United Nations
Educational, Scientific and
Cultural Organization

World Commission on
the Ethics of Scientific Knowledge
and Technology (COMEST)

SHS/COMEST/EXTWG-ETHICS-AI/2019/1
Paris, 26 February 2019
Original: English

PRELIMINARY STUDY ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

Building on the work of COMEST on Robotics Ethics (2017) and on the Ethical Implications of the Internet of Things (IoT), this preliminary study is prepared by a COMEST Extended Working Group on Ethics of Artificial Intelligence.

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

PRELIMINARY STUDY ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

TABLE OF CONTENT

PRELIMINARY STUDY ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

INTRODUCTION

1. The world is facing a rapid rise of 'Artificial Intelligence' (AI). Advancements in this field are introducing machines with the capacity to learn and to perform cognitive tasks that used to be limited to human beings. This technological development is likely to have substantial societal and cultural implications. Since AI is a cognitive technology, its implications are intricately connected to the central domains of UNESCO: education, science, culture, and communication. Algorithms have come to play a crucial role in the selection of information and news that people read, the music that people listen to, and the decisions people make. AI systems are increasingly advising medical doctors, scientists, and judges. In scientific research, AI has come to play a role in analysing and interpreting data. Furthermore, the ongoing replacement of human work by intelligent technologies demands new forms of resilience and flexibility in human labour. Public thinkers like Stephen Hawking have even voiced the fear that AI could bring an existential threat to humankind, because of its potential to take control of many aspects of our daily lives and societal organization.

2. In the 1950s, the term 'artificial intelligence' was introduced for machines that can do more than routine tasks. As computing power increased, the term was applied to machines that have the ability to learn. While there is not one single definition of AI, it is commonly agreed upon that machines which are based on AI, or on 'cognitive computing', are potentially capable of imitating or even exceeding human cognitive capacities, including sensing, language interaction, reasoning and analysis, problem solving, and even creativity. Moreover, such 'intelligent machines' can demonstrate human-like learning capabilities with mechanisms of self-relation and self-correction, on the basis of algorithms that embody 'machine learning' or even 'deep learning', using 'neural networks' that mimic the functioning of the human brain.

3. Recently, large multinational tech companies in many regions of the world have started to invest massively in utilizing AI in their products. Computing power has become large enough to run highly complicated algorithms and to work with 'big data': huge sets of data that can be used for machine learning. These companies have access to almost unlimited computing power and also to data collected from billions of people to 'feed' AI systems as learning input. Moreover, via their products, AI is rapidly gaining influence in people's daily lives and in professional fields like healthcare, education, scientific research, communications, transportation, security, and art.

4. This profound influence of AI raises concerns that could affect the trust and confidence people have in these technologies. Such concerns range from the possibility of criminality, fraud and identity theft to harassment and sexual abuse; from hate speech and discrimination to the spreading of disinformation; and more generally from the transparency of algorithms to the possibilities of trusting AI systems. Since many of these problems cannot be addressed by regulation alone, UNESCO has been proposing multi-stakeholder governance as an optimum modality to involve the various actors in the formulation and implementation of norms, ethics and policy, as well as the empowerment of users.

5. Because of its profound social implications, many organizations and governments are concerned about the ethical implications of AI. The European Commission has formed a High Level Expert Group on AI comprising representatives from academia, civil society, industry, as well as a European AI Alliance, which is a forum engaged in broad and open

discussion on all aspects of AI development and its impacts. The European Group on Ethics in Science and New Technologies has issued a *Statement on AI, Robotics, and Autonomous Systems* (EGE, 2018). The European Commission has published a *Communication on AI for Europe* (EC, 2018) and the Council of Europe has produced various reports on AI and has formed a Committee of Experts to work on the *Human rights dimensions of automated data processing and different forms of artificial intelligence*. The IEEE organization has formed a Global Initiative on Ethics of Autonomous and Intelligent Systems. The OECD has initiated the 'Going Digital' project, which aims to help policymakers in all relevant policy areas better understand the digital revolution that is taking place across different sectors of the economy and society as a whole. The OECD has also created an expert group (AIGO) to provide guidance in scoping principles for artificial intelligence in society. ITU and the WHO have established a Focus Group on "Artificial intelligence for Health". Furthermore, many countries have initiated reflection on their ethical and political orientation towards AI, like the Villani report in France (Villani et al., 2018); the House of Lords report in the UK (House of Lords, 2017); and the report of the Executive Office of the President of the USA (2016).

6. UNESCO has a unique perspective to add to this debate. AI has implications for the central domains of UNESCO's work. Therefore, in addition to the many ethical guidelines and frameworks that are currently being developed by governments, companies, and societal organizations, UNESCO can bring a multidisciplinary, universal and holistic approach to the development of AI in the service of humanity, sustainable development, and peace.

7. In this regard, there are several existing frameworks and initiatives to build on. Firstly, there is the *human rights* framework, which formed the basis of the 2003 World Summit on the Information Society's (WSIS) Geneva Declaration of Principles, stating that "the use of ICTs and content creation should respect human rights and fundamental freedoms of others, including personal privacy, and the right to freedom of thought, conscience, and religion in conformity with relevant international instruments" (WSIS, 2003). WSIS (2005) proposes a multi-stakeholder approach that calls for an effective cooperation of all stakeholders, including Governments, the private sector, civil society, international organizations, and the technical and academic communities. In the WSIS follow-up process, UNESCO has adopted this multi-stakeholder approach and 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).

8. Second, there is the framework of *Internet Universality* and the associated *R.O.A.M. principles* as approved by the 38th General Conference in 2015 (UNESCO, 2015b). These principles cover Human Rights, Openness, Accessibility and Multi-stakeholder participation, and have emerged from the UNESCO "Keystones" study for the 38th General Conference (UNESCO, 2015a). In the "Connecting the Dots" outcome document of this conference, UNESCO commits to promoting human rights-based ethical reflection, research and public dialogue on the implications of new and emerging technologies and their potential societal impacts. Moreover, the 18th Session of the Intergovernmental Council of UNESCO's Information For All Programme (IFAP) examined and approved the Code of Ethics for the Information Society, which was elaborated by the IFAP Working Group on Information Ethics.

9. To investigate the ethical implications of AI, this study will first explain what Artificial Intelligence is, how it works, and how it is different from other technologies. The second section will investigate the ethical aspects of AI, taking the UNESCO domains of education, science, culture, and communication as a starting point, as well as the global-

ethical dimensions of peace, cultural diversity, gender equality, and sustainability. This investigation should be seen as an exploration, not as a comprehensive analysis, ranging from cultural diversity to trust in science, from artistic creativity to critical thinking, and from AI-based decision-making to the role of AI in developing countries. The third section of this preliminary study will sketch out the central dimensions that proper ethical reflection on AI should have from the perspective of UNESCO.

I. WHAT IS AI?

I.1. Definition

10. The idea of 'artificial intelligence' (AI) – as the idea of 'artificially created' and 'intelligent' beings, machines or tools – is scattered throughout human history. Its various forms can be found in both Western and non-Western religions, mythologies, literature and philosophical traditions. As such, these instances testify to the perennial curiosity of humankind with such entities, and despite the expression of this curiosity through culturally diverse appearances, it is something shared or cross-cultural. Today, the fascination with AI – including its ethical dimensions – is amplified by its development and real-world applications.

11. Any examination of the ethical implications of AI need a clarification of its possible meanings. The term was coined in 1955, by John McCarthy, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon. The 'study of artificial intelligence' was planned "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 2006 [1955], p.12). As the field developed and diversified in the decades to come, the number of meanings of 'AI' increased and there is no universally agreed upon definition today. Various definitions of AI are related to different disciplinary approaches such as computer science, electrical engineering, robotics, psychology or philosophy.

12. Despite the multitude and diversity of definitions of AI, there is certain consensus, at the most general level, that its two aspects can be distinguished: one usually labelled as 'theoretical' or 'scientific' and the other one as 'pragmatic' or 'technological'.

13. To talk about 'theoretical' or 'scientific' AI is about "using AI concepts and models to help answer questions about human beings and other living things" (Boden, 2016, p.2). 'Theoretical' or 'scientific' AI thus naturally interconnects with disciplines like philosophy, logic, linguistics, psychology and cognitive science. It deals with questions like: What is meant by 'intelligence' and how to distinguish 'natural' from 'artificial' intelligence? Is symbolic language necessary for thought processes? Is it possible to create 'strong AI' (*genuine* intelligence of the same kind and level of generality as human intelligence) as opposed to 'weak AI' (intelligence that only *mimics* human intelligence and is able to perform a limited number of narrowly defined tasks)? Although questions like these are theoretical or scientific, they involve a number of metaphysical or spiritual concerns (e.g. about human uniqueness or the freedom of will) which themselves have indirect, but nonetheless serious, ethical implications.

14. 'Pragmatic' or 'technological' AI is engineering-oriented. It draws on various branches of AI – textbook examples are natural language processing, knowledge representation, automated reasoning, machine learning, deep learning, computer vision and robotics (Russell and Norvig, 2016, p.2-3) – in order to create machines or programs capable of independently performing tasks that would otherwise require human

intelligence and agency. 'Pragmatic' or 'technological' AI became remarkably successful as they are combined with ICT (information and communications technology). AI innovations are used today in many areas of modern life, such as transport, medicine, communication, education, science, finance, law, military, marketing, customer services or entertainment. These innovations do raise direct ethical concerns, ranging from the disappearance of traditional jobs, over responsibility for possible physical or psychological harm to human beings, to general dehumanization of human relationships and society at large. At the moment, no AI system can be considered as a general-purpose intelligent agent that can perform well in a wide variety of environments, which is a proper ability of human intelligence.

15. One of the particularities of AI concerns its 'unfamiliarity' to us humans in a sense that the way its intelligence works seems strange and mysterious to us. The essence of this 'unfamiliarity' is what one might call 'performance without awareness'. High-functioning AI such as AlphaGo or Watson can perform impressively without recognizing what it is doing. AlphaGo defeated a number of Go masters without even knowing that it was playing a human game called Go. Watson answered devilish questions so fast, which most humans have difficulties with even understanding in the given time. However, Watson is not 'answering' in the human sense; rather it 'computes' the probabilities of several candidate answers based on its automated analysis of an available database. AlphaGo and Watson perform brilliantly without being aware of what they are doing.

16. There are certainly important philosophical questions about whether the 'play' of AlphaGo and the 'answer' of Watson are 'genuine' or not. An ethically more crucial fact is however that we humans are not used to this kind of intelligence. Whenever we are confronted with impressive works of art, literature and science, we naturally consider the 'conscious' intelligence behind them. We recognize the unique character of Beethoven behind his 9th symphony, and the overwhelming searching mind behind Goedel's incompleteness theorem. The simple fact that we should not apply this familiar rule of thumb in regard to brilliant performances when we interact with high-functioning AI poses serious social and ethical challenges. As we are used to interacting emotionally and socially with behaviourally intelligent agents, we naturally interact emotionally and socially with 'high-functioning AI without awareness', such as so-called 'emotion' or 'social robots'- for example 'smart home assistant' (Alexa, Siri, Google assistant). At the current level of technological development, high-functioning AI without awareness cannot properly reciprocate complicated emotional and social expectations of human agents, while its external behaviour coupled with human imagination could generate an 'unrealistic' hope of genuine interactions with humans. It is important for us to remember that the seemingly 'emotional' mind of AI is much more of our imagination rather than of reality. There is general agreement that artificially intelligent systems do not have awareness in the experiential human sense, even if they can answer questions about the context of their actions. It is important not to equate experience with intelligence, even though some experts have suggested that recent developments in AI might also be a reason to re-examine the importance of this experience or awareness for being human. If experience is at the core of being human, ethical considerations must ensure that this is protected and enhanced through the use of AI rather than side-lined or disempowered. However, it may be that our experience with high-functioning AI without awareness can still influence our interactions with ordinary humans with awareness.

I.2. How does AI work?

17. To be able to perform the tasks of a human mind, an AI machine needs to be able

to sense the environment and to collect data dynamically, to process it promptly and to respond – based on its past 'experience', its pre-set principles for decision-making and its anticipation about the future. However, the technology behind AI is a standard ICT: it is based on collecting/acquiring data, storing, processing and communicating it. The unique features of cognitive machines come from quantities, which are transformed into qualities. AI technology is based on the following components:

a. *Dynamic data*. The system needs to be exposed to changing environments and to all relevant data acquired by various sensors, to classify and to store it, and to be able to process it promptly.
b. *Prompt processing*. Cognitive machines must react promptly. AI therefore needs to have reliable, fast and strong computing and communication resources.
c. *Decision-making principles*. AI decision-making is based on machine learning algorithms. Therefore, its response to a specific task depends on its 'experience' – that is, on the data it has been exposed to. The algorithms behind the decisions made by cognitive machines are based on some general principles the algorithm obeys and tries to optimize, given the data it is provided with.

The present ability to efficiently integrate dynamic data acquisition and machine-learning algorithms for prompt decision-making enables the creation of 'cognitive machines'.

I.3. How is AI different from other technologies?

18. Most 20th century technologies are model-driven. That is, scientists study nature and suggest a scientific model to describe it, and technology is advanced based on such models. For example, understanding the propagation of electromagnetic waves is the basis for the technology of wireless telecommunication. Modelling of the human brain is, however, a task which still seems far from being at a stage where a cognitive machine can be model-based. Therefore, AI is built on a different approach: a data-driven approach.

19. The data-driven approach is at the core of *machine learning*, which is commonly based on 'artificial neural networks' (ANNs). ANNs are formed by a series of nodes conceptually similar to brain neurons interconnected through a series of layers. The nodes of the input layer receive information from the environment, where, at each node, a non-linear transformation is applied. Such systems 'learn' to perform tasks by considering examples (labelled data), generally without being programmed with any task-specific rules or models. Deep learning, to conclude, is based on ANNs of several layers, which enables the machine to recognize complex concepts such as human faces, human bodies, speech understanding and all types of images classification.

20. The key issue in the ability of AI to show human-like capabilities is its scalability. The performance of AI machines depends on the data to which they are exposed, and for best performance, access to relevant data should be borderless. There may be technical limitations to the access to data, but the way data are selected and classified is also a socio-cultural issue (Crawford, 2017). Classification is culture-specific and a product of history, and may create bias in the decisions made by the algorithm. If the same machine is exposed to diverse sets of data, its bias can be reduced but not completely suppressed (Executive Office of the President, 2016). It is important to point out that in order to comply with what is mandated in Article 27 of the Universal Declaration of Human Rights – stating that every human being is entitled to the benefits of scientific progress – and to ensure

diversity in data sets available for AI, it is relevant to promote the capacity building of states, both in terms of human skills and infrastructure.

21. AI technology has matured under the drive of multinational companies that are not confined to local and national constraints. Moreover, to ensure prompt processing and reliability of the systems, the actual location of computing processes is distributed and the location of an AI machine is not defined by the place where it operates. Practically, AI is based on cloud technology, where the location of the storage and processing units can be anywhere. AI technology is characterized by the following:

a. While many of its applications are in the public sphere, AI technology is developed and led by multinational companies, most of them operating in the private sector and less obligated to the public good.
b. AI is not confined to a tangible location. This poses a challenge as regards how to regulate AI technology nationally and internationally.
c. The technology is based on accessibility to personal as well as public data.
d. AI technologies are not neutral, but inherently biased due to the data on which they are trained, and the choices made while training with the data.
e. AI and cognitive machine decisions cannot be fully predictable or explainable. Rather than operating mechanistically or deterministically, AI software learns from dynamic data as it develops and incorporates real-world experience into its decision-making.

II. ETHICAL CONSIDERATIONS

22. Artificial Intelligence has substantial societal and cultural implications. As many information technologies do, AI raises issues of freedom of expression, privacy and surveillance, ownership of data, bias and discrimination, manipulation of information and trust, power relations, and environmental impact in relation to its energy consumption. Moreover, AI brings specifically new challenges that are related to its interaction with human cognitive capacities. AI-based systems have implications for human understanding and expertise. Algorithms of social media and news sites can help to spread disinformation and have implications for the perceived meaning of ‘facts’ and ‘truth’, as well as for political interaction and engagement. Machine learning can embed and exacerbate bias, potentially resulting in inequality, exclusion and a threat to cultural diversity. The scale and the power generated by AI technology accentuates the asymmetry between individuals, groups and nations, including the so-called ‘digital divide’ within and between nations. This divide may be exacerbated due to lack of access to fundamental elements such as algorithms for learning and classification, data to train and to evaluate the algorithms, human resources to code, set up the software, and prepare the data, as well as computational resources for storage and processing of data.

23. As a result, Artificial Intelligence requires careful analysis. From UNESCO’s perspective, the most central ethical issues regarding Artificial Intelligence concern its implications for culture and cultural diversity, education, scientific knowledge, and communication and information. In addition to this, given UNESCO’s global orientation, the global-ethical themes of peace, sustainability, gender equality, and the specific challenges for Africa also deserve separate attention.

II.1. Education

24. Artificial Intelligence challenges the role of education in societies in many respects. Firstly, AI requires a rethinking of the societal role of education. The labour displacement caused by some forms of AI requires, among other measures, the retraining of employees, and a new approach to formulate the final qualifications of educational programmes. Moreover, in a world of AI, education should empower citizens to develop new forms of critical thinking, including 'algorithm awareness' and the ability to reflect on the impact of AI on information, knowledge, and decision-making. A second field of ethical questions regarding AI and education concerns its role in the educational process itself, as an element of digital learning environments, educational robotics, and systems for 'learning analytics', all of which require responsible development and implementation. Finally, engineers and software developers should be appropriately trained to ensure responsible design and implementation of AI.

II.1.1. The societal role of education

25. One of the main societal concerns regarding AI is labour displacement. The speed of change that AI is bringing presents unprecedented challenges (Illanes et al., 2018). It will involve, in the near future, the need to retrain large numbers of workers, and will have deep implications for the career paths students will need to follow. According to a McKinsey panel survey of 2017, "executives increasingly see investing in retraining and "upskilling" existing workers as an urgent business priority" (Illanes et al., 2018).

26. AI, therefore, will urge societies to rethink education and its social roles. Traditional formal education provided by universities might no longer be enough in the rise of digitized economies and AI applications. Until now, the standard education model has typically been to provide 'core knowledge' (Oppenheimer, 2018) and has focused on formal literacies like reading, writing and mathematics. In the 21st century, information and knowledge are omnipresent, demanding not only 'data literacy' that allows students to read, analyse and efficiently manage this information but also 'AI literacy' to enable critical reflection on how intelligent computer systems have been involved in the recognition of information needs, selection, interpretation, storage and representation of data.

27. Moreover, in a continuously developing labour market, the educational system can no longer aim to educate people for one specific profession. Education should enable people to be versatile and resilient, prepared for a world in which technologies create a dynamic labour market, and in which employees need to re-school themselves on a regular basis. Current ideas about 'lifelong learning' might need to be up-scaled into a model of continuous education, including the development of other types of degrees and diplomas.

II.1.2. AI in teaching and learning

28. Open educational resources (OER) have been an important addition to the learning landscape with the free availability of high quality lectures and other teaching resources through the internet. The potential of OERs to impact the education of people from across the world is unparalleled, but has yet to be fully realised as the limited completion rates for massive open online courses (MOOCs) demonstrates. The wide variety and depth of resources available has given rise to two problems. Firstly, the problem of finding the right resource for either an individual learner or a teacher wishing to reuse a resource in their own teaching materials. This has led to the second problem of reducing diversity through some resources becoming very popular at the expense of other potentially more relevant but less accessible content.

29. An example here is the Horizon 2020 project “X5GON” (Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network: https://www.x5gon.org/). This project, funded by the European Union, is developing Artificial Intelligence methods to enable both learners and teachers to identify resources that match their learning goals, taking into account the specifics of their situation. For example, a teacher in Africa might be directed to lectures that present a topic based on local and indigenous knowledge that is appropriate for the particular cultural and local context, but equally would enable a learner from elsewhere interested in understanding specific African challenges to find relevant African content potentially translated from a local language.

30. In this way, AI can potentially address both of the above-identified problems. The first problem is tackled through assisting in the identification of the resources that are better matched to the learner’s or teacher’s needs through modelling their interests and goals, while at the same time exploiting an enriched representation of the huge repositories of OERs available throughout the world. By tuning recommendations to the individual learner or teacher, it further addresses the second problem, as the recommendations will no longer default to the most popular resource on a particular topic. There is the further potential to link learners from different cultures to enhance cross-cultural sharing of ideas and hence supporting mutual understanding and respect.

II.1.3. Educating AI engineers

31. The development of future technologies is in the hands of technical experts. Traditionally, engineers are educated to develop products to optimize performance using minimum resources (power, spectrum, space, weight etc.), under given external constrains. Over the past decades, the ethics of technology has developed various methods to bring ethical reflection, responsibility and reasoning to the design process. In the context of AI, the term ‘ethically aligned design’ (EAD) has been developed to indicate design processes that explicitly include human values (IEEE, 2018).

32. It is most important to apply Ethically Aligned Design in AI and other autonomous, intelligent systems (AIS) because this makes it possible to address ethical issues at a moment when the technology can still be adapted. A good example is ‘privacy by design’. Privacy can be violated less if not all data is stored but only that which is required for a specific task. An example of this is crowd counting, i.e. counting people in a crowd based on photos. In this case, if the photo is pre-processed to extract only the contours (edges) of the figures, people will remain unrecognizable and the counting algorithm will perform well without violating privacy. Similarly, AI developers can consider other ethical issues such as the prevention of algorithmic bias and traceability, minimizing the ability to misuse the technology, and explainability of algorithmic decisions.

33. Global engineering education today is largely focused on scientific and technological courses that are not intrinsically related to the analysis of human values overtly designed to positively increase human and environmental wellbeing. It is most important to change this and to educate future engineers and computer scientists for ethically aligned design of AI systems. This requires an explicit awareness of the potential societal and ethical implications and consequences of the technology-in-design, and of its potential misuse. The IEEE (a global organization of more than 400,000 electrical engineers) already promotes this issue via its global initiative on the ethics of autonomous and intelligent systems (https://ethicsinaction.ieee.org/). Addressing this issue is also a matter of ensuring active efforts for gender inclusion as well as social and cultural diversity of engineers, and for a holistic application of societal and ethical implications of AI system

design. Occasions for dialogue between engineers and the public should be encouraged in order to facilitate communication on the needs and visions of society, and on how engineers really work and conduct research in their everyday activities.

II.2. AI and Scientific Knowledge

34. In the field of scientific practice, AI is likely to have profound implications. In the natural and social sciences as well as in the life sciences and environmental sciences, it challenges our concepts of scientific understanding and explanation in a fundamental way. This also has implications on how we apply scientific knowledge in social contexts.

II.2.1. AI and scientific explanation

35. Because of the increasingly powerful forms of machine learning and deep learning, AI challenges existing conceptions of satisfactory scientific explanation as well as what we can naturally expect from predictably successful scientific theories. In the conventional view of science, the so-called deductive-nomological model, a proper scientific explanation is able to make correct predictions of specific phenomena based on scientific laws, theories and observations. For instance, we can legitimately say that we explain how the moon is moving around the earth in terms of Newtonian mechanics only when we are able to employ Newtonian mechanics in a deductive way to predict the lunar orbit. Such predictions are typically based on causal understanding, or on a unifying understanding of seemingly disparate phenomena.

36. In contrast to this, AI can reliably produce impressively accurate predictions based on data sets without giving us any causal or unifying explanation of its predictions. Its algorithms do not work with the same semantic concepts that humans employ to achieve scientific understanding of a phenomenon. This gap between successful predictions on the one hand and satisfactory scientific understanding on the other is likely to play a key role in scientific practice, as well as in decision-making based on AI.

37. This might have implications for trust in science, which is typically based on the scientific method that explains different phenomena in a systematic and transparent way, making its predictions rational and evidence-based. The apparent success of machine learning algorithms to deliver comparable results without such a scientifically justified model could have implications for the public perception and evaluation of science and scientific research.

38. Moreover, research shows that the quality of machine learning depends heavily on the available data used to train the algorithms. But since most AI applications are developed by private companies, there is not always enough transparency about these data, in contrast to the traditional scientific method that warrants the validity of results by requiring replicability, i.e. the possibility to reproduce them by repeating the same experiments.

II.2.2. AI, life sciences and health

39. Within the life sciences and medicine in particular, the development of AI technologies has significantly transformed the health care and bioethics landscape over the years. They can bring positive effects, like more precision in robotic surgery, and better care for autistic children, but at the same time, they raise ethical concerns, such as the cost they bring within the context of scarcity of resources in the health care system and the transparency they should bring in order to respect the autonomy of patients.

40. From an individual perspective, AI is bringing a new way of dealing with health and medical issues for the lay public. The use of internet sites and the multiplication of mobile phone software applications for self-diagnosis have given people the opportunity to generate health diagnoses without the participation of a health professional. This might have implications for medical authority and for the acceptance of self-medication, including the dangers it entails. It also changes the doctor-patient relationship, and calls for some kind of regulation without hindering innovation and autonomy.

41. AI technologies might free up time for health providers to dedicate to their patients, for instance by facilitating data entry and deskwork, but at the same time, they might replace the holistic and human elements of care. The well-known technology Watson for Oncology by IBM is a breakthrough in cancer treatment, but also raises important questions regarding the character and expectations of medical expertise and education, and the responsibilities of doctors working with the system. Similar concerns are raised with the development of chatbots for people seeking psychological help and counselling, apps for early detection of episodes of psychiatric diseases, or AI systems for producing psychiatric diagnoses on the basis of information collected from people's activity on social media and the Internet – which obviously also has important implications for privacy. In addition, in the case of the elderly, AI-based technologies such as assistive social robots are being introduced which can be useful on medical grounds for patients with dementia for example, but also raise concerns about reduced human care and the resulting social isolation.

42. AI also brings a new dimension to the ongoing discussion about 'human enhancement' versus 'therapy'. There are initiatives to integrate AI with the human brain using a 'neural interface': a mesh growing with the brain, which would serve as a seamless brain-computer interface, circulating through the host's veins and arteries (Hinchliffe, 2018). This technological development has important implications for the question of what it means to be human, and what 'normal' human functioning is.

II.2.3. AI and environmental science

43. AI has the potential to be beneficial to environmental science through a number of different applications. It can be used to process and interpret data within ecology, systems biology, bioinformatics, space and climate research, thus enhancing scientific understanding of processes and mechanisms. Improved recycling, environmental monitoring and remediation, and more efficient energy consumption can have direct environmental benefits. AI in agriculture and farming can lead to improved crop production (e.g., automated fertilization and irrigation) and animal welfare, and reduced risks from disease, pests, or weather threats. On the other hand, AI could lead to changes in human perceptions of nature, either positively by enhancing human awareness of beauty or independency, or negatively through increased 'instrumentalization' of nature or separation between humans and animals or the environment.

44. For all applications, the potential benefits need to be balanced against the environmental impact of the entire AI and IT production cycle. This includes mining for rare-earth elements and other raw materials, the energy needed to produce and power the machines, and the waste generated during production and at the end of life cycles. Increased AI is likely to add to the growing concerns about the increasing volumes of e-waste and the pressure on rare-earth elements generated by the computing industry. In addition to the environmental and health impacts, e-waste has important socio-political implications, especially related to the export to developing countries and vulnerable populations (Heacock et al., 2015).

45. Disaster risk management is an area where AI can aid in the prediction and response to environmental hazards such as tsunamis, earthquakes, tornadoes and hurricanes. A concrete example is the UNESCO G-WADI Geoserver application (Water and Development Information), which is being used to inform emergency planning and management of hydrological risks, such as floods, droughts and extreme weather events. Its support system PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) is a satellite-based precipitation retrieval algorithm, providing near-real time information. The PERSIANN Cloud Classification System CCS algorithm (accessible at: http://hydis.eng.uci.edu/) has been optimized for observing extreme precipitation, particularly at very high spatial resolution and is being widely used globally to track storms. It also offers an iRain mobile application (http://en.unesco.org/news/irain-new-mobile-app-promote-citizen-science-and-support-water-management) where crowdsourcing gives opportunities for engaging citizen scientists in data collection.

46. Interestingly, even private companies have recently contributed to disaster management. One such example is Google's AI-enabled flood forecasting project (https://www.blog.google/products/search/helping-keep-people-safe-ai-enabled-flood-forecasting/). In this regard, the development of AI technologies that could bring potential benefit for disaster management should be encouraged.

II.2.4. AI and social sciences

47. Broadly speaking, social science research aims at finding out the causal structure of personal and social interactions. As most social phenomena are multiply influenced by a number of causal factors, social scientists typically rely on statistical analysis of the relevant empirical data to determine prominent causal factors and the strength of their effects. While doing so, it is crucial to distinguish mere statistical correlations from genuine causal connections. Certainly AI has clear potential to help social scientists navigate huge data sets to come up with plausible causal mechanisms as well as verify the validity of the proposed ones. On the other hand, AI can 'overfit' the data, and put forward 'pseudo' causal relations when there is none. This possibility could lead to social controversies especially when the proposed causal relations are ethically sensitive such as suggestions of racial differences of intelligence. Here again we should not accept AI's 'conclusions' automatically without human evaluation.

II.2.5. AI-based decision-making

48. AI methods can potentially have a huge impact in a wide range of areas, from the legal professions and the judiciary to aiding the decision-making of legislative and administrative public bodies. For example, they can increase the efficiency and accuracy of lawyers in both counselling and litigation, with benefits to lawyers, their clients and society as a whole. Existing software systems for judges can be complemented and enhanced through AI tools to support them in drafting new decisions (CEPEJ, 2018).

49. A key issue in such uses is the nature and interpretation of the results of algorithms, which are not always intelligible to humans[1]. This issue can be expanded to the wider field of data-driven decision-making. Being able to analyse, process and categorize very large

[1] As K.D. Ashley states: "since a Machine Learning (ML) algorithm learns rules based on statistical regularities that may surprise humans, its rules may not necessarily seem reasonable to humans. […] Although the machine-induced rules may lead to accurate predictions, they do not refer to human expertise and may not be as intelligible to humans as an expert's manually constructed rules. Since the rules the […] algorithm infers do not necessarily reflect explicit legal knowledge or expertise, they may not correspond to a human expert's criteria of reasonableness." (Ashley, 2017, p.111)

amounts of potentially rapidly-evolving data of very different natures, an AI engine is seen to be capable of proposing – and if allowed, making – decisions in complex situations. Examples of such uses discussed in this report include environmental monitoring, disaster prediction and response, anticipation of social unrest and military battlefield planning.

50. The validity of an AI-driven decision however should be treated with caution. Such a decision is not necessarily fair, just, accurate or appropriate. It is susceptible to inaccuracies, discriminatory outcomes, embedded or inserted bias and limitations of the learning process. Not only does a human have a much larger 'world view', but he or she also has a tacit knowledge that will outperform AI in critical and complex situations, such as battlefield decisions. Ideally, a decision would be the one a human would make if he or she had been able to process the mountain of data in a reasonable time. However, humans have different capabilities and make decisions based on fundamentally different decision-making architectures, including sensitivity to potential bias.

51. It is highly questionable that AI will – at least in the near future – have the capacity to cope with ambiguous and rapidly evolving data, or to interpret and execute what human intentions would have been if the human could have coped with complex and multifaceted data. Even having a human 'in the loop' to moderate a machine decision may not be sufficient to produce a 'good' decision: as cognitive AI does not make decisions in the same way as humans would, the human would not be equipped with the knowledge and information she or he would need in order to decide if the data-driven action fulfils the human's intentions. Moreover, the stochastic behaviour of cognitive AI, together with the human's consequent inability to know why a particular choice has been made by the system, means the choice is less likely to be trusted.

52. A cautionary tale that illustrates some of the problems of using AI to assist decision-making in social contexts is the Allegheny Family Screening Tool (AFST), a predictive model used to forecast child neglect and abuse in Allegheny, Pennsylvania (see https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/07/AFST-Frequently-Asked-Questions.pdf). The tool was put in place with the belief that data-driven decisions would provide the promise of objective, unbiased decisions that would solve the problems of public administration with scarce resources. The Authority that implemented this tool may have been well intentioned. However, recent research has argued that the AFST tool has harmful implications for the population it hoped to serve (Eubanks, 2018b, p.190; Eubanks, 2018a). It oversamples the poor and uses proxies to understand and predict child abuse in a way that inherently disadvantages poor working families. It thus exacerbates existing structural discrimination against the poor and has a disproportionately adverse impact on vulnerable communities.

53. In some contexts, employing AI as a (either human-assisted or fully autonomous) decision maker might even be seen as a pact with the devil: in order to take advantage of the speed and large data ingestion and categorization capabilities of an AI engine, we will have to give up the ability to influence that decision. Moreover, the effects of such decisions can be profound, especially in conflict situations.

II.3. Culture and Cultural Diversity

54. AI is likely to have substantial implications for culture and artistic expression. Although still in its infancy, we are beginning to see the first instances of artistic collaboration between intelligent algorithms and human creativity, which might eventually bring important challenges for the rights of artists, the Cultural and Creative Industries (CCI), and the future of heritage. At the same time, the role of algorithms in online

streaming media and in machine translation is likely to have implications for cultural diversity and language.

II.3.1. Creativity

55. Artificial Intelligence is increasingly connected to human creativity and artistic practice: ranging from 'autotune' software that automatically corrects the pitch of the voices of singers, to algorithms helping to create visual art, compose music or write novels and poetry. Creativity, understood as the capacity to produce new and original content through imagination or invention, plays a central role in open, inclusive and pluralistic societies. For this reason, the impact of AI on human creativity deserves careful attention. While AI is a powerful tool for creation, it raises important questions about the future of art, the rights and remuneration of artists and the integrity of the creative value chain.

56. The case of the 'Next Rembrandt' – in which a brand-new Rembrandt painting was produced using AI and a 3D printer – is a good illustration (Microsoft Europe, 2016). Works of art like this require a new definition of what it means to be an 'author', in order to do justice to the creative work of both the 'original' author and of the algorithms and technologies that produced the work of art itself. This raises another question: What happens when AI has the capacity to create works of art itself? If a human author is replaced by machines and algorithms, to what extent copyrights can be attributed at all? Can and should an algorithm be recognized as an author, and enjoy the same rights as an artist?

57. Although AI is clearly capable of producing 'original' creative works, people are always involved in the development of AI technologies and algorithms, and often in the creation of artworks that serve as the inspiration for AI-generated art. From this perspective, AI can be seen as a new artistic technique, resulting in a new type of art. If we want to preserve the idea of authorship in AI creations, an analysis of the various authors 'behind' each work of art, and their relationships with each other, need to be made. Accordingly, we need to develop new frameworks to differentiate piracy and plagiarism from originality and creativity, and to recognize the value of human creative work in our interactions with AI. These frameworks are needed to avoid the deliberate exploitation of the work and creativity of human beings, and to ensure adequate remuneration and recognition for artists, the integrity of the cultural value chain, and the cultural sector's ability to provide decent jobs.

II.3.2. Cultural diversity

58. AI also has a close relation to cultural diversity. While it has the potential to positively impact the cultural and creative industries, not all artists and entrepreneurs have the skills and resources to use AI-based technologies in the creation and distribution of their work. The commercial logic of large platforms may lead to an increased concentration of cultural supply, data and income in the hands of only a few actors, with potential negative implications for the diversity of cultural expressions more generally, including the risk of creating a new creative divide, and an increasing marginalization of developing countries.

59. As these platforms develop into the dominant means of enjoying works of art, it is crucial to ensure diversity and fair access to these platforms for artists from all genres and backgrounds. In this context, artists from developing countries require special consideration. Artists and cultural entrepreneurs should have access to the training, financing opportunities, infrastructure and equipment necessary to participate in this new cultural sphere and market.

60. Moreover, the algorithms used by media streaming companies such as Spotify and Netflix have a major influence on the selection of music and movies that people enjoy. Because these platforms not only make works of art available, but also *suggest* works of art for their users to enjoy, it is important that their algorithms are designed in such a way that they do not privilege specific works of art over others by limiting their suggestions to the most dominant works of a particular genre, or to the most popular choices of users and their peers. Other institutions have expressed similar concerns (ARCEP, 2018*).* Transparency and accountability of these algorithms are essential for ensuring access to diverse cultural expressions and active participation in cultural life.

61. Also in its relation to cultural heritage, AI can play an important role. AI can be used, for instance, to monitor and analyse changes to heritage sites, in relation to development pressures, climate change, natural disasters and armed conflicts. It can also be used to monitor the illicit trafficking of cultural objects and the destruction of cultural property, and to support data collection for recovery and reconstruction efforts.

II.3.3. Language

62. In our rapidly globalizing world, the machine-powered translation of languages is likely to play an increasingly important role. Because of this, AI will have a substantial impact on language and human expression, in all dimensions of life. This fact brings with it a responsibility to deal carefully with 'natural' languages (as opposed to artificial languages or computer code) and their diversity. Language, after all, is the basis for human identity, social cohesion, education, and human development. Since its founding, UNESCO has recognized the importance of language in promoting access to quality education, building inclusive knowledge societies and transmitting cultural heritage and expressions (UNESCO, 2002).

63. A central element of the complex relationship between AI and language is the intermediary role of 'formal languages' (languages with words derived from an alphabet). AI technologies often require that words and sentences expressed in any of the many natural languages used around the world have to be translated into formal languages that can be processed by computers. The translation of many natural languages into formal languages is not a neutral process, because every translation from natural language into formal language results in the 'loss' of meaning, given the fact that not all the specificities and idiosyncrasies of languages can be entirely formalized.

64. A second element is the translation between natural languages, which takes place via these formal languages. There are several intrinsic problems with machine translations: words can have different meanings in different languages, and there can be a lack of linguistic or conceptual correspondence between languages. In these cases, translation is very difficult, if not technically impossible. In addition, the contextual and cultural connotations of words and expressions are not always fully translatable. Although greatly improved in recent years, at least for more common languages, automatic translation or machine translation is often too unreliable to be used, for instance, in technical fields where lexical and conceptual precision is crucial, or in cultural expression and literature.

65. These two aspects of machine translation have important implications, not only for the quality of translation and the risk of inter-language misunderstanding, but also for linguistic diversity. It is very likely that machine translation, at least in the short term, will be primarily developed for the main world languages, especially English. The technology requires large data sets compiled from human-made translations. Such data sets are often not available in significant numbers for less spoken languages. At the same time, this

technology can also play a positive role, allowing people to express themselves in less widely spoken languages.

66. An analogous process has actually already taken place with radio. While commercial radio largely produces content in widely spoken languages, thus reinforcing the cultures embodied in dominant languages, community broadcasters often generate content in local languages, thus enhancing pluralism and diversity in the media. As the UNESCO handbook on Community Media states: "[Community media is] present in all regions of the world as social movements and community-based organizations have sought a means to express their issues, concerns, cultures and languages" (UNESCO, 2013, p.7). Mass media, therefore, can actually help to preserve languages and cultural diversity.

67. Similarly, machine translation has already been used as a tool to foster diversity and protect indigenous languages. For instance, in Australia, a researcher from the ARC *Centre of Excellence for the Dynamics of Language* has recorded nearly 50,000 hours of spoken word. To process these recordings, linguists needed to select short segments of the recording that might include key sections of grammar and vocabulary, by listening to the recordings and transcribing them. Without AI, this would have taken roughly 2 million hours. So far, this usage of AI has facilitated the modelling of 12 indigenous languages spoken in Australia, including *Kunwok, Kriol, Mangarayi, Nakkara, Pitjantjatjara, Warlpiri, Wubuy*, among others (O'Brien, 2018).

68. These examples show that AI, like any technology, should be developed and used in ways that do not threaten cultural diversity but rather preserve it. If we want to preserve multilingualism and interoperability among different languages, adequate technical and financial resources should be made available to make this possible (Palfrey and Gasser, 2012; Santosuosso and Malerba, 2015).

II.4. Communication and information

69. Artificial Intelligence plays an increasingly important role in the processing, structuring and provision of information. Automated journalism and the algorithmic provision of news on social media are just a few examples of this development, raising issues of access to information, disinformation, discrimination, freedom of expression, privacy, and media and information literacy. At the same time, attention is needed for new digital divides between countries and within different social groups.

II.4.1. Disinformation

70. AI can strengthen the free flow of information and journalistic activity, but it can also be used to spread disinformation, which is sometimes referred to using the contested term 'fake news'. Recent examples, such as the Cambridge Analytica affair, have shown that algorithms that were designed to avoid human political bias in deciding which content will appear prominently on social media can be taken advantage of for deliberately promoting the spreading of fabricated, manipulative and divisive content to specific target groups. In some cases, this content may include information fraudulently formatted as news, and may also include content that serves as emotive propaganda.

71. This can have negative effects on norms of civil and informed discussion, on social trust and public debate or even on democratic processes. The existence of different, sometimes polarized opinions is a regular feature of any open and democratic society that offers a free and open public space. Social media algorithms, however, may exacerbate the polarization of opinions by intensifying and amplifying emotional content via 'likes',

‘shares’, ‘retweets’, auto-completion in search queries, and other forms of online recommendations and engagement, resulting in so-called ‘filter bubbles’ and ‘echo chambers’ instead of providing an infrastructure for discussion and debate. Persons sharing the same ‘bubble’ may be exposed to filtered content of information and in return, the open public space can become characterized with more and more homogenized opinion groups which are at the same time more and more polarized to each other.

72. Although some big social media companies are beginning to recognize the problem and the need to address it in a multi-stakeholder way, which includes civil society together with state regulators, the solutions still seem to be unclear. One way to explore solutions is to use the UNESCO R.O.A.M. framework (Rights, Openness, Accessibility to all, Multi-stakeholder participation) to systematically identify where improvements can be made and how these interrelate with the totality of principles at stake.

73. Sometimes, the moderation of content can be justified precisely as a means to avoid spreading disinformation and content that incites violence, hatred and discrimination, as well as a means to prevent aggressive personal communication. The filtering may be done by humans, but is often assisted or even automated via AI algorithms. The particular challenge in this case is not just to identify the offending content, but also to avoid the filter being too inclusive and consequently incurring accusations of automated censorship and restriction on legitimate speech. Response to disinformation and ‘hate speech’ should be based on international freedom of expression standards and in line with UN conventions and declarations on the issue (Article 19, 2018a).

II.4.2. Data Journalism and Automated Journalism

74. The recent emergence of functionally powerful AI has implications for journalism in several different ways. On the one hand, the growing possibilities to use data and computer tools in journalistic research can strengthen journalistic work. On the other hand, AI might also take over some journalistic tasks. Routine tasks for which lots of ‘practice data’ are available are the first candidates to be mimicked by AI, and a substantial part of journalistic work is in fact routine: collecting and selecting relevant data, summarizing the results and describing them in a clear way. AI is already performing relatively simple, fixed-format jobs of article writing, in areas where continuous updates are needed, like market reviews or sports reporting. This development is ambivalent: it can also free journalists up to do higher end work in interpretation, analysis, verification and presentation of news.

75. Automated news writing without human intervention or supervision is a reality that is often hidden to the reader. As early as in 2006, some news services (e.g. Thomson Financial) announced the use of computers to generate stories based on data, in order to deliver information to their users in a fast manner. In 2014, Wibbitz (Israel) won the Netexplo Grand Prix at the UNESCO/Netexplo Forum, proposing an app that enables news channels to easily create videos using text content from the internet, providing a summary of the main ideas of the text. In recent times, a number of major mainstream media are using ‘robot journalism’: Le Monde, Press Association, Xinhua, to name a few, have reported to use natural language generation algorithms to cover different journalistic topics.

76. Media content production and dissemination increasingly delegate analytical and decision-making authority attributed to sophisticated algorithms. Media organizations increasingly rely on algorithms that analyse user preferences and media consumption patterns (personalization). Applied to journalism, algorithms are then called to analyse specific geographic communities for demographic, social, and political variables in order to produce the most relevant information for these communities, including weather

forecasts and sports reports. This practice has the potential to sustain local journalism and newspapers. In this way, AI can help strengthen business models for journalism.

77. At the same time, AI-based journalism raises issues of liability, transparency and copyright. Liability can be an issue when it is complicated to determine the fault in algorithm-based reporting, for instance in cases of defamation. Transparency and credibility are issues when consumers do not or cannot realize when content is machine-generated, from which sources it comes and how verified or even false the information is – with current discussions about 'deep fakes' as extreme cases. Copyright is an upcoming issue, since AI-generated content depends ever less on human input, which is reason for some to argue that some form of copyright liability should be attributed to the algorithms themselves.

78. To address these challenges, many argue that journalists and editors should engage with the technologists who build the algorithms. An example of this is the recent launch of an open-source platform by Quartz AI Studio, a US-based project to help journalists use machine learning supporting them in various tasks.

II.5. AI in Peace-Building and Security

79. In line with UNESCO's mission and mandate to promote and build peace, this study also wants to investigate the role of Artificial Intelligence in matters of peace-building and security. The fact that this includes the potential military use of AI in no way weakens its commitment to peace.

80. AI is argued to be capable of analysing, processing and categorizing very large amounts of rapidly evolving data of very different natures (Payne, 2018; Roff, 2018; Gupta, 2018). '*Hard'* data would include satellite and other surveillance imagery, signals and electronic intelligence, while *'soft'* data could include reports, documents, newsfeeds, social media inputs and political and sociological data. AI is advertised as being capable of categorizing this massive amount of data to identify external and internal threats, discover the objectives and strategies of actors, interpret complex and multifaceted intentions underpinning their activities, and strategies about how to pre-empt or counter predicted actions.

81. Such a situational awareness tool could be a powerful instrument for conflict prevention and resolution (Spiegeleire et al., 2017). It could give insight into the drivers of human endeavour and their outcomes, with possible application in deradicalization. Learning-enabled 'anticipatory intelligence' might foresee the development of social unrest and societal instability, and suggest means of prevention. Deeper insights into the drivers of conflict could nudge potential agents of conflict away from realizing malign intentions. We might be able to detect social pathologies at an early stage, find out which actions might de-escalate a threatening situation, or discover effective non-inflammatory routes to counter attempts to whip up sectarian frenzy. At a societal level, by tracking and helping us understand the dynamics that strengthen or weaken societal resilience, AI may be able to lead us to a more resilient society, and to help us move towards a more peaceful, conflict-free world.

82. On the negative side, *AI will transform the nature and practice of conflict,* with a consequential impact on society that will reach far beyond strictly military issues (Payne, 2018; Spiegeleire et al., 2017). Not only will it change how explosive force is used by increasing the effectiveness of deployment of weapons systems, but also AI promises to dramatically improve the speed and accuracy of everything from military logistics, intelligence and situational awareness to battlefield planning and execution/operations.

The AI system itself might be used to make its own suggestions of actions to be taken: it could create a set of orders that exploit enemy weaknesses that it has identified from its own analysis, or, from finding patterns in enemy/insurgent actions, devise countermeasures to predicted aggressive action. It might also do its own 'war gaming' to probe the likely responses to particular actions.

83. The speed with which such planning tools could operate would increase the ability to act under rapidly changing situations. One can envisage, for example, the development of algorithmic response to coordinated attack by e.g. drone swarms and other uninhabited assets such as incoming missiles. The speed of AI-enabled response can be seen as an incentive to use it and hence be potentially destabilising. Or indeed disastrous, as past examples of machine warnings being thankfully not being acted on by an intervening human commander have demonstrated. Nevertheless, a State that does not go down this AI response route would be at a major disadvantage, thus encouraging proliferation of the capability.

84. The possibility exists of the AI-assisted decision-making machine implementing its own attack and kill decisions without human intervention – for example, a fully autonomous weapon. The idea of such a *non-human* entity having specific agency could radically change our understanding of politics at the widest levels. Moreover, the closeness of potential military uses of AI to its civilian development ('ease of weaponisation') means it is not a discretely bounded category, a characteristic which complicates both the ethics and the regulation of its development and application.

85. While AI might be considered to be just another revolution in military affairs that allows armed forces to do similar things with similar tools, perhaps its real 'revolutionary' potential (Payne, 2018; Spiegeleire et al., 2017) is in *transforming the concept of 'armed force'* into one whose weapons are more subtle than explosive devices. The power of AI in conflict lies not only in enhancing physical technologies, but also in redefining what 'armed force' might be.

86. We are already seeing this in the cyber context, where AI gives it both defence and attack capability. Through pattern matching, deep learning and observing deviations from normal activity, software vulnerabilities can be detected and then weaponized to avoid defences. Deep neural networks may detect and prevent intrusions. In order to be effective, cyber defences will have to operate at speed and by implication have a high degree of autonomy.

87. Propaganda is another weapon that AI has empowered. The ease of faking voices, images and news, and propagating them to selected audiences, threatens social engineering and (mis)-shaping of public opinion. In essence, AI makes it easier to lie persuasively and enhance forgery. The consequent threat to trust in the integrity of information increases the potential for miscalculation of a perceived adversary's intention both tactically and strategically.

88. AI also empowers economic sabotage and critical infrastructure disruption. By moving radio and electronic warfare into cognitive mode, AI could be critical in interfering access to the electromagnetic spectrum. Systems are already marketed, which use machine learning, 'intelligent' algorithms and adaptive signal processing.

89. Finally, with respect to *internal* state security, the use of data set analysis and face recognition implies a new relationship between society and the institutions charged with protecting it. This obviously has significant ethical implications.

II.6. AI and Gender Equality

90. AI systems have significant implications for gender equality, since they may reflect existing societal biases, with the potential to exacerbate them. Most AI systems are built using datasets that reflect the real world – one which can be flawed, unfair, and discriminatory (Marda, 2018). Recently, a hiring tool used by Amazon was found to be sexist, as it prioritized male applicants for technical jobs (Reuters, 2018). Such systems can be dangerous, not only because they perpetuate gender inequalities in society, but also because they embed these inequalities in opaque ways, while at the same time being hailed as 'objective' and 'accurate' (O'Neil, 2018).

91. These inequalities are primarily a result of how machines learn. Indeed, as machine learning relies on the data it is fed, particular attention is needed to promote gender-sensitive data as well as gender-disaggregated data collection. In the case of Amazon's hiring tool, the bias emerged because the tool was learning from previous Amazon candidates – who were predominantly male – and had 'learnt' that male applicants should be preferred over female applicants (Short, 2018). Paying attention to biased data would therefore help limit the blindspot of how AI systems can best be suited and designed for both men and women. Additionally, applying gender-disaggregated data to AI analytics represents an opportunity to better grasp gender issues we face today.

92. It is important to note that gender inequalities begin at the early stages of conceptualising and designing AI systems. The gender disparity in technical fields is well known and apparent (Hicks, 2018), from wage gaps to promotions (Brinded, 2017). This is generally known as the 'leaky pipeline', with female participation in tech and engineering dropping 40% between when students graduate, to when they become executives in the field (Wheeler, 2018). The low share of women in the AI workforce – and in digital skills development in general – means that women's voices are not equally represented in the decision-making processes that go into the design and development of AI systems. As a result, we risk building these technologies only for some demographics (Crawford, 2016).

93. Moreover, the biases that people carry in their everyday lives can be reflected and even amplified through the development and use of AI systems. The 'gendering' of digital assistants, for example, may reinforce understandings of women as subservient and compliant. Indeed, female voices are routinely chosen as personal assistance bots, mainly fulfilling customer service duties, whilst the majority of bots in professional services such as the law and finance sectors, for example, are coded as male voices. This has educational implications with regards to how we understand 'male' vs 'female competences, and how we define authoritative versus subservient positions. Further, the notion of 'gender' in AI systems is often a simple choice - male or female. This ignores and actively excludes transgender individuals, and can discriminate against them in humiliating ways (Costanza-Chock, 2018).

II.7. Africa and AI Challenges

94. Africa, like other developing regions, is facing the acceleration of the use of information technologies and AI. The new digital economy that is emerging presents important societal challenges and opportunities for African creative societies.

95. Concretely, in terms of infrastructure connectivity, Africa has a very large deficit and is significantly behind other developing regions; domestic connections, regional links and continuous access to electricity are a big handicap. Infrastructure services are paid at

a high price even if more and more Africans - even in town's slums - have their own mobile phones.

96. The development issues that African countries face are numerous. The human rights framework and the Sustainable Development Goals (SDGs) provide a consistent way to orient the development of AI. Therefore, how can AI technology and knowledge be shared and oriented through the priorities defined by developing countries themselves? These include challenges such as infrastructure, skills, knowledge gaps, research capacities and availability of local data, as expressed during the UNESCO Forum on Artificial Intelligence in Africa that took place at the Mohammed VI Polytechnic University, in Benguerir, Morocco, on 12 and 13 December 2018.

97. The role of women is crucial. As very dynamic economic agents in Africa, women carry out the majority of agricultural activities, hold one-third of all businesses and may represent, in some countries, up to 70% of employees. They are the main levers of the domestic economy and family welfare, and play an absolutely indispensable leadership role in their respective communities and nations. By placing gender equality at the centre of its strategy for promoting development in Africa, the African Development Bank recognizes the fundamental role of gender parity in achieving inclusive growth and in the emergence of resilient societies. Access to education, to AI literacy and more globally to information and communication technologies (ICTs) are key elements to empower women in order to avoid their marginalization.

98. With particular attention to scientific research, science, technology, engineering and mathematics, together with education for citizenship based on values, rights and obligations, AI should be integrated into national development policies and strategies by drawing on endogenous cultures, values and knowledge in order to develop African economies.

III. STANDARD-SETTING INSTRUMENT

III.1. Declaration vs Recommendation

99. The Working Group carefully examined two of UNESCO's normative tools – the Declaration and the Recommendation –, which were linked to the analyses of the two first sections of this preliminary study on the Ethics of AI. The Working Group also drew on COMEST's previous experience, which initiated the 2017 Declaration of Ethical Principles in relation to Climate Change and participated in the revision of the 2017 Recommendation on Science and Scientific Researchers. The Working Group weighed the pros and cons of each of these two normative tools.

100. With regard to the proposal for a Declaration on the Ethics of Artificial Intelligence, the Working Group noted the very recent increase in the number of declarations of ethical principles on AI in 2018. The *Montreal Declaration for a Responsible Development of AI* (University of Montreal, 2018), the *Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems* (Amnesty International and Access Now, 2018), and the Declaration of the Future of Life Institute on the *Asilomar AI Principles* (Future of Life Institute, 2017) come from different initiatives and are supported by various organizations (universities, governments, professional associations, companies, NGOs). We must add to this set of declarations several ethical proposals such as: the *Ethics Guidelines for Trustworthy AI* from the European Commission's High-Level Expert Group on AI, which is based on Human Rights; and the second document of the IEEE (currently under consultation) on *Ethically Aligned Design: A Vision for Prioritizing Human Well-being*

with Autonomous and Intelligent Systems, which is addressed to engineers and aimed to embed values into Autonomous Intelligent Systems. All these initiatives are positive as they initiate discussions on AI ethics at different levels.

101. Nevertheless, the Working Group concluded that there was a great heteronomy in the principles and in the implementation of the values promoted by one or the other. This heteronomy is both the consequence of the definition that has been chosen for AI, and of the objectives that are being sought - governance, education for engineers, public policy. The question is as follows: would a UNESCO Declaration on the Ethics of AI allow this heteronomy to be federated under a few guiding principles that would respond in a comprehensive manner to the ethical issues of AI, as well as to UNESCO's specific concerns in the fields of education, culture, science and communication? The Working Group believes that this could be possible, but with the risk that during the process leading to the Declaration, Member States would essentially agree on some general, abstract and non-binding principles, since it is a Declaration. In such a perspective, would a UNESCO Declaration on the Ethics of AI bring added value vis-a-vis other ongoing declarations and initiatives? It is questionable that such instrument will immediately establish itself as an international reference, in a context of competition between ethical frameworks, at a time when technologies are emerging and their uses not yet stabilized.

102. The Working Group therefore considered whether a Recommendation would then be a more appropriate tool in the current situation. At the international level, the European level and the national political context for several countries, there is a move towards similar forms of regulation with respect to the digital economy, but also taking into account the relations between the two major digital powers - USA and China. The increase of criticisms concerning the non-transparency, biases or ways of acting by big companies, or the rise of popular mistrust in the face of cyber-attacks are creating a new political climate that is having an impact on the development of AI. The digital regulation movement, initiated by the European Union on the protection of personal data, could therefore be extended to an international level in emerging fields such as AI. However, at this level, the tools are still in their early stages of development, although the OECD's strategy through its Artificial Intelligence Expert Group (AIGO) emphasizes responsibility, security, transparency, protection and accountability:

> The OECD supports governments through policy analysis, dialogue and engagement and identification of best practices. We are putting significant efforts into work on mapping the economic and social impacts of AI technologies and applications and their policy implications. This includes improving the measurement of AI and its impacts, as well as shedding light on important policy issues such as labour market developments and skills for the digital age, privacy, accountability of AI-powered decisions, and the responsibility, security and safety questions that AI generates. (OECD, 2019)

103. OECD public policy priorities are more a matter of AI governance and good practice. It seems here that UNESCO's approach could be complementary at the international level to the OECD’s, but with a focus on aspects that are generally neglected such as culture, education, science and communication. These dimensions directly affect people and populations in their daily lives and in their individual and collective aspirations. UNESCO's approach for a Recommendation on AI Ethics would be presented as a complementary alternative to a vision of economic governance. The Working Group therefore believes that by initiating a Recommendation, although it requires more time and energy than a Declaration, UNESCO would be able to distinguish itself not only in terms of ethical content but also through specific proposals to Member States. One of the aims

is to empower and strengthen the capacity of States to intervene in key areas that are impacted by the development of AI, such as culture, education, science and communication.

104. The Recommendation should contain two dimensions. The first is the affirmation of a number of basic principles for an Ethics of AI. The second is the outlining of specific proposals to help States monitor, and regulate the uses of AI in the areas under UNESCO's mandate through the reporting mechanism of the Recommendation, as well as identify ethical assessment tools to review on a regular basis their policies for guiding the development of AI. In this regard, UNESCO would be uniquely positioned to provide a multidisciplinary perspective, as well as a universal platform for the development of a Recommendation on the Ethics of AI. Specifically, UNESCO would be able to bring together both developed and developing countries, different cultural and moral perspectives, as well as various stakeholders within the public and private spheres into a truly international process for elaborating a comprehensive set of principles and proposals for the Ethics of AI.

105. The next section identifies some of these proposals.

III.2. Suggestions for a standard-setting instrument

106. On the basis of its analysis of the potential implications of Artificial Intelligence for society, the Working Group would like to suggest a number of elements that could be included in an eventual Recommendation on the Ethics of AI. These suggestions embody the global perspective of UNESCO, as well as UNESCO's specific areas of competence.

107. First of all, the Working Group would like to suggest a number of generic principles for the development, implementation and use of AI. These principles are:

a. **Human rights**: AI should be developed and implemented in accordance with international human rights standards.
b. **Inclusiveness**: AI should be inclusive, aiming to avoid bias and allowing for diversity and avoiding a new digital divide.
c. **Flourishing**: AI should be developed to enhance the quality of life.
d. **Autonomy**: AI should respect human autonomy by requiring human control at all times.
e. **Explainability**: AI should be explainable, able to provide insight into its functioning.
f. **Transparency**: The data used to train AI systems should be transparent.
g. **Awareness and literacy**: Algorithm awareness and a basic understanding of the workings of AI are needed to empower citizens.
h. **Responsibility**: Developers and companies should take into consideration ethics when developing autonomous intelligent system.
i. **Accountability**: Arrangements should be developed that will make possible to attribute accountability for AI-driven decisions and the behaviour of AI systems.
j. **Democracy**: AI should be developed, implemented and used in line with democratic principles.
k. **Good governance**: Governments should provide regular reports about their use of AI in policing, intelligence, and security.

l. **Sustainability**: For all AI applications, the potential benefits need to be balanced against the environmental impact of the entire AI and IT production cycle.

108. More specifically, the Working Group would like to point out some central ethical concerns regarding the specific focus of UNESCO:

a. **Education**: AI requires that education fosters AI literacy, critical thinking, resilience on the labour market, and educating ethics to engineers.

b. **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.

c. **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.

d. **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.

e. **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.

f. **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.

g. **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.

h. **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.

BIBLIOGRAPHY

AI Now. 2016. *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*. New York, The White House and the New York University's Information Law Institute. Available at: https://ainowinstitute.org/AI_Now_2016_Report.pdf

Ajunwa, I., Crawford, K., and Schultz, J. 2017. Limitless Worker Surveillance. *California Law Review.* No. 735, pp. 101-142.

Allen, G. and Chan, T. 2017. Artificial Intelligence and National Security. *Harvard Kennedy School, Belfer Center for Science and International Affairs.* Online. Available at: https://www.belfercenter.org/publication/artificial-intelligence-and-national-security

Amnesty International and Access Now. 2018. *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems.* Toronto, RightsCon 2018. Available at: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

ARCEP (Autorité de régulation des communications électroniques et des postes). 2018. *Smartphones, tablets, voice assistants... Devices, the weak link in achieving an open Internet*. Paris, ARCEP. Available at: https://www.arcep.fr/uploads/tx_gspublication/rapport-terminaux-fev2018-ENG.pdf

Article 19. 2018a. *Free speech concerns amid the "fake news" fad*. Online. Available at: https://www.article19.org/resources/free-speech-concerns-amid-fake-news-fad/

Article 19. 2018b. *Privacy and Freedom of Expression in the Age of Artificial Intelligence*. Online. Online. Available at: https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf

Ashley, K.D. 2017. Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge, Cambridge University Press.

Boden, M.A. 2016. *AI: Its Nature and Future*. Oxford, Oxford University Press.

Brinded, L. 2017. "Robots are going to turbo charge one of society's biggest problems", *QUARTZ* (28 December 2017). Online. Available: https://qz.com/1167017/robots-automation-and-ai-in-the-workplace-will-widen-pay-gap-for-women-and-minorities/

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. and Anderson, H. 2018. *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Available at: https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

Bunnin, N. and Yu, J. 2008. *The Blackwell dictionary of western philosophy*. John Wiley & Sons.

Butterfield, A., Ngondi, G.E. and Kerr, A. eds. 2016. *A dictionary of Computer Science*. Oxford, Oxford University Press.

Costanza-Chock, S. 2018. “Design justice, AI, and escape from the matrix of domination”, *Journal of Design and Science.* Online. Available at: https://jods.mitpress.mit.edu/pub/costanza-chock

Crawford, K. 2016. “Artificial Intelligence’s White Guy Problem”, *The New York Times* (Opinion, 25 June 2016). Online. Available at: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Crawford, K. 2017. ‘The Trouble with Bias’, NIPS 2017 Keynote. Available at: https://www.youtube.com/watch?v=fMym_BKWQzk

Cummings, M. L., Roff, H. M., Cukier, K., Patakilas, J. and Bryce, H. 2018. *Artificial Intelligence and International Affairs: Disruption Anticipated.* Chatham House Report. Available at: https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf

Brookfield Institute and Policy Innovation Hub (Ontario). 2018. *Policymakers: Understanding the Shift*. Online. Available at: https://brookfieldinstitute.ca/wp-content/uploads/Brookfield-Institute_-The-AI-Shift.pdf

Eubanks, V. 2018a. “A Child Abuse Prediction Model Fails Poor Families”, *WIRED*. Online. Available at: https://www.wired.com/story/excerpt-from-automating-inequality/

Eubanks, V. 2018b. *Automating Inequality: How high tech tools profile, police, and punish the poor*. New York, St. Martin’s Press.

European Commission (EC). 2018. *Artificial Intelligence for Europe.* 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. Brussels, European Commission. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN

European Commission for the Efficiency of Justice (CEPEJ). 2018. *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment.* Strasbourg, CEPEJ. Available at: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

European Group on Ethics in Science and New Technologies (EGE). 2018. *Statement on AI, Robotics, and Autonomous System*. Brussels, European Commission. Available at: https://publications.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382

Executive Office of the President (USA). 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Washington, D.C., Executive Office of the President. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Frankish, K. and Ramsey, W.M. eds. 2014. *The Cambridge handbook of artificial intelligence*. Cambridge, Cambridge University Press.

Future of Life Institute. 2017. *Asilomar AI Principles.* Cambridge, Future of Life Institute. Available at: https://futureoflife.org/ai-principles/?cn-reloaded=1

Gupta, D.K. 2018. "Military Applications of Artificial Intelligence", *Indian Defence Review* (22 March 2019). Online. Available at: http://www.indiandefencereview.com/military-applications-of-artificial-intelligence/

Heacock, M., Kelly, C.B., Asante, K.A., Birnbaum, L.S., Bergman, Å.L., Bruné, M.N., Buka, I., Carpenter, D.O., Chen, A., Huo, X. and Kamel, M. 2015. "E-waste and harm to vulnerable populations: a growing global problem", *Environmental health perspectives*, Vol. 124, No. 5, pp. 550-555.

Hicks, M. 2018. "Why tech's gender problem is nothing new", *The Guardian* (12 October 2018). Online. Available at: https://amp.theguardian.com/technology/2018/oct/11/tech-gender-problem-amazon-facebook-bias-women?__twitter_impression=true

Hinchliffe, T. 2018. "Medicine or poison? On the ethics of AI implants in humans", *The Sociable*. Online. Available at: https://sociable.co/technology/ethics-ai-implants-humans/

House of Lords. 2017. *AI in the UK: ready, willing and able?* London, House of Lords Select Committee on Artificial Intelligence. Available at: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Illanes, P., Lund, S., Mourshed, M., Rutherford, S. and Tyreman, M. 2018. *Retraining and reskilling workers in the age of automation*. Online, McKinsey Global Institute. Available at: https://www.mckinsey.com/featured-insights/future-of-work/retraining-and-reskilling-workers-in-the-age-of-automation

Institute of Electrical and Electronic Engineers (IEEE). 2018. *Ethically Aligned Design – Version 2 for Public Discussion*. New Jersey, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: https://ethicsinaction.ieee.org/

Laplante, P.A. 2005. *Comprehensive dictionary of electrical engineering*. Boca Raton, CRC Press.

Latonero, M. 2018. *Governing Artificial Intelligence: Upholding Human Rights & Dignity*. Data & Society. Available at: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Marda, V. 2018. "Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making", *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*. Online. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3240384

Matias, Y. 2018. Keeping people safe with AI-enabled flood forecasting. *The Keyword* (24 September 2018). Online. Available at: https://www.blog.google/products/search/helping-keep-people-safe-ai-enabled-flood-forecasting/

Matsumoto, D.E. 2009. *The Cambridge dictionary of psychology.* Cambridge, Cambridge University Press.

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E. 2006 [1955]. "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence", *AI Magazine*, vol. 27, no. 4, pp.12-14.

Microsoft Europe. 2016. "The Next Rembrandt", *Microsoft News Centre Europe*. Online. Available at: https://news.microsoft.com/europe/features/next-rembrandt/

National Science and Technology Council (USA). 2016. *The National Artificial Intelligence Research and Development Strategic Plan*. Washington, D.C., National Science and Technology Council. Available at: https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

O'Brien, A. 2018. "How AI is helping preserve Indigenous languages", *SBS News*. Online. Available at: https://www.sbs.com.au/news/how-ai-is-helping-preserve-indigenous-languages

O'Neil, C. 2018. "Amazon's Gender-Biased Algorithm Is Not Alone", *Bloomberg Opinion* (16 October 2018). Online. Available at: https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone

OECD. 2019. *Going Digital.* Paris, OECD. Available at: http://www.oecd.org/going-digital/ai/

Oppenheimer, A. 2018. *¡Sálvese quien pueda!: El futuro del trabajo en la era de la automatización*. New York, Vintage Espanol.

Palfrey, J.G. and Gasser, U. 2012. *Interop: The Promise and Perils of Highly Interconnected Systems*. New York, Basic Books.

Payne, K. 2018. "Artificial Intelligence: A Revolution in Strategic Affairs?", *Survival*, Vol. 60, No. 5, pp. 7-32.

Peiser, J. 2019. "The Rise of the Robot Reporter", *The New York Times* (5 February 2019). Online. Available at: https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html

Reuters. 2018. "Amazon ditched AI recruiting tool that favored men for technical jobs", *The Guardian* (11 October 2018). Online. Available at: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Roff, H.M. 2018. 2018. "COMPASS: a new AI-driven situational awareness tool for the Pentagon?", *Bulletin of the Atomic Scientists* (10 May 2018). Online. Available at: https://thebulletin.org/2018/05/compass-a-new-ai-driven-situational-awareness-tool-for-the-pentagon/

Rosenberg, J.M. 1986. *Dictionary of artificial intelligence and robotics*. New York, John Wiley & Sons.

Russell, S.J. and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*, 3rd ed. Harlow, Pearson.

Santosuosso, A. and Malerba, A. 2015. "Legal Interoperability As a Comprehensive Concept in Transnational Law", *Law, Innovation and Technology*, Vol. 5, No. 1, pp. 51-73.

Short, E. 2018. "It turns out Amazon's AI hiring tool discriminated against women", *Siliconrepublic* (11 October 2018). Online. Available at: https://www.siliconrepublic.com/careers/amazon-ai-hiring-tool-women-discrimination

Spiegeleire, S. De, Maas, M. and Sweijs, T. 2017. *Artificial Intelligence and the Future of Defence*. The Hague, The Hague Centre for Strategic Studies.

UNI Global Union. 2016. Top 10 principles for ethical artificial intelligence. Switzerland, UNI Global Union. Available at: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

UNICEF. 2017. *Children in a Digital World*. New York UNICEF. Available at: https://www.unicef.org/publications/files/SOWC_2017_ENG_WEB.pdf

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2002. *UNESCO Universal Declaration on Cultural Diversity: a vision, a conceptual platform, a pool of ideas for implementation, a new paradigm*. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000127162

UNESCO. 2013. *Community Media: A Good Practice Handbook*. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000215097

UNESCO. 2015a. *Keystones to foster inclusive knowledge societies: access to information and knowledge, freedom of expression, privacy and ethics on a global internet.* Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000232563

UNESCO. 2015b. *Outcome document of the "CONNECTing the Dots: Options for Future Action" Conference*. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000234090

University of Montreal. 2018. *Montreal Declaration for a Responsible Development of AI*. Montreal, University of Montreal. Available at: https://www.montrealdeclaration-responsibleai.com/

Vernon, D. 2014. *Artificial cognitive systems: A primer*. Cambridge, MIT Press.

Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A.-C., Levin, F. and Rondepierre, B. 2018. *For A Meaningful Artificial Intelligence: Towards a French and European Strategy*. Paris. Available at: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

Wheeler, T. 2018. ”Leaving at Lightspeed : the number of senior women in tech is decreasing”, *OECD Forum* (23 March 2018). Online. Available: https://www.oecd-forum.org/users/91062-tarah-wheeler/posts/31567-leaving-at-lightspeed-the-number-of-senior-women-in-tech-is-decreasing

World Summit on the Information Society (WSIS). 2003. *Declaration of Principles. Building the Information Society: A global challenge in the new Millenium*. Geneva, WSIS. Available at: http://www.itu.int/net/wsis/docs/geneva/official/dop.html

WSIS. 2005. *Tunis Agenda for the Information Society*. Tunis, WSIS. Available at: http://www.itu.int/net/wsis/docs2/tunis/off/6rev1.html

ANNEX: COMPOSITION OF THE COMEST EXTENDED WORKING GROUP ON ETHICS AND AI

1. **Prof. (Mr) Peter-Paul VERBEEK (Co-Coordinator)**
   Professor of Philosophy of Technology at the University of Twente, Netherlands
   Member of COMEST (2016-2019)

2. **Prof. (Mrs) Marie-Hélène PARIZEAU (Co-Coordinator)**
   Professor, Faculty of Philosophy, Université Laval, Québec, Canada
   Member of COMEST (2012-2019)
   Chairperson (2016-2019) and Vice-Chairperson (2014-2015) of COMEST

3. **Prof. (Mr) Tomislav BRACANOVIĆ**
   Research Associate, Institute of Philosophy, Zagreb, Croatia
   Member of COMEST (2014-2021)
   Rapporteur of COMEST (2018-2019)

4. **Mr John FINNEY**
   Emeritus Professor of Physics, Department of Physics and Astronomy, London, United Kingdom
   Coordinator of the Working Group on Scientific Ethics, Pugwash Conference on Science and World Affairs
   Ex-officio Member of COMEST

5. **Mr Javier JUAREZ MOJICA**
   Commissioner, Board of the Federal Telecommunications Institute of Mexico, Mexico City, Mexico
   Member, OECD Expert Group on AI (AIGO)
   Member of COMEST (2018-2021)

6. **Mr Mark LATONERO**
   Research Lead, Data and Human Rights, Data & Society

7. **Ms Vidushi MARDA**
   Digital Programme Officer at ARTICLE 19

8. **Prof. (Ms) Hagit MESSER-YARON**
   Professor of Electrical Engineering and former Vice-President for Research and Development, University of Tel Aviv, Tel Aviv, Israel
   Member, Executive Committee, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
   Member of COMEST (2016-2019)

9. **Dr (Mr) Luka OMLADIC**
   Lecturer, University of Ljubljana, Ljubljana, Slovenia
   Member of COMEST (2012-2019)

10. Prof. (Mrs) Deborah OUGHTON
Professor and Research Director, Centre for Environmental Radioactivity, Norwegian University of Life Sciences
Member of COMEST (2014-2021)

11. Prof. (Mr) Amedeo SANTOSUOSSO
Founder and Scientific Director, European Center for Law, Science and new Technologies (ECLT), University of Pavia, Pavia, Italy,
President, First Chamber, Court of Appeal of Milan, Italy
Member of COMEST (2018-2021)

12. Prof. (Mr) Abdoulaye SENE
Environmental sociologist, Coordinator for “Ethics, Governance, Environmental and Social Responsibility”, Environmental Sciences Institute, Cheikh Anta Diop University, Dakar, Senegal
Member of COMEST (2012-2019)
Vice-Chairperson of COMEST (2016-2019)

13. Prof. (Mr) John SHAWE-TAYLOR
UNESCO Chair in Artificial Intelligence, University College of London and Chair of the Knowledge 4 All Foundation

14. Mr Davide STORTI
Programme Specialist, Section for ICT in Education, Science and Culture, Communication and Information Sector, UNESCO

15. Prof. (Mr) Sang Wook YI
Professor of Philosophy, Hanyang University, Seoul, Republic of Korea
Member of COMEST (2018-2021)