. Accountability principles are present in 97% of documents in the dataset.	Berkman Klein Center 2020

	Principle	*Key message*	*Sources*
	Accountability	The responsibility for an AI's decisions and actions should never be delegated to the AI. People should take responsibility for following ethical principles when working with AI and be held accountable when AIs break ethical principles and voluntary obligations.	WEF Principles Development Tool 2020
10	**Democracy**	AI should be developed, implemented and used in line with democratic principles.	COMEST 2019
	Democratic participation principle	AI systems must meet intelligibility, justifiability, and accessibility criteria, and must be subjected to democratic scrutiny, debate, and control.	Montreal Declaration 2018
11	**Good governance**	Governments should provide regular reports about their use of AI in policing, intelligence and security.	COMEST 2019
	Multi-stakeholder governance	Building on the successful partnerships that have evolved since WSIS between governments, the private sector, the technical and professional community, and civil society to foster the Internet's growth and use for peace, prosperity, social equality and sustainable development.	UNESCO General Conference 2015 decision on the Internet Universality
	Establish global governance mechanism	Establish multi-stakeholder Decent Work and Ethical AI governance bodies on global and regional levels. The bodies should include AI designers, manufacturers, owners, developers, researchers, employers, lawyers, civil society organisations and trade unions.	UNI Global Union 2017
12	**Sustainability**[29]	i. For all AI applications, the potential benefits need to be balanced against the environmental impact of the entire AI and IT production cycle. ii. AI should be developed in a sustainable manner taking into account the entire AI and IT production cycle. iii. AI can be used for environmental monitoring and risk management, and to prevent and mitigate environmental crises.	COMEST 2019

[29] Suggested by this background document as a foundational value in a different formulation.

	Principle	***Key message***	***Sources***
	Sustainable development principle	The development and use of AI systems must be carried out so as to ensure a strong environmental sustainability of the planet.	Montreal Declaration 2018
	Sustainability	To the extent that is referenced, sustainability calls for development and deployment of AI to consider protecting the environment, improving the planet's ecosystem and biodiversity, contributing to fairer and more equal societies and promoting peace. Ideally, AI creates sustainable systems that process data sustainably and whose insights remain valid over time.	The global landscape of AI ethics guidelines, Nature 2019
Other relevant principles or sub-principles			
13	**Safety and security**	These principles express requirements that AI systems be safe, performing as intended, and also secure, resistant to being compromised by unauthorized parties. Safety and Security principles are present in 81% of documents in the dataset.	Berkman Klein Center 2020
	Do no harm principle	Board members emphasized the importance of incorporating the "do no harm" principle at the outset when designing solutions.	Ethics of AI Context from CEB and HLCP 2020
	'Do not harm' principle	'Do not harm' principle is a red line for robots. As many technologies, a robot has the potentiality for 'dual-use'. Robots are usually designed for good and useful purposes (to diminish harmfulness of work for example), to help human beings, not to harm or kill them. In this regard, Isaac Asimov's formulation of this principle (three laws) is still accurate (see paragraph 18. If we are morally serious about this ethical principle, then we have to ask ourselves whether armed drones and autonomous weapons should be banned.	COMEST 2017
	Prudence principle	The development and use of AI systems must not contribute to lessening the responsibility of human beings when decisions must be made.	Montreal Declaration 2018
	Principle of quality and security	With regard to the processing of judicial decisions and data, using certified sources and intangible data with models conceived in a multi-disciplinary manner, in a secure technological environment	CoE Ethical Charter 2018
	Robustness, security and safety	i. AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions,	G20 AI Principles 2019 = OECD AI Principles 2019

	Principle	*Key message*	*Sources*
		they function appropriately and do not pose unreasonable safety risk. ii. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.	
	Non-maleficence	References to non-maleficence occur significantly more often than references to beneficence and encompass general calls for safety and security or state that AI should never cause foreseeable or unintentional harm. More granular considerations entail the avoidance of specific risks or potential harms—for example, intentional misuse via cyberwarfare and malicious hacking—and suggest risk-management strategies. Harm is primarily interpreted as discrimination, violation of privacy or bodily harm. Less frequent characterizations include loss of trust or skills; "radical individualism"; the risk that technological progress might outpace regulatory measures; and negative impacts on long-term social well-being, infrastructure, or psychological, emotional or economic aspects.	The global landscape of AI ethics guidelines, Nature 2019
	Safety	Deliberate or inadvertent harm caused by AIs should be prohibited, prevented and stopped.	WEF Principles Development Tool 2020
14	**Gender**	Gender bias should be avoided in the development of algorithms, in the datasets used for their training, and in their use in decision-making.	COMEST 2019
		All artificial intelligence-related capacity-building programming by United Nations entities should be gender transformative. Gender	CEB 2019

	Principle	*Key message*	*Sources*
		and age transformative approaches need to be embedded in all artificial intelligence-related capacity-building programming and decision-making processes. The particular effects of artificial intelligence on women and girls, and on the increasing digital gender and age divide, should also be taken into account.	
		Specifically preventing the development or intensification of any discrimination between individuals or groups of individuals	CoE Ethical Charter 2018
		In the design and maintenance of AI and artificial systems, it is vital that the system is controlled for negative or harmful human-bias, and that any bias be it gender, race, sexual orientation or age is identified and is not propagated by the system.	UNI Global Union 2017
15	**Age** (young and elderly)	Young people have valid concerns relating to ethical issues of AI. As such, they should be included, in all their diversity, in all discussions on the ethical principles of AI and their concerns and considerations taken into account.	UNESCO Operational Strategy on Youth (2014-2021)
16	**Privacy**	Principles under this theme stand for the idea that AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it. Privacy principles are present in 97% of documents in the dataset.	Berkman Klein Center 2020
	Value of privacy	Various protection schemes and regulations have been implemented in many countries to limit access to personal data in order to protect the privacy of individuals. However, the advent of Big Data changes the way data are collected and how they are processed (use of algorithm in profiling). The scale is much wider and the uses are expanding (e.g. commercial, state security and surveillance, research, etc.), and so are the forms of intrusion. Robots are devices that can collect data through sensors and that can use Big Data through deep learning. Therefore, collection and use of data need to be scrutinized in the design of robots, using an approach that balances the aim of the robot and the protection of privacy. Some data may be more sensitive than others; therefore a mix of approaches such as legislation,	COMEST 2017

	Principle	*Key message*	*Sources*
		professional regulations, governance, public surveillance, etc. is necessary in order to maintain public trust in and good use of robots.	
	Protection of privacy and intimacy principle	Privacy and intimacy must be protected from AI systems intrusion and data acquisition and archiving systems.	Montreal Declaration 2018
	Privacy	Ethical AI sees privacy both as a value to uphold and as a right to be protected. While often undefined, privacy is frequently presented in relation to *data protection and data security*. A few sources link privacy to freedom or trust.	The global landscape of AI ethics guidelines, Nature 2019
	Privacy	AIs and people with AI responsibilities should protect personal and client data. Those who gather or share data with AIs or from AIs should seek and respect the preferences of those whom the data is about, including their preference to control the data.	WEF Principles Development Tool 2020
17	**Solidarity principle**	The development of AI systems must be compatible with maintaining the bonds of solidarity among people and generations.	Montreal Declaration 2018
	Solidarity	Solidarity is mostly referenced in relation to the implications of AI for the labour market. Sources call for a strong social safety net. They underline the need for redistributing the benefits of AI in order not to threaten social cohesion and respecting potentially vulnerable persons and groups. Lastly, there is a warning of data collection and practices focused on individuals that may undermine solidarity in favour of “radical individualism”.	The global landscape of AI ethics guidelines, Nature 2019
	Principle of solidarity and social justice	Any ethically permissible application should not increase disadvantage, discrimination or division in society. This principle is one of the two guiding principles proposed by the Nuffield Council, alongside the principle that any intervention should be consistent with the welfare of the future person. The French and German councils also emphasise the ethical concepts of non-maleficence and beneficence. In addition, the Deutscher Ethikrat recommends consideration of the ethical concepts of human dignity, protection of life and integrity, freedom, naturalness and responsibility.	Joint statement on the ethics of heritable human genome editing 2020

	Principle	*Key message*	*Sources*
18	**Value of justice** (Equality)	The value of justice is related to inequality. The extensive use of industrial robots and service robots will generate higher unemployment for certain segments of the work force. This raises fears concerning rising inequality within society if there are no ways to compensate, to provide work to people, or to organize the workplace differently. Work is still a central element of social and personal identity and recognition. The value of justice is also related to non-discrimination. Roboticists should be sensitised to the reproduction of gender bias and sexual stereotype in robots. The issue of discrimination and stigmatisation through data mining collected by robots is not a trivial issue. Adequate measures need to be taken by States.	COMEST 2017
	Justice, fairness and equity	Justice is mainly expressed in terms of *fairness* and of *prevention, monitoring or mitigation of unwanted bias and discrimination*, the latter being significantly less referenced than the first two by the private sector. Whereas some sources focus on justice as *respect for diversity, inclusion* and *equality*, others call for a *possibility to appeal or challenge decisions* or the *right to redress and remedy*. Sources also emphasize the importance of *fair access to AI*, *data* and the *benefits of AI*. Issuers from the public sector place particular emphasis on AI's *impact on the labour market*, and the ne*ed to address democratic or societal issues*. Sources focusing on the risk of biases within datasets underline the importance of acquiring and processing accurate, complete and diverse data especially training data.	The global landscape of AI ethics guidelines, Nature 2019
	Equity principle	The development and use of AI systems must contribute to the creation of a just and equal society.	Montreal Declaration 2018
	Equality	AIs should make only fair decisions consistent with human rights.	WEF Principles Development Tool 2020
19	**Holistic approach**	Artificial intelligence should be addressed in an ambitious and holistic manner, promoting the use of artificial intelligence as a tool in the implementation of the Goals, while also addressing emerging ethical and human rights, decent work, technical and socioeconomic challenges.	CEB 2019

	Principle	*Key message*	*Sources*
20	**Trust**	References to trust include calls for trustworthy AI research and technology, trustworthy AI developers and organizations, trustworthy “design principles”, or underline the importance of customers’ trust. Calls for trust are proposed because a culture of trust among scientists and engineers is believed to support the achievement of other organizational goals, or because overall trust in the recommendations, judgments and uses of AI is indispensable for AI to “fulfil its world changing potential”. This last point is contradicted by one guideline explicitly warning against excessive trust in AI.	The global landscape of AI ethics guidelines, Nature 2019
21	**Freedom**	Whereas some sources specifically refer to the freedom of expression or informational self-determination and “privacy-protecting user controls”, others generally promote freedom, empowerment or autonomy. Some documents refer to autonomy as a positive freedom, specifically the freedom to flourish, to self-determination through democratic means, the right to establish and develop relationships with other human beings, the freedom to withdraw consent, or the freedom to use a preferred platform or technology. Other documents focus on negative freedom—for example, freedom from technological experimentation99, manipulation or surveillance. Freedom and autonomy are believed to be promoted through transparency and predictable AI55, by not “reducing options for and knowledge of citizens”, by actively increasing people’s knowledge about AI, giving notice and consent or, conversely, by actively refraining from collecting and spreading data in absence of informed consent.	The global landscape of AI ethics guidelines, Nature 2019
22	**Dignity**	While dignity remains undefined in existing guidelines, save one specification that it is a prerogative of humans but not robots, there is frequent reference to what it entails: dignity is intertwined with human rights or otherwise means avoiding harm, forced acceptance, automated classification and unknown human–AI interaction. It is argued that AI should not diminish or destroy, but respect, preserve or even increase human dignity. Dignity is believed to be preserved if it is respected by AI developers in the	The global landscape of AI ethics guidelines, Nature 2019

	Principle	*Key message*	*Sources*
		first place and promoted through new legislation, through governance initiatives, or through government issued technical and methodological guidelines.	
23	**Remediation**	Those with AI responsibilities should seek to be educated by people affected by their AIs. Workers, customers and others affected should have fair means to seek assistance or redress should AI endanger their livelihood, reputation or physical well-being.	WEF Principles Development Tool 2020
24	**Professionalism**	AI researchers, scientists and technicians should follow high scientific and professional standards.	WEF Principles Development Tool 2020

ANNEX 4: SUMMARY TABLES OF A SELECTION OF POSSIBLE POLICY ACTIONS TO GUIDE REFLECTION

	Policy action	*Key message*	*Sources*
Potentially relevant overarching policy actions for all stakeholders			
1	**Promoting research**	• Participating in interdisciplinary research on how AI ethics intersects with human rights, openness, accessibility and multistakeholder governance, and promoting Open Access publishing of the research results. • Using UNESCO's Internet Universality indicators to measure human Rights, Openness, Accessibility and Multi-stakeholder participation and to thereby map and improve the ecosystem in which AI and its ethics are developed, applied and governed. • Assessing algorithmic discrimination in order to protect the right to equality of all, in particular of historically marginalized populations.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
2	**Putting human rights upfront**	Applying human rights norms that can inform more specific ethical guidelines for rights to expression, privacy, and participation in public life.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
3	**Promoting transparency**	As a basis for ethical process on AI, facilitating the development of norms and policies for improving openness and transparency in AI algorithms through elements of ex-ante information disclosure and ex-poste monitoring of algorithmic decision-making.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
4	**Educating about cost-benefit and inequalities**	• Raising awareness of ownership and access to big data, AI skills and technologies, and the issues of who benefits, as well as harms such as marginalization or manipulation of human agency. • Upholding open market competition to prevent monopolization of AI and advance the ethics of inclusion, while also requiring adequate safeguards against violation of ethical practices by market-driven factors. • Working to reduce digital divides, including gender divides, in regard to AI access, and establishing	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019

	Policy action	*Key message*	*Sources*
		independent monitoring mechanisms, as key for leaving no one behind.	
5	**Practising multistakeholder governance**	• Motivating for process ethics through encouraging and enabling active participation in AI governance from all stakeholder groups, including but not limited to Governments, the Private Sector, Technical Community, Civil Society, Academia, International organizations and Media. • Ensuring gender equality, linguistic and regional diversity as well as the inclusion of youth and marginalized groups in multi-stakeholder ethical dialogues on AI issues.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
6	**Mainstreaming AI ethics**	Integrating discussion of AI ethical issues into relevant events such as UN international days around press freedom, disability, and universal access to information, and drawing in networks linked to UNESCO as well Category 2 institutes, NGOs, UNESCO intergovernmental bodies, UNESCO National Commissions.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
Potentially relevant policy actions for governments			
7	**Investing in AI research and development**	i. Governments should consider long-term public investment, and encourage private investment, in research and development, including inter-disciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues. ii. Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.	G20 AI Principles 2019 = OECD AI Principles 2019
8	**Fostering a digital ecosystem for AI**	Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem includes	G20 AI Principles 2019 = OECD AI Principles 2019

	Policy action	*Key message*	*Sources*
		in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.	
9	**Shaping an enabling policy environment for AI**	a) Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate. b) Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.	G20 AI Principles 2019 = OECD AI Principles 2019
10	**Building human capacity and preparing for labor market transformation**	a) Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills. b) Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programs along the working life, support for those affected by displacement, and access to new opportunities in the labor market. c) Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits from AI are broadly and fairly shared.	G20 AI Principles 2019 = OECD AI Principles 2019
11	**International co-operation for trustworthy AI**	a) Governments, including developing countries and with stakeholders, should actively cooperate to advance these	G20 AI Principles 2019 = OECD AI Principles 2019

	Policy action	*Key message*	*Sources*
		principles and to progress on responsible stewardship of trustworthy AI. b) Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate. They should encourage international, cross sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI. c) Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI. d) Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.	
12	**Establishing impact assessment**	Introduce ethical impact assessment on national level to predict consequences, avoid harmful consequences, facilitate participation and address societal challenges in line with the principles.	-
13	**Ensuring education**	• Planning AI in education policies • AI for education management and delivery • AI to empower teaching and teachers • AI for learning and learning assessment • Development of values and skills for life and work in the AI era • AI for offering lifelong learning opportunities for all • Promoting equitable and inclusive use of AI in education • Gender-equitable AI and AI for gender equality • Ensuring ethical, transparent and auditable use of education data and algorithms • Monitoring, evaluation and research • Financing, partnership and international cooperation	Beijing Consensus 2019

	Policy action	***Key message***	***Sources***
		• Leverage AI to promote quality education with a special focus on STEM, scientific research and innovation, as well as to continue to strengthen education for citizenship based on values, rights and duties. • Promote the disciplines of sciences, as well as media and AI literacy, which contribute to the development of critical thinking and the acquisition of the skills required to understand and use AI responsibly.	
	Increase artificial intelligence-related human capacity by supporting high quality and inclusive education, learning and training policies and programmes as well as reskilling and retraining of workers, including women and girls	Human capacity-building, including education and reskilling, is a critical element of efforts to ensure employability of workers and ensuring that no one is left behind. Taking into consideration the requirements of the bottom billion has to ensure that the most marginalized and those that are most vulnerable to the risks and barriers presented by artificial intelligence, including women and the elderly, are empowered. In this regard, a key strategy is to enrich and diversify the knowledge base of the labour force and promote shared mindsets that enable enterprises and organizations to rapidly adopt and diffuse new artificial intelligence technologies, and thus shape the future of work and make progress towards the Goals. This strategy needs to address learning in schools and workplaces, social networks such as families and communities, occupational and organizational networks, while also using digital platform and artificial intelligence tools. These aspects are further elaborated in the strategies on the future of learning and education and the future of work.	CEB 2019
	AI Education, training and re-skilling	• Support universities and technical training institutes to educate and train more students in AI and associated fields, thereby strengthening AI talent availability. • Encourage and support the acquisition of coding skills and computer science literacy for citizens through proactive policies for education, technical and vocational training, including for lifelong learning.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019

	Policy action	*Key message*	*Sources*
		• Strengthen gender diversity in AI research both in academia and the private sector. • Collaborate with universities and research centres, including through student training, doctoral research grants, sharing data and computing resources for research and development. • Create and strengthen mechanisms for research collaboration, mobility of researchers and mentorship opportunities for students between universities across the world, with special focus on North-South and South-South exchanges, and on gender parity. • Strengthen access to AI knowledge by offering high quality open educational resources in multiple languages and formats accessible by persons with disabilities. Update university curricula dynamically with state of the art research developments and methodologies, including through regular education and skills needs assessment in partnership with the private sector and other stakeholders.	
14	**Ensuring systemic changes and processes**	Governmental action, oversight, more interdisciplinary or otherwise diverse workforce, better inclusion of civil society or other relevant stakeholders, increased attention to the distribution of benefits.	The global landscape of AI ethics guidelines, Nature 2019
15	**Providing for technical measures and governance strategies.**	From interventions at the level of AI research, design, technology development and/or deployment to lateral and continuous approaches. Technical solutions: in-built data quality evaluations or security and privacy by design, industry standards. Governance strategies: active cooperation across disciplines and stakeholders, compliance, oversight processes and practices (tests, monitoring, audits and assessments by internal units, customers, users, independent third parties or governmental entities, often geared towards standards for AI implementation and outcome assessment). Many imply that damages may be unavoidable, in which case risks should be assessed, reduced and mitigated, and the	The global landscape of AI ethics guidelines, Nature 2019

	Policy action	*Key message*	*Sources*
		attribution of liability should be clearly defined. Several sources mention potential “multiple” or “dual use”, take explicit position against military application or simply guard against the dynamics of an “arms race”.	
16	**Ensuring responsibility and accountability**	Despite widespread references to “responsible AI” , responsibility and accountability are rarely defined. Nonetheless, specific recommendations include acting with “integrity” and clarifying the attribution of responsibility and legal liability, if possible upfront, in contracts or, alternatively, by centring on remedy. In contrast, other sources suggest focusing on the underlying reasons and processes that may. lead to potential harm. Yet others underline the responsibility of whistleblowing in case of potential harm, and aim at promoting diversity or *introducing ethics into science, technology, engineering and mathematics education. Very different actors are named as being responsible and accountable for AI’s actions and decisions*: AI developers, designers, “institutions” or “industry”. Further divergence emerged on whether AI should be held accountable in a human-like manner87 or whether humans should always be the only actors who are ultimately responsible for technological artifacts.	The global landscape of AI ethics guidelines, Nature 2019
17	**Providing for technical solutions; more research and awareness; regulatory approaches**	Technical solutions: differential privacy, privacy by design, data minimization and access control. Regulatory approaches: legal compliance, certificates, creation or adaptation of laws and regulations to accommodate the specificities of AI.	The global landscape of AI ethics guidelines, Nature 2019
18	**Ensuring beneficence**	Strategies for the promotion of good include aligning AI with human values, advancing “scientific understanding of the world”, minimizing power concentration or, conversely, using power “for the benefit of human rights”, working more closely with “affected” people, minimizing conflicts of interests, proving beneficence through customer demand and feedback, and developing new metrics and measurements for human well-being.	The global landscape of AI ethics guidelines, Nature 2019
19	**Ensuring trust**	Suggestions for building or sustaining trust include education, reliability, accountability, processes to monitor and evaluate the integrity of AI systems over time, and tools and techniques	The global landscape of AI ethics guidelines, Nature 2019

	Policy action	*Key message*	*Sources*
		ensuring compliance with norms and standards. Whereas some guidelines require AI to be transparent understandable or explainable in order to build trust, another one explicitly suggests that, instead of demanding understandability, it should be ensured that AI fulfils public expectations. Other reported facilitators of trust include “a Certificate of Fairness”, multi-stakeholder dialogue, awareness about the value of using personal data, and avoiding harm.	
20	**Ensuring sustainability**	AI should be designed, deployed and managed with care to increase its energy efficiency and minimize its ecological footprint. To make future developments sustainable, corporations are asked to create policies ensuring accountability in the domain of potential job losses and to use challenges as an opportunity for innovation.	The global landscape of AI ethics guidelines, Nature 2019
21	**Ensuring gender-sensitive approach**	• Adopt sustained, varied and life-wide approaches; • Establish incentives, targets and quotas; • Embed ICT in formal education • Support engaging experiences; • Emphasize meaningful use and tangible benefits; • Encourage collaborative and peer learning; • Create safe spaces and meet women where they are; • Examine exclusionary practices and language; • Recruit and train gender-sensitive teachers; • Promote role models and mentors; • Bring parents on board; • Leverage community connections and recruit allies; • Support technology autonomy and women’s digital rights; • Use universal service and access funds; • Collect and use data, and set actionable indicators and targets.	AI systems must be equipped with an ethical black box
22	**AI systems must be equipped with an ethical black box**	The ethical black box should not only contain relevant data to ensure system transparency and accountability, but also clear	UNI Global Union 2017

	Policy action	*Key message*	*Sources*
		data and information on the ethical considerations built into the system.	
23	**Ensuring human centred AI**	• Develop adequate policy and regulatory frameworks to address the human rights challenges posed by the development and application of AI, providing mechanisms for preventing human rights violations, as well as for transparency, accountability and remedy processes. • Evaluate if existing regulation against discrimination enables an individual to seek remedy for algorithmic discrimination. • Develop norms and policies for improving openness, transparency and accountability in automated decisions taken by AI systems through methods such as ex-ante information disclosure and ex-post monitoring of automated decision-making. • Ensure policies that provide for affordable broadband access and avoid interferences with connectivity such as Internet shut downs, throttling or arbitrary filtering and blocking. • Motivate more active participation to discuss AI policies at national and supra-national levels from all stakeholder groups, including but not limited to: i) government, ii) private sector, iii) technical community, iv) civil society, v) academia, vi) international organizations, and vii) media. • Organize multi-stakeholder fora and events for AI issues and policies and integrate multi-stakeholder participation in monitoring and correcting where there are unexpected outcomes that are problematic.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
24	**Ensuring democracy**	• Take effective measures to ensure that algorithms are not exploited to impede the right to free elections. • Support the UN Plan of Action on the Safety of Journalists and the Issue of Impunity and address the AI-assisted attacks on journalists and media workers.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019

	Policy action	*Key message*	*Sources*
25	**Use of AI in public service delivery**	• Ensure that the public sector's use of AI in decision-making is transparent and consistent with human rights obligations. • Establish guidelines and policies for openness, transparency and ac-countability in the use and deployment of automated decision-ma-king systems, including for use by the government.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
26	**Ensuring open markets**	• Facilitate open market competition to prevent monopolization of AI and follow the United Nations 'Guiding Principles on Business and Human Rights' for human rights based best practices for businesses.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
27	**Data policies**	• Ensure adequate safeguards are put in place with respect to open data in order to protect against the infringement of the right to privacy. • Create open repositories for publicly funded or owned data and re-search including the creation of platforms for open government data. • Develop standards for interoperability between data sets while strengthening data commons and the availability of data for machine learning.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
28	**Fighting digital divide**	• Work to reduce digital divides, including gender divides, in AI access, and establish mechanisms for continuous monitoring of the differences in access. • Ensure that individuals, groups and countries that are least likely to have access to AI are active participants in multi-stakeholder dialogues on the digital divide by emphasizing the importance of gender equality, linguistic and regional diversity as well as the inclusion of youth and marginalized groups.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
29	**Ensuring infrastructure**	• Strengthen the infrastructure and support needed for AI-related re-search and development at universities and research centres.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019

	Policy action	*Key message*	*Sources*
		• Strengthen access to AI-specific computational hardware, including through funding support and providing need-based access to centralized computing resources.	

	Stakeholder group	*Key message*	*Sources*
Policy actions for other stakeholders cross-cutting all principles			
1	**Private sector and technical community**	• for ethical process, conduct human rights risk and impact assessments of AI applications to ensure that these do not interfere with human rights. • develop self-regulation norms for ethical practices in deployment of AI to avoid risky or anti-competitive behaviour in pursuit of market advantage. • Increase transparency reporting, in order to enhance ethical decision-making and participation • provide greater access to affordable connectivity, hardware and software needed for running AI programs. • more actively involved in national and international level policymaking concerned with AI, and engage other actors in their internal governance issues such as defining terms of service and operating procedures.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
2	**Academia**	• engage in rights-oriented research on the social, economic and political effects of AI content personalization, including the consequences of manipulation of human agency • support the development of open data standards (while safeguarding privacy) and ensure interoperability between different data sets while strengthening data commons and the availability of data for machine learning. • improve access to AI algorithms for learning through the creation of research repositories and by offering online education for AI.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019

	Stakeholder group	*Key message*	*Sources*
		• conduct research to support the institutionalization and sustainability of multi-stakeholder governance experiences.	
3	**Civil society**	• advocate that AI development and use must respect the ethic of human rights • act as a watchdog against hidden operations of AI and demand greater transparency in regard to funding and use of the technologies. • support the development of AI content and resources in formats and languages that render the information about AI more widely available.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
4	**Media actors**	• investigate and report on abuses and biases of AI as well as the benefits, and harness AI to strengthen journalism and media development. investigate and report on abuses and biases of AI as well as the benefits, and harness AI to strengthen journalism and media development. • participate actively in, and provide coverage of, governance processes for AI.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019
5	**UNESCO, UN agencies and other international organizations**	• convene ongoing dialogues about AI to ensure that ethics and norms of human rights are kept aloft and strengthened, and not be ignored or eroded. • continue to foster the Open Data movement by helping establish Open Data Standards and Open Data Repositories for AI through its network of partners and Category 2 Centres. • support Member States to enhance AI research capacity in general, and in the areas of communication-information in particular, through stimulating relevant trainings, education policy development, academic exchanges and through CI's intergovernmental programmes. • offer a forum for international and multistakeholder cooperation.	Steering AI and Advanced ICTs for Knowledge Societies, UNESCO 2019

ANNEX 5: SOURCES TO BE CONSIDERED

1. UNESCO:
    a. Preliminary Study on the Ethics of Artificial Intelligence (2019)
    b. Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective (2019)
    c. Final Report on the International Conference on Artificial Intelligence and Education. Planning Education in the AI Era: Lead the Leap (2019)
    d. Beijing Consensus on Artificial Intelligence and Education (2019)
    e. Artificial Intelligence in Education. Compendium of Promising Initiatives (2019)
    f. Artificial Intelligence for Sustainable Development: Synthesis Report (2019)
    g. I'd blush if I could: closing gender divides in digital skills through education (2019)
    h. Two-Eyed AI: A Reflection on Artificial Intelligence (2019)
    i. Bangkok Statement on the Ethics of Science and Technology and Sustainable Development (2019)
    j. Outcome Statement of the Forum on Artificial Intelligence in Africa, Benguérir (2018)
    k. Human Decisions: Thoughts on AI (2018)
    l. Report of COMEST on Robotics Ethics (2017)
    m. Principles for governing the Internet: a comparative analysis (2015)

2. The United Nations System:
    a. The Age of Digital Interdependence, Report of the UN Secretary-General's High-level Panel on Digital Cooperation (2019)
    b. Executive Summary of The Age of Digital Interdependence, Report of the UN Secretary-General's High-level Panel on Digital Cooperation (2019)
    c. United Nations Activities on Artificial Intelligence, ITU (2019)
    d. The right to privacy in the digital age. Resolution adopted by the Human Rights Council (2019)
    e. A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence, CEB/2019/1/Add.3 (2019)
    f. Policy Inputs for the Young UN Policy Lab (2018)
    g. AI for Good Global Summit Report, ITU (2017)

3. Other international organizations:
    a. Council of Europe:
        i. Declaration Decl(13/02/2019)1 on the manipulative capabilities of algorithmic processes (2019)
        ii. Recommendation on Artificial Intelligence and Human Rights "Unboxing artificial intelligence: 10 steps to protect human rights" (2019)
        iii. European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment, Council of Europe (2018)
        iv. Addressing the impacts of Algorithms on Human Rights: Draft Recommendation of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (2018)
        v. Recommendation n°2102(2017) about Technological convergence, artificial intelligence and human rights (2017)
    b. EU:
        i. White Paper on Artificial Intelligence – A European approach to excellence and trust, European Commission (2020)
        ii. Trustworthy AI Assessment List, High-Level Expert Group on AI (2020)
        iii. EU guidelines on ethics in artificial intelligence: Context and implementation, European Parliamentary Research Service (2019).

iv. Statement on Artificial Intelligence, Robotics and "Autonomous" Systems (incl. Ethical principles and democratic prerequisites), European Group on Ethics in Science and New Technologies, EGE (2018)
v. Ethics Guidelines for Trustworthy AI: Working Document for stakeholders' consultation, European Commission's High-Level Expert Group on AI (2018)
vi. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe, COM (2018) 237 final.

c. OECD:
   i. OECD AI Policy Observatory (2020)
   ii. Artificial Intelligence in Society (2019)
   iii. Recommendation of the Council on Artificial Intelligence (2019)
   iv. Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO), (2019)

4. Member States sources:
   a. Australia
      i. Australia's Ethics Framework, Department of Industry, Science, Energy and Resources (2019)
      ii. Australia's 2025 Digital Transformation Strategy (Vision 2025) (2018)
   b. Belgium: AI 4 Belgium (2019)
   c. Brazil: Brazilian Digital Transformation Strategy (2018)
   d. Canada:
      i. A set of guiding principles to ensure effective and ethical AI (2019)
      ii. Canada's Directive on Automated Decision-Making (2019)
      iii. Pan-Canadian AI Strategy, CIFAR (2017)
   e. China:
      i. Next Generation Artificial Intelligence Development Plan, State Council (2017)
   f. Czech Republic: National Artificial Intelligence Strategy of the Czech Republic (2019)
   g. Denmark: Danish National Strategy for Artificial Intelligence (2019)
   h. Estonia: National Artificial Intelligence Strategy 2019-2021 (2019)
   i. Finland:
      i. Leading the way into the age of artificial intelligence (2019)
      ii. Work in the age of artificial intelligence - four perspectives on economy, employment, skills and ethics (2018)
      iii. Finland's age of artificial intelligence (2017)
   j. France: Strategy for a Meaningful Artificial Intelligence (2018)
   k. Germany:
      i. Opinion of the Data Ethics Commission (2019)
      ii. Artificial Intelligence Strategy: AI Made in Germany (2018)
      iii. Automated and Connected Driving, BMVI Ethics Commission report (2017)
   l. India: National Strategy for Artificial Intelligence (2017)
   m. Italy: Artificial Intelligence at the service of the citizen, the Agency for Digital Italy (2018)
   n. Japan:
      i. Social Principles of Human-Centric AI (2019)
      ii. AI Strategy 2019 (2019)
      iii. Artificial Intelligence Technology Strategy, Strategic Council for AI Technology (2017)

iv. The Japanese Society for Artificial Intelligence Ethical Guidelines, JSAI (2017)
v. AI R&D Principles, Ministry of Internal Affairs and Communications (MIC) (2017)

o. Korea: Mid- to long-term master plan in preparation for the intelligent information society, Interdepartmental Exercise of the Korean Government (2017)
p. Lithuania: Lithuanian Artificial Intelligence Strategy: A vision of the future, Ministry of Economy (2019)
q. Luxembourg: Artificial Intelligence : a strategic vision for Luxembourg (2019)
r. New Zealand:
i. Artificial Intelligence: Shaping a Future New Zealand (2018)
ii. Government Use of Artificial Intelligence in New Zealand (2018)
s. Norway: The National Strategy for Artificial Intelligence (2020)
t. Portugal: AI Portugal 2030 (2019)
u. Russia: National Strategy for the development of Artificial Intelligence (2019)
v. Singapore:
i. Model Artificial Intelligence Governance Framework: Second Edition (2020)
ii. Compendium of Use Cases: Practical Illustrations of the Model AI Governance Framework (2020)
iii. Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations (2020)
iv. National Artificial Intelligence Strategy: Advancing our Smart National Journey (2019)
v. Summary of National Artificial Intelligence Strategy: Advancing our Smart National Journey (2019)
w. Spain: RDI Strategy in Artificial Intelligence, Ministry of Science, Innovation and Universities (2019)
x. Sweden:
i. National Approach for Artificial Intelligence (2018)
ii. Artificial Intelligence in Swedish Business and Society (2018)
y. United Arab Emirates: UAE Strategy for Artificial Intelligence (2017)
z. United Kingdom:
i. A guide to using artificial intelligence in the public sector (2020)
ii. Code of conduct for data-driven health and care technology, Department of Health and Social Care (2019)
aa. United States: National Artificial Intelligence Research and Development Strategic Plan (2019)

5. Declarations:
a. Montreal Declaration for a Responsible Development of AI, University of Montreal (2018)
b. Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, Amnesty International and Access Now (2018)
c. Declaration of the Future of Life Institute on the Asilomar AI Principles, Future of Life Institute (2017)

6. Other sources:
a. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI, Berkman Klein Center (2020)
b. Artificial Intelligence: Consumer Experiences in New Technology, Consumers International (2019)
c. Global Technology Governance: A Multistakeholder Approach, World Economic Forum (2019)

d. Ethically Aligned Design, First Edition: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, IEEE, 2019.
e. G20 Ministerial Statement on Trade and Digital Economy and G20 AI Principles, (2019)
f. Indigenous AI (2019)
g. The Future Computed: AI & Manufacturing, Microsoft (2019)
h. SAP's Guiding Principles for Artificial Intelligence, SAP (2018)
i. Principles for AI Ethics, SAMSUNG (2018)
j. Sony Group AI Ethics Guidelines, Sony (2018)
k. Harmonious Artificial Intelligence Principles, HAIP (2018)
l. Universal Guidelines for Artificial Intelligence, The Public Voice (2018)
m. OpenAI Charter, OpenAI (2018)
n. AI at Google: Our Principles, Google (2018)
o. Microsoft AI Principles, Microsoft (2018)
p. Principles for Trust and Transparency, IBM (2018)
q. G7 Common Vision for the Future of AI, (2018)
r. Developing AI for Business with Five Core Principles, Sage (2017)
s. Principles for Algorithmic Transparency and Accountability by ACM, USACM (2017)
t. Top 10 Principles for Ethical Artificial Intelligence, UNI Global Union (2017)
u. DeepMind Ethics & Society Principles, DeepMind (2017)
v. AI Policy Principles, ITI (2017)
w. Tenets of Partnership on AI, PAI (2016)

7. UNESCO recommendations as examples:
   a. Recommendation on Open Educational Resources (OER), 25 November 2019
   b. Recommendation on Science and Scientific Researchers, 13 November 2017
   c. Recommendation on Adult Learning and Education, 13 November 2015
   d. Recommendation concerning technical and vocational education and training (TVET), 13 November 2015
   e. Recommendation concerning the protection and promotion of museums and collections, their diversity and their role in society, 17 November 2015
   f. Recommendation concerning the preservation of, and access to, documentary heritage including in digital form, 17 November 2015
   g. Recommendation on the Historic Urban Landscape, including a glossary of definitions, 10 November 2011
   h. Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace, 15 October 2003
   i. Recommendation concerning the Status of Higher-Education Teaching Personnel, 11 November 1997
   j. Recommendation on the Recognition of Studies and Qualifications in Higher Education, 13 November 1993
   k. Recommendation on the Safeguarding of Traditional Culture and Folklore, 15 November 1989
   l. Revised Recommendation concerning the International Standardization of Statistics on the Production and Distribution of Books, Newspapers and Periodicals, 1 November 1985
   m. Recommendation concerning the Status of the Artist, 27 October 1980
   n. Recommendation for the Safeguarding and Preservation of Moving Images, 27 October 1980
   o. Recommendation concerning the International Standardization of Statistics on the Public Financing of Cultural Activities, 27 October 1980
   p. Revised Recommendation concerning International Competitions in Architecture and Town Planning, 27 November 1978

q. Recommendation concerning the International Standardization of Statistics on Science and Technology, 27 November 1978
r. Revised Recommendation concerning the International Standardization of Educational Statistics, 27 November 1978
s. Recommendation for the Protection of Movable Cultural Property, 28 November 1978
t. Recommendation concerning the International Exchange of Cultural Property, 26 November 1976
u. Recommendation concerning the Safeguarding and Contemporary Role of Historic Areas, 26 November 1976
v. Recommendation concerning the International Standardization of Statistics on Radio and Television, 22 November 1976
w. Recommendation on the Legal Protection of Translators and Translations and the Practical Means to improve the Status of Translators, 22 November 1976
x. Recommendation on Participation by the People at Large in Cultural Life and their Contribution to It, 26 November 1976
y. Recommendation concerning Education for International Understanding, Co-operation and Peace and Education relating to Human Rights and Fundamental Freedoms, 19 November 1974
z. Recommendation concerning the Protection, at National Level, of the Cultural and Natural Heritage, 16 November 1972
aa. Recommendation concerning the International Standardization of Library Statistics, 13 November 1970
bb. Recommendation concerning the Preservation of Cultural Property Endangered by Public or Private works, 19 November 1968
cc. Recommendation concerning the Status of Teachers, 5 October 1966
dd. Recommendation concerning the International Standardization of Statistics Relating to Book Production and Periodicals, 19 November 1964
ee. Recommendation on the Means of Prohibiting and Preventing the Illicit Export, Import and Transfer of Ownership of Cultural Property, 19 November 1964
ff. Recommendation concerning the Safeguarding of Beauty and Character of Landscapes and Sites, 11 December 1962
gg. Recommendation concerning the Most Effective Means of Rendering Museums Accessible to Everyone, 14 December 1960
hh. Recommendation against Discrimination in Education, 14 December 1960
ii. Recommendation on International Principles Applicable to Archaeological Excavations, 5 December 1956

ANNEX 6: PAST, ONGOING AND FUTURE INITIATIVES RELATED, EITHER DIRECTLY OR INDIRECTLY, TO THE ETHICAL, LEGAL AND SOCIAL IMPLICATIONS OF AI WITHIN THE UN SYSTEM

International Labour Organization (ILO)

- **ILO research program on Technologies and the Future of Work** addresses the impact of technology, including artificial intelligence (AI) on jobs, employment, decent work, productivity, inequality and sustainable development.
- **The Report of the Global Commission on the Future of Work "Work for a brighter future"** (January 2019) subscribes to a "human-in-command" approach to AI that ensures that the final decisions affecting work are taken by human beings, not algorithms. It also calls for the establishment of "an international governance system for digital labour platforms".

International Organization for Migration (IOM)

- IOM leads **an inter-agency group on Data Science, Artificial Intelligence and Ethics**, which established inter-agency peer review mechanisms for mathematical AI models and ethics.
- IOM co-leads, with OCHA and UNHCR, the **IASC RG1 Sub-Group on Data Responsibility**, tasked with developing "Joint System-Wide Operational Guidance on Data Responsibility in Humanitarian Action"
- IOM co-organized with the German Federal Foreign Office (FFO) an **interagency workshop on "Forecasting Human Mobility in Contexts of Crises"**, touching on diverse aspects of data science, including machine learning and artificial intelligence.
- IOM funded the "The Signal Code: Ethical Obligations for Humanitarian Information Activities", published by the Harvard Humanitarian Initiative in 2018.

International Telecommunications Union (ITU)

- **The AI for Good Global Summit** seeks to ensure trusted, safe and inclusive development of AI technologies and equitable access to their benefits.
- **The ITU/WHO AI for Health Focus Group** serves as a benchmarking framework for AI-enabled healthcare solutions so that they can be deployed responsibly and in the right context of use for all.
- **The ITU-UNESCO Broadband Commission for Sustainable Development's Working Group on AI and Global Health** facilitates advocacy efforts such as to generate knowledge on successes, challenges, and lessons learned from AI solutions in health.
- **The ITU Telecommunication Standardization Sector (ITU-T) Focus Group on AI for autonomous and assisted driving (FG-AI4AD)** supports standardization activities of AI evaluation in autonomous and assisted driving.

United Nations High Commissioner for Human Rights (OHCHR)

- **The OHCHR-UN Global Pulse Conference on a Human Rights-based approach to AI**
- The **High Commissioner for Human Rights' Report on the Right to Privacy in the Digital Age** (A/HRC/39/29) addressed the rise of data-driven technologies and made recommendations for rights-protective measures.
- **An expert seminar on the impact of AI on the enjoyment of the right to privacy** will be organized in 2020, with a thematic report on this topic to the Human Rights Council in September.
- OHCHR works closely with the Advisory Committee of the Human Rights Council on addressing human rights-issues related to digital technology, including AI.
- OHCHR also provides input into the work of several treaty bodies concerning AI (e.g. the **draft General Recommendation on racial profiling** of the **Committee on the Elimination of Racial Discrimination** and the **draft General Comment on the right of peaceful assembly** of the **Human Rights Committee**.

- **B-Tech project on the application of the UN Guiding Principles on Business and Human Rights** to the development and use of digital technologies including AI.

United Nations Department of Economic and Social Affairs (UN DESA)

- **The 2018 World Economic and Social Survey (WESS) on** Frontier Technologies for Sustainable Development analyzed (1) efficiency gains and equity and ethical concerns in relation to AI-based decision-making systems both in the public and private sector, and (2) production of targeted advertisements, manipulation of human emotion and spread of misinformation, including hatred.
- **A paper entitled** "Artificial Intelligence: Opportunities and Challenges for the Public Sector" addresses "*Ethical considerations for policy makers in the era of AI-centric approach*"
- **Global Working Group (GWG) on Big Data** has a task team on Privacy Preserving Techniques. This team produced a UN Handbook on Privacy-Preserving Computation Techniques.
- 2018 United Nations E-Government Survey **Chapter 8 entitled "Fast-evolving technologies in e-government: Government Platforms, Artificial Intelligence and People"** discusses transformative technologies, such as data analytics, artificial intelligence including cognitive analytics, robotics, bots, high-performance and quantum computing.
- **UN Technology Facilitation Mechanism (TFM)**
  - **Multi-Stakeholder Forum on Science, Technology and Innovation for the SDGs ("STI Forum")** is the premier UN space for discussions on STI for the SDGs, including such cross-SDG issues such as emerging technologies and their sustainable development impact.
  - **Interagency Task Team on STI for the SDGs (IATT),** through its work stream 10 ("Analytical work on emerging technologies and the SDGs"), have worked towards assessing the impacts of rapid technological change on the SDGs, including through UN expert group meetings to discuss the economic, societal and environmental impacts and ethical dimensions of artificial intelligence. Core principles and recommendations on responsible AI was suggested by experts in the contexts of the work under TFM and in particular the IATT's subgroup on new and emerging technologies.
- **The Commission for Social Development address** "Innovation and interconnectivity for social development" as an emerging issue, while **"Socially just transition towards sustainable development: the role of digital technologies on social development and well-being of all"** will be a priority theme for its 59th session in 2021
- Annual observance of the **International Day of Persons with Disabilities** on 3 December 2014 under the theme "*Sustainable Development: The promise of technology*".
- A roundtable on **"Technology, digitalization and information and communications technology for the empowerment and inclusion of persons with disabilities"** was organized at the 12th session of the Conference of States Parties to the Convention on the Rights of Persons with Disabilities in 2019
- **DESA-ITU side event on "Why it Matters: AI for Older Persons"** (18 April 2019) at the 10th working session of the General Assembly's open-ended working group for the purpose of strengthening the protection of the human rights of older persons.
- DESA will examine the potential for further research, including in collaboration with young researchers for 2021 and beyond to create a **youth research collaborative to investigate further the potential impacts of AI from a youth perspective**
  - DESA will produce a **policy paper focusing on the potential socioeconomic impacts of digital technologies on with a particular focus on youth**, given that they will experience much of the changes driven by AI
  - The **2021 World Youth Report has the theme "Safe and Inclusive Digital Spaces for Youth"**, which will explore issues around online data management,

disinformation, health and wellbeing, cybersecurity, and human rights etc. in the context of increasing youth engagement in digital spaces mediated by AI.

United Nations Conference on Trade and Development (UNCTAD)

- **Technology and Innovation Report (TIR)**
  - **TIR 2018 "Harnessing Frontier Technologies for Sustainable Development**" explored how harnessing frontier technologies could be transformative in achieving the Sustainable Development Goals (SDGs).
  - **Forthcoming TIR 2020** will outline the state-of-the-art debate and critically examine the possibility of frontier technologies (including AI) widening existing inequalities and creating new ones.
- UN Secretary-General's Report for the UN Commission on Science and Technology for Development (CSTD) on **"Harnessing rapid technological change for inclusive and sustainable development" (E/CN.16/2020/2)** and on **"Impact of rapid technological change on sustainable development" (E/CN.16/2019/2)** discussed the need for a consistent public policy response to the normative challenges posed by frontier technologies, notably Artificial Intelligence.
- Session entitled **"Structural transformation, Industry 4.0 and inequality: Science, technology and innovation policy challenges"**, at the Eleventh session of the Investment, Enterprise and Development Commission

United Nations Environment Programme (UNEP)

- **Science Policy Business Forum of UNEP** has been holding discussions and consultations related to the implications around data and AI.
- **Partnership with Google Earth Engine and the EU JRC** to deploy machine-learning algorithms to detect global surface freshwater from open source satellite images as a baseline data set for indicator SDG 6.6.1.
- **Partnership with Global AI** on using document scraping techniques to assess the compliance of Corporate Sustainability Reports to certain standards.

United Nations Educational, Scientific and Cultural Organization (UNESCO)

- As a follow-up of World Summit on the Information Society's (WSIS), UNESCO has taken responsibility for the implementation of the Action Lines on Access (C3), E-Learning (C7), Cultural diversity (C8), Media (C9), and Ethical dimension of the information society (C10).
- Member States of UNESCO has adopted the **framework of "Internet Universality"** and the associated **"R.O.A.M. principles" (Human Rights, Openness, Accessibility and Multi-stakeholder participation)** in 2015; a new publication entitled **"Steering AI and Advanced ICTs for Knowledge Societies: a ROAM perspective"** was launched at the Internet Governance Forum in 2019.
- UNESCO's Information For All Programme (IFAP) examined and approved the **Code of Ethics for the Information Society**
- UNESCO's World Commission on Ethics of Scientific Knowledge and Technology has prepared a **Preliminary Study on Ethics of Artificial Intelligence**, which triggered the decision of UNESCO Member States to elaborate a **Recommendation on the Ethics of Artificial Intelligence**
- UNESCO has also organized a series of events addressing the ethical, legal and social implications of AI. Some of major events include:
  - **Roundtables on "Artificial Intelligence: Reflection on its complexity and impact on our society"** (Paris, September 2018 & December 2019);
  - **Workshop on "Artificial Intelligence for Human Rights and SDGs: Fostering Multi-Stakeholder, Inclusive and Open Approaches"** (Paris, November 2018);
  - **Forum on artificial intelligence in Africa** (Ben Guérir, December 2018);
  - **Debate on Ethics of New Technologies and Artificial Intelligence "Tech Futures: Hope or Fear?"** (Paris, January 2019);

- **UNESCO Conference "Principles for AI: Towards a Humanistic Approach?"** (March 2019);
- **International Conference on Artificial Intelligence and Education** (Beijing, May 2019), with **Beijing Consensus on Artificial Intelligence and Education** as the outcome document;
- **Youth Voices and the Future of Artificial Intelligence: Towards a Human-Centered Approach** (Paris, November 2019).

United Nations Framework Convention on Climate Change (UNFCCC)

- General consideration of the use of AI in relation to climate action is being explored in the context of the UNFCCC's Resilience Frontiers initiative to further the exploration of frontier issues, as launched by the United Nations Chief Executives Board for Coordination.

United Nations Population Fund (UNFPA)

- Since 2018, **GRID3 (Geo-Referenced Infrastructure and Demographic Data for Development)** works with countries to generate, validate, and use geospatial data on population, settlements, infrastructure, and subnational boundaries in regions where an updated snapshot of populations and population distribution is needed and/or significant migration has occurred.
- **"Testing ECHO: Amplifying citizens' voices for the SDGs"** is an initiative led by UNFPA's Colombia Country Office, which is developing a tool powered by AI to promote citizens' participatory planning and awareness about the SDGs through real-time guided public discussion.

United Nations Industrial Development Organization (UNIDO)

- The ethical issues have been raised in the different discussions on the Fourth Industrial Revolution (4IR) , including at the Global Manufacturing and Industrialization Summit
- **International Conference on Ensuring Industrial Safety: the Role of Governments, Regulations and Standards (Vienna, May 2019)** discussed the implications of several 4IR technologies like AI on industrial safety and security (safe production, safe data transfer, safe human-robots/machine interactions)

United Nations Office on Drugs and Crime (UNODC) and United Nations Interregional Crime and Justice Research Institute (UNICRI)

- **UNODC's illicit crop monitoring programme** is piloting the use of AI (machine learning and deep learning) for detection of illicit crops on satellite images.
- **Fourth Workshop of the Fourteenth United Nations Congress on Crime Prevention and Criminal Justice (Kyoto, April 2020)** is expected to include the issue of the ethical considerations, as well as procedural and human rights safeguards, in the use of technology, including artificial intelligence and robotics, against crime as one of the sub-topics of discussion
- **Global Judicial Integrity Network** raises awareness about the implications of AI use in judiciaries through different events and advocacy methods.
- **The Centre for Artificial Intelligence and Robotics of UNICRI** has been working on AI since 2015, exploring the ethical, legal and social implications of advances in AI as they pertain to its mandate.
  - **UNICRI-INTERPOL annual Global Meeting on AI for law enforcement** since 2018
  - **Panel discussions on AI and Law Enforcement** at Tallinn Digital Summit in 2019 and at the 14th United Nations Congress on Crime Prevention and Criminal Justice
  - UNICRI and INTERPOL released a **Report on AI for Law Enforcement** in April 2019, which includes, inter alia, analysis of the ethical, legal and social implications and,
  - UNICRI and INTERPOL will explore the development of a toolkit for the responsible use of AI by law enforcement in 2020

United Nations Secretary-General's High Level Panel on Digital Cooperation

- **Report of the High-level Panel on Digital Cooperation** provides recommendations on how the international community could work together to optimize the use of digital technologies and mitigate the risks. **Recommendation 3C of the Report** has direct relevance to the ethics of artificial intelligence.

United Nations University (UNU)

- UNU Centre for Policy Research (UNU-CPR) in New York has worked on digital technology since 2013, this including contribution to the preparation of the **Secretary-General's Strategy on New Technologies** and the **report of the High-Level Panel on Digital Cooperation**. UNU-CPR also hosts the online thought leadership and engagement platform AI & Global Governance
- UNU-CPR has published a report entitled The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI in 2019, examining how the multilateral system can better understand and anticipate the risks that will come from AI convergence with cyber and biotechnologies.
- UNU Institute in Macau will be assembling a research team consisting of post-doctoral fellows and senior researchers well-known in the field of AI & ethics, focusing on the Global South. In particular, the Institute is setting up a **consortium on AI for social inclusion** to bring together experts in higher education institutes and other experts in AI policy, governance, design and deployment.

World Health Organization (WHO)

- WHO has established an expert group to develop a **Guidance Document on Ethics and Governance of Artificial Intelligence for Health**.

World Intellectual Property Organization (WIPO)

- WIPO has started an open process to discuss the **legal and policy implications of AI on IP**, with a list of the main questions and issues being developed concerning the impact of AI on IP policy. Outcome of the questionnaire may form the basis for future structured discussions.