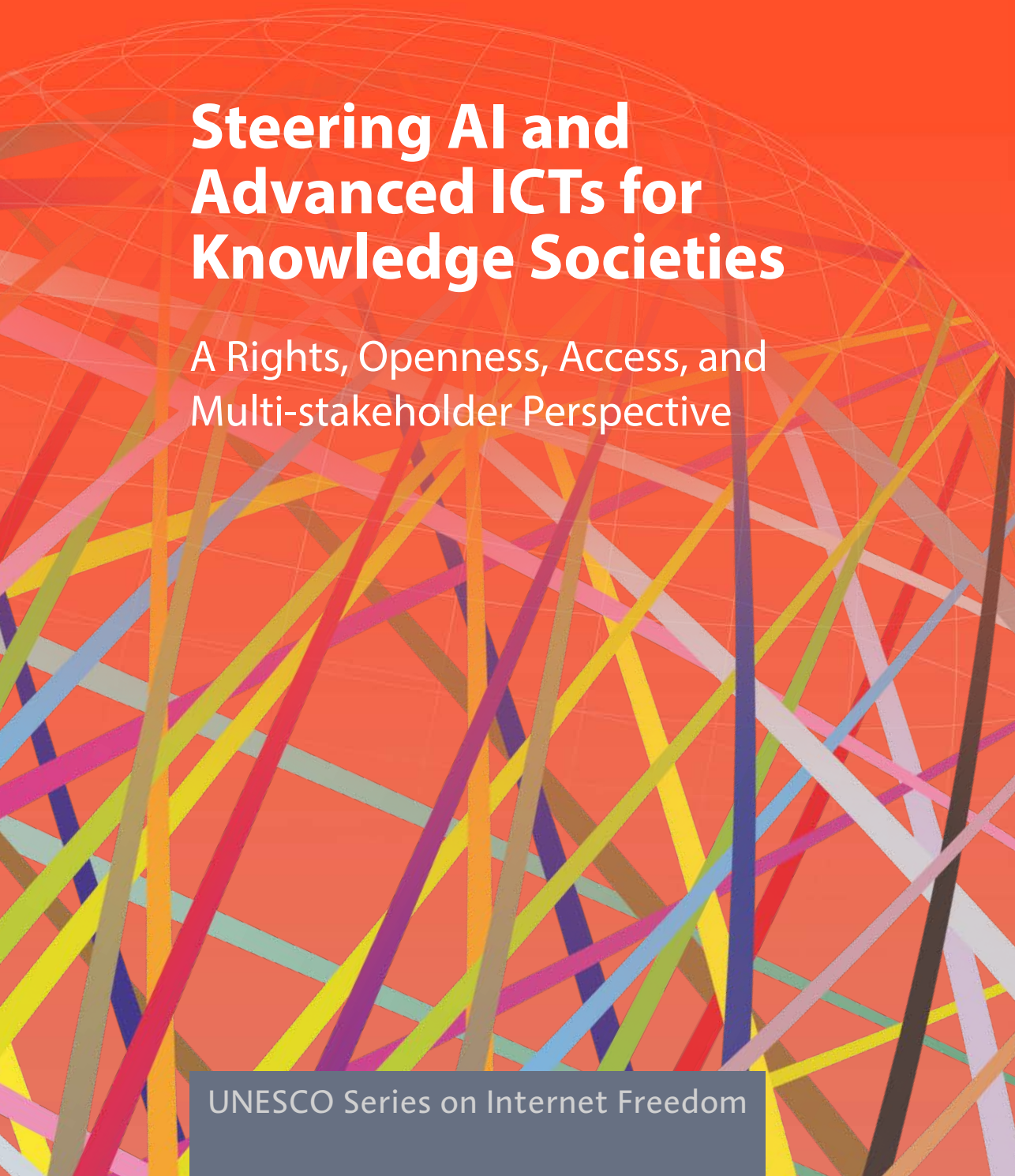
UNESCO Publishing

United Nations
Educational, Scientific and
Cultural Organization

Steering AI and Advanced ICTs for Knowledge Societies

A Rights, Openness, Access, and Multi-stakeholder Perspective



UNESCO Series on Internet Freedom

Published in 2019 by the United Nations Educational, Scientific and Cultural Organization
7 place de Fontenoy, 75352 Paris 07 SP, France

ISBN 978-92-3-100363-9

This publication is authored by:
Xianhong Hu, Bhanu Neupane, Lucia Flores Echaiz, Prateek Sibal, and Macarena Rivera Lam.

The full report and summary leaflet are available at:
https://en.unesco.org/unesco-series-on-internet-freedom

Graphic design, cover design, and typeset: Josselyn Guillarmou
Chapters illustrations: © Josselyn Guillarmou
Cover illustration: © Shuttershock/greiss design
Copy editor: Yahia Dabbous

Printed by UNESCO
Printed in France

UNESCO Communication and Information Sector

STEERING AI AND ADVANCED ICTS FOR KNOWLEDGE SOCIETIES

A Rights, Openness, Access and Multi-stakeholder Perspective

UNESCO Series on Internet Freedom

CONTENTS

LIST OF FIGURES, BOXES AND TABLES

FIGURES

BOXES

TABLES

PREFACE

Artificial Intelligence – abbreviated as AI – can help to pave the way for new opportunities in terms of sustainable development, including when it comes to UNESCO's fields of competence. Developments in the field already have a direct impact on our work in the areas of education, natural and human sciences, culture, as well as communication and information.

UNESCO has long played a leading role in international standard-setting and cooperation, and our Member States have now recognized the need for us to develop ethical principles for AI. In this way, we can help to ensure that technological development is aligned with a human-centered vision, with human rights being respected and sustainable development being advanced. This is why UNESCO engages in analysis and reflection on AI, including from the basis of our mandate to promote freedom of expression and build inclusive knowledge societies.

Recognizing that there are no simple answers about what the future holds for humanity, this research report is a contribution to the wider debate about the ethics and governance of AI. It is an attempt to 'steer' clear of both technological utopianism, and dystopian thinking. Instead of technological determinism and its implication of inevitability, UNESCO gives attention to the role of human agency and human-centred values in the development of AI and other advanced information and communication technologies (ICTs).

Our starting point is with recognizing AI as an opportunity to achieve the United Nations Sustainable Development Goals (SDGs), and to construct knowledge societies. For UNESCO, these kinds of societies are based upon free expression, access to information, quality education and respect for cultural and linguistic diversity. They represent a vision to which we can aspire, and a beacon as we walk the path in shaping AI's role for humanity.

This mission applies to UNESCO's work around not only AI, but also to developments such as the Internet of Things, blockchain, biometrics and algorithmic decision-making. While we examine opportunities, we also seek to identify and mitigate risks such as those posed by arbitrary and bulk surveillance, profiling and violations of privacy and equality. This is important in assessing the potential impacts of continued digitalization on education, the sciences, culture, and communication and information, as well as on employment, equality and empowerment.

Research is essential if we are to understand how AI and other advanced technologies are being used to influence so much of our everyday lives. It is the foundation upon which we can build the knowledge we need to shape technological evolution and to leverage AI's potential in positive ways.

This study frames its assessment of AI through UNESCO's Internet Universality ROAM framework agreed by our Member States in 2015. It therefore covers how AI and advanced ICTs will impact **Human Rights**, **Openness** and **Access**, and how a **Multi-stakeholder** approach underpins work to address both the challenges and opportunities presented by AI.

The ROAM principles can help elaborate the values needed to orientate ethical and rights-based development and deployment of AI in ways that mitigate risks and achieve the SDGs. In this light, this study also offers a set of options for action that can serve as inspiration for the development of new and ethical policy frameworks and other actions, whether by States in their different fields of work, diverse actors in the private sector, members of academia and the technical community, and civil society. We hope the insights in these pages will help steer AI towards making an important contribution to building inclusive knowledge societies that leave no one behind.

Moez Chakchouk
Assistant Director-General for Communication and Information
UNESCO

ACKNOWLEDGMENTS

We sincerely thank the peer reviewers for their time and insightful suggestions that have helped improve the publication:

- Olubayo Adekanmbi, Chief Transformation Officer, MTN Nigeria
- Alex Comninos, Researcher, Research ICT Africa
- Hugo Cyr, Dean, Faculty of Political Science and Law, Université du Québec à Montréal
- Eileen Donahoe, Executive Director, Global Digital Policy Incubator, Stanford University
- Jaco du Toit, Programme Specialist, UNESCO
- Helani Galpaya, CEO, LIRNEasia
- Damiano Giampaoli, Programme Specialist, UNESCO
- Marcus Goddard, Partner, Intelligence, NETEXPLO Observatory
- Helen Hester, Associate Professor of Media and Communication, University of West London
- Joe Hironaka, Programme Specialist, UNESCO
- Joe F. Khalil, Associate Professor, Northwestern University in Qatar
- Nnenna Nwakanma, Chief Web Advocate. World Wide Web Foundation
- Michael J. Oghia, Advocacy and Engagement Manager, Global Forum for Media Development (GFMD)
- Julia Pohle, Senior Researcher, WZB Berlin Social Science Center,
- Rachel Pollack, Associate Programme Specialist, UNESCO
- Jan Rydzak, Associate Director for Program, Global Digital Policy Incubator, Stanford University
- Moses M. Thiga, Lecturer, Computer Science and IT, Kabarak University
- Tim Unwin, Chairholder, UNESCO Chair in ICT4D, Royal Holloway, University of London

We would like to thank Alexandra Agay, Antoine Martin, Bérénice Chaumont, Maud Barret, Valentin Leblanc, Melissa Tay Ru Jein and Joo Eun Chae for their research assistance.

We would like to thank colleagues Kelly Christine Wong and Yahia Dabbous for their help in compiling the publication.

We would also like to thank other colleagues from the Communication and Information Sector, the Natural Sciences Sector, the Culture Sector, the Education Sector, the Social and Human Sciences Sector, the Division for Gender Equality and the Priority Africa and External Relations Sector who are part of UNESCO's Task Team on AI for their support.

We pay tribute to the late Mr Indrajit Banerjee for his vision and support in co-initiating this research.

GLOSSARY

This glossary provides the broad meanings of key terms as used in this study. Attribution to authors for ideas does not constitute an endorsement of their definitions.

Algorithm

A set of step-by-step instructions for solving a problem (Negnevitsky, 2011)

Algorithmic decision-making

A form of decision-making based on outputs from algorithms (Andersen, 2018).

Anonymization

The process of irreversibly removing personal identifiers, direct and indirect, which may lead to an individual being identified (Article 29 Data Protection Working Party, 2014)

Artificial intelligence (AI)

While there is no one single definition of 'artificial intelligence' (AI), this publication tends to define AI as an ensemble of advanced ICTs that enable "machines capable of imitating certain functionalities of human intelligence, including such features as perception, learning, reasoning, problem solving, language interaction, and even producing creative work" (COMEST, 2019).

Artificial narrow intelligence (ANI)

The ability of machines to resemble human capabilities in narrow domains, with different degrees of technical sophistication and autonomy (ARTICLE 19 & Privacy International, 2018).

Weak AI or artificial narrow intelligence (ANI) is the form of AI that humanity has achieved so far – machines that are capable of performing certain precise tasks autonomously but without consciousness, within a framework defined by humans and following decisions taken by humans alone (UNESCO, 2018d).

Artificial general intelligence (AGI)

The overarching, and as yet unachieved, goal of a system that displays intelligence across multiple domains, with the ability to learn new skills, and which mimic or even surpass human intelligence (ARTICLE 19 & Privacy International, 2018).

Strong AI or AGI thus refers to a machine that has consciousness and is capable of providing human-like response (UNESCO, 2018d).

Artificial neural network (ANN)

An information-processing paradigm inspired by the structure and functions of the human brain. An ANN consists of a number of simple and highly interconnected processors, called neurons, which are analogous to the biological neurons in the brain. The neurons are connected by weighted links that pass signals from one neuron to another. While in a biological neural network, learning involves adjustments to the synapses, ANNs learn through repeated adjustments of the weights. These weights store the rules needed to solve specific problems (Negnevitsky, 2011).

Automated decision-making

A process of making a decision by automated means. It usually involves the use of automated reasoning to aid or replace a decision-making process that would otherwise be performed by humans. It does not necessarily involve the use of AI but will generally involve the collection and processing of data (CoE CHR/Rec(2019)1, 2019).

Bias

An inclination or prejudice for or against a person or group, especially in a way that is considered to be unfair (societal definition); the difference between the estimated—or predicted—value and the true value – in other words, the difference between what a system predicts and what actually happens (statistical definition) (Andersen, 2018).

Big data

Datasets that are too large or complex for traditional data processing software to analyze (Andersen, 2018). Most AI systems rely on the collection, processing and sharing of such big data in order to perform their functions.

Black box

A model that is opaque to its user. Although the model can produce correct results, how these results are produced is unknown. An example of a black box is a neural network. To understand the relationships between inputs and outputs of a black box, sensitivity analysis can be used (Negnevitsky, 2011).

Bots

Software applications that run automated tasks, increasingly powered by machine learning (Andersen, 2018).

Cloud

A metaphor describing network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand (ITU, 2014).

Collingridge dilemma

The problem that when change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult and time consuming (Collingridge, 1980).

Data

Facts, measurements, or observations. Also, a symbolic representation of facts, measurements, or observations (Negnevitsky, 2011).

Database

A collection of structured data (Negnevitsky, 2011).

Data mining

Extraction of information and knowledge from data. Also, the exploration and analysis of large amounts of data in order to discover meaningful patterns and rules. The ultimate goal of data mining is to discover information and knowledge (Negnevitsky, 2011).

Deep learning

This technique enables a machine to independently recognize complex variations. An example is automated scouring and classifying of millions of images picked from the Internet that have not been comprehensively labelled by humans. The result of a combination of learning algorithms and formal neural networks and the use of massive amounts of data, deep learning powers AI (UNESCO, 2018d).

Information and communication technologies (ICTs)

Diverse set of technological tools and resources used to transmit, store, create, share or exchange information. These technological tools and resources include software, computers, the Internet (websites, blogs and emails), live broadcasting technologies (radio, television and webcasting), recorded broadcasting technologies (podcasting, audio and video players and storage devices) and telephony (fixed or mobile, satellite, visio/video-conferencing, etc.) (UNESCO Institute of Statistics, 2019).

Information for All Programme (IFAP)

IFAP is a unique UNESCO intergovernmental programme established in 2001. Through IFAP, member and partner governments pledge to harness the new opportunities of the information age to create equitable societies through better access to information.

Intelligence

The ability to learn and understand, to define problems and to make decisions to solve them. A machine is thought to be intelligent if it can achieve human-level performance in some cognitive task (Negnevitsky, 2011).

Internet intermediaries

The term 'internet intermediaries' commonly refers to a wide, diverse and rapidly evolving range of service providers that facilitate interactions on the internet between natural and legal persons by offering and performing a variety of functions and services. Some connect users to the internet, enable the processing of information and data, or host web-based services, including for user-generated content. Others ag-

gregate information and enable searches; they give access to, host and index content and services designed and/or operated by third parties. Some facilitate the sale of goods and services, including audio-visual services, and enable other commercial transactions, including payments (CoE, 2018).

Internet of Things (IoT)

A global infrastructure that enables advanced services by interconnecting (physical and digital) things based on existing and evolving interoperable information and communication technologies (ITU, 2012).

Knowledge societies

Knowledge societies encompass the ability to identify, produce, process, transform, disseminate and use information to build and apply knowledge for human development. They require an empowering social vision that encompasses plurality, inclusion, solidarity and participation. Four principles that are essential for the development of an equitable Knowledge Society are: i) cultural diversity, ii) equal access to education, iii) universal access to information and iv) freedom of expression (UNESCO, 2005).

Machine learning

An adaptive mechanism that enables computers to learn from experience, learn by example and learn by analogy. Learning capabilities improve the performance of an intelligent system over time. Machine learning is the basis of systems that can adapt their response continuously (Negnevitsky, 2011).

Metadata

Data used to define, contextualize or characterize data (CoE, 2018).

Open data

Databases that are publicly available for anyone to access, use and share.[1]

Personal data

Information relating to an identified or identifiable natural person, directly or indirectly, by reference to one or more elements specific to that person (CoE, 2018).

Personal data processing

Any operation or set of operations performed using automated processes and applied to personal data or sets of data, such as collection, recording, organization, structuring, storage, adaptation or modification, retrieval, consultation, use, communication by transmission, dissemination or any other form of making available, linking or interconnection, limitation, erasure or destruction (CoE, 2018).

---

1 Open Data Institute: https://theodi.org/article/what-is-open-data-and-why-should-we-care/

Profiling

The processing of personal data for the purpose of evaluating certain aspects of a natural person's life, in particular to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements" (CoE, 2018).

Pseudonymization

The processing of personal data in a manner such that the personal data can no longer be attributed to a specific data subject without the use of additional information. Such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (GDPR art. IV, cl. 4, 2016).

ROAM Principles

UNESCO's ROAM principles (Rights, Openness, Accessibility and Multi-stakeholder participation) are a lens through which to assess the inclusiveness of the development and deployment of AI technologies, highlighting the relevance of human rights, as well as the importance of openness, accessibility and multi-stakeholder participation.[2]

ROAM-X Indicators

ROAM-X Indicators for Internet Universality is a research instrument which contains 303 indicators (109 identified as core ones) addressing categories of ROAM (Rights, Openness, Accessibility, Multi-stakeholder) as well as contextual and cross-cutting (X) indicators to address gender equality and the needs of children and young people, economic dimensions, trust and security, as well as legal and ethical aspects of the Internet. National assessments of Internet development may be conducted using these international standards, endorsed for voluntary application by the 31st Council Session of UNESCO's International Programme for the Development of Communication (IPDC) in November 2018.[3]

Supervised learning

A type of machine learning that requires an external teacher, who presents a sequence of training examples to the Artificial Neural Networks. Each example contains the input pattern and the desired output pattern to be generated by the network. The network determines its actual output and compares it with the desired output from the training example. If the output from the network differs from the desired output specified in the training example, the network weights are modified (Negnevitsky, 2011).

---

2 Based on UNESCO's Internet Universality framework as adopted by General Conference resolution 38 C/53 on the Outcome Document 'Connecting the Dots: Option for Future Action' in 2015. https://en.unesco.org/system/files/private_documents/234090e-1_0.pdf

3 Decisions taken by the 31st Council Session of the International Programme for the Development of Communication (IPDC), 21-22 November 2018. https://en.unesco.org/system/files/private_documents/266235eng.pdf

Technological determinism

As per the International Encyclopedia of the Social and Behavioral Sciences, 'technological determinism' is a term used to describe a set of claims made about the relationship between what we generally call 'technology' and 'society.' Two meanings have come into use: (1) an internal, technical logic determines the design of technological artifacts and systems; and (2) the development of technological artifacts and systems determines broad social changes. The two meanings are often conjoined in the claim that an autonomous technology (in both its development and use) shapes social relations (Kline, 2015).

Turing test

A test designed by the computer scientist Alan Turing to determine whether a machine could pass a behaviour test for intelligence. Turing defined the intelligent behaviour of a computer as the ability to achieve human-level performance in cognitive tasks. During the test, a human interrogates two conversational partners, a machine and a human via a neutral medium such as a remote terminal. The computer passes the test if the interrogator cannot distinguish the machine from the human (Negnevitsky, 2011).

UNITWIN/UNESCO Chairs Programme

Launched in 1992, the UNITWIN/UNESCO Chairs Programme, which involves over 700 institutions in 116 countries, promotes international inter-university cooperation and networking to enhance institutional capacities through knowledge sharing and collaborative work. The programme supports the establishment of UNESCO Chairs and UNITWIN Networks in key priority areas related to UNESCO's fields of competence – i.e. in education, the natural and social sciences, culture and communication.

World Summit on the Information Society (WSIS)

The UN General Assembly Resolution 56/183 (21 December 2001) endorsed the holding of the World Summit on the Information Society (WSIS) in two phases. The World Summit on the Information Society (WSIS), held in two phases, Geneva in December 2003 and Tunis in November 2005, was the first opportunity for the international community to assess the potential of new information and communication technologies (ICTs) for achieving internationally-agreed development goals, and to consider the new challenges which they presented. The four WSIS outcome documents-the Geneva Declaration of Principles, Geneva Plan of Action, Tunis Commitment and Tunis Agenda for the Information Society – set out a vision of a people-centered, inclusive and development-oriented Information Society that would enhance the opportunities and quality of life for people worldwide and facilitate sustainable development (CSTD/UNCTAD, 2015).

EXECUTIVE SUMMARY

UNESCO's mandate to build inclusive knowledge societies is centered on its efforts to promote freedom of expression and access to information, alongside quality education and respect for cultural and linguistic diversity. The digital transformation underway in society is touching all spheres of human activity, and it is timely to reflect on the key challenges and opportunities created by digital technologies like artificial intelligence (AI).

The title of this publication is a call for 'Steering AI and Advanced ICTs for Knowledge Societies' from the perspective of human Rights, Openness, Access and Multi-stakeholder governance (the ROAM principles). Such steering should also support gender equality and Africa, the two global priorities of UNESCO. Technological change and advancement is important for sustainable development, yet belief in technological determinism risks neglecting social, economic and other drivers. Instead, the challenge is to harness human agency to shape the trajectory of AI and related information and communication technologies (ICTs).

While there is no single definition of 'artificial intelligence', this publication focuses on what UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) describes as "machines capable of imitating certain functionalities of human intelligence, including such features as perception, learning, reasoning, problem solving, language interaction, and even producing creative work" (COMEST, 2019).

AI and its constitutive elements of data, algorithms, hardware, connectivity and storage exponentially increase the power of ICT. This is a major opportunity for sustainable development, with concomitant risks that also need to be addressed. To steer AI accordingly, we need to recognize the uneven but dynamic distribution of AI power across multiple and dispersed centres within governments, the private sector, the technical community, civil society and other stakeholders worldwide. It is for this reason that multi-stakeholder engagement around AI is vital. This perspective aligns with the approach to ICT governance as per the World Summit on the Information Society (WSIS) principles and processes that are led by the United Nations (UN).

Using the Internet Universality framework and indicators,[1] this publication explores the multiple implications for AI and how the ROAM principles can steer the development and usage of AI from the following dimensions:

1 More information on the Internet Universality Indicators is available at: https://en.unesco.org/themes/internet-universality-indicators

- Human Rights such as freedom of expression, privacy and equality;
- Openness with regard to knowledge, open data, and open and pluralistic markets;
- Inclusive Access regarding research, human resources, data and hardware;
- Multi-stakeholder governance;
- Cross-cutting issues: Gender equality and Africa.

In summary, the study's findings are structured according to these six dimensions. It should be noted that many of the illustrations cited in the text concern experiences in particular areas of the world which are comparatively advanced in terms of AI development and application. This derives from the availability of data and secondary research materials for the study, and it serves to alert other geographical areas as to the kinds of issues that they may come to experience. Therefore, the study is not a global review of AI.

1 HUMAN RIGHTS IMPLICATIONS AND RIGHTS-BY-DESIGN APPROACH FOR AI DEVELOPMENT

As part of the UN family, UNESCO stands for human rights. From algorithms that are designed and used to shape the way our social media news feed is shown, to those profiling users and curating the information they receive and thus affecting voting choices in elections, AI already influences all human rights including those relevant to communication and information both positively and negatively:

Right to freedom of expression

- Online content personalization by AI offers relevant content and connections. But it may manipulate how people use their right to seek information and their right to form an opinion. This could weaken the pluralism of ideas and the degree of exposure to verified information.
- AI is being used to de-emphasize or remove online content that is discriminatory or that incites hatred and violence. However, automated content moderation also risks blocking legitimate free expression, and levels of transparency and the existing channels for redress appeal are inadequate.

Right to privacy

- Privacy is infringed when AI involves opaque data collection, de-anonymization, unauthorized third-party data sharing, and the tracking and profiling of individuals. However, AI could also help monitor violations and abuses of personal privacy.
- Data protection based on consent and transparency is vital in AI, but the availability of this protection is uneven across the world, and it does not deal with the full scope of privacy concerns.

Journalism and media development (as exercises of right to expression)

- AI can be used to strengthen journalism in its operations of gathering, verifying, analyzing and distributing information.
- However, AI is also being used with the side-effect of weakening the institutions of journalism and reducing their diversity by facilitating the migration of advertising to data-rich Internet intermediaries. Elements of AI also play a role in many digital attacks on journalists, their devices and their websites.
- AI can be used to disseminate false content deliberately fabricated with harmful intent and it is used to overshadow journalistic content by amplifying such disinformation. However, AI could also be used to help identify covert co-ordination of online campaigns.

Right to equality and right to participation in public life

- It remains a challenge to eliminate bias in automated decision-making systems, which poses risks for the equal enjoyment of human rights by women and children, as well as minorities, indigenous groups, persons with disabilities, groups facing discrimination based on their gender identity and gender expression, and economically disadvantaged people. However, value-explicit and well-trained systems based on unbiased data may diminish the risk of human biases in certain decisions.
- Participation in public life has been enabled by ICT, but at the same time AI systems have been used to profile and manipulate voters' access to information and to influence their voting choices, through behavioural manipulation and spread of disinformation by means such as micro-targeting of users.

2 OPENNESS AND AI

UNESCO advocates for open access to scientific research, open data, open educational resources and open science to ensure equal access and opportunities. This is in order to strengthen universal access to information, to bridge information inequalities and to promote transparency. Openness in AI raises challenges and opportunities.

Explainability and transparency for the 'Black-Box' problem of AI

- The 'Black-Box' problem of AI systems, understood as the opacity in how AI systems make decisions raises concerns regarding transparency and accountability in automated decision-making. Several solutions, both technical and operational, concerning transparency in the use of automated decision-making and generating explanations for why the decisions have been taken have been proposed to address the 'black-box' problem. But these can come at odds with intellectual property issues.
- Norms of disclosure and transparency are useful for clarifying the intended purpose algorithms but are insufficient to resolve the opacity problem of AI. However, AI may also be harnessed to explain, at least in part, its own workings, and its results can be audited.

Open data

- As opposed to proprietary data sets, open data repositories play an important role in reducing entry barriers for inclusive development of AI.
- Openly available data, however, raises concerns with respect to privacy because of potential de-anonymization of individuals through triangulation based on different public data sets.

Open markets

- Open and pluralistic markets are a way to foster innovation in AI development and for efficient allocation of resources.
- At the same time, in order to gain a larger market share, firms may choose practices not in conformity with the UN Guiding Principles on Business and Human Rights, and thereby depart from the ethical practices necessary for the safe and beneficial use of AI.

3 ACCESS AND AI

The ability for everyone to access and contribute information, ideas and knowledge is essential for inclusive knowledge societies. Access to information and knowledge can be promoted by increasing awareness of the possibilities offered by AI among all stakeholders. These possibilities include development of free and open-source software in order to improve skills, co-operation and competition, access by users, diversity of choice, and to enable all users to develop solutions that best meet their requirements.

Access to research

- There is a strong increase in the number of research publications on AI and associated technologies. However, there is significant divide with respect to the quality, and research output varies across countries. Inequalities in access to AI research are growing between both countries and research institutions.
- Where they exist, national policies and international support for AI related research help in strengthening the research output in developing countries and provide a base for local innovation to grow on.

Access to knowledge, education and human resources

- Access to education and training for development and implementation of AI remains limited in many countries. There is a need for strengthening capacities and infrastructure within institutions providing AI education and training.
- Leading research and development centres attract global talent, often resulting in brain drain from some countries, as well as brain drain from academia to the private sector. Efforts to increase availability of human resources include local initiatives to upgrade skills of existing employees, to crowd-source solutions, thereby leveraging a wider knowledge pool to solve problems, and to offer AI service platforms without costly investment in infrastructure and human resources.
- People need to understand their own engagement with AI in order for the technologies to be accessible to all. Media and Information Literacy concerning AI and other digital technologies is far from universal, but will be needed in order to empower and inform people.

Access to data

- Technology firms and state actors access large amounts of user data and use this data to train algorithms, but this unequal access to data creates entry barriers for new entrants, including start-up firms. Academic institutions and research centres face challenges in accessing high quality data available to private sector firms.

- Data Commons based upon open data repositories can enable the training of algorithms that may strengthen access to data for inclusive development of AI.

Access to connectivity and hardware

- Development of AI depends on the availability of broadband, cloud storage and specialized computational hardware that can run algorithms on processors designed to perform large quantities of calculations.
- Emerging cloud-based solutions combined with affordable and universal broadband connections reduce the need for large overheads or fixed-cost investments for smaller AI developers and users.

4 MULTI-STAKEHOLDER APPROACH FOR AI GOVERNANCE

All stakeholders – from governments, the private sector, the technical community, intergovernmental organizations (IGOs), civil society, academia, to individual users – are increasingly impacted by AI and have a common interest in defining how AI is governed.

Effective multi-stakeholder processes are:

- Inclusive
- Diverse
- Collaborative
- Transparent
- Equal
- Well-informed
- Flexible and relevant
- Safe and private
- Accountable and legitimate
- Responsive
- Timely

Fora for AI multi-stakeholder discussions include participatory legislative and regulatory policy debates; national, regional and international AI cooperation frameworks; and technology company consultations to develop terms of service and operating procedures.

AI AND GENDER EQUALITY

Gender equality is important to ensure that all people, without discrimination based on sex, gender and sexual orientation, enjoy the right to access, participate and contribute to society. Building on the rich literature on the relationship between gender and technology, this chapter recognizes that AI-powered technologies may both set back and push forward the struggle for gender equality. AI brings challenges such as lack of representativeness of the AI workforce, algorithmic discrimination, subservient 'female' voice assistants and sex robots, and 'deepfake' pornography, which may perpetuate negative gender stereotypes and disadvantage women and LGBTI individuals. Simultaneously, members of the AI community have begun to devise remedies to these challenges, some of which make use of AI and associated technologies.

AI AND AFRICA

The importance of science, technology and innovation is well recognized by African countries and forms an essential part of the African Union's vision 2063. However, there are significant capacity, infrastructure and governance challenges in building a strong enabling environment for AI development. Increasing numbers of African governments are cognizant of these challenges and are taking initiatives, some through AI specific policies to empower the private sector, researchers, and civil society to harness AI for development. The speed and scale of the initiatives to date are limited. Nevertheless, many actors within the private sector, the technical community and civil society are actively trying to address the immediate challenges of access to knowledge, skills, mentorship and business opportunities.

INTRODUCTION

1. CONCEPTUALIZATION OF AI IN THIS PUBLICATION

While there is no one single definition of 'artificial intelligence' (AI), this publication focuses on the combination of technologies that enable what UNESCO's COMEST calls "machines capable of imitating certain functionalities of human intelligence, including such features as perception, learning, reasoning, problem solving, language interaction, and even producing creative work" (COMEST, 2019). This understanding is closer to the narrow scope of AI so-called 'artificial narrow intelligence' (ANI) which means the ability of machines to resemble human capabilities in narrow domains, with different degrees of technical sophistication and autonomy (ARTICLE 19 & Privacy International, 2018). For simplicity, at times the publication also uses the term AI and automated/algorithmic decision-making interchangeably, while keeping in mind however that AI proper should be understood as a wider complex of technologies, relations and practices that include deep-learning.

We are often oblivious to our data footprints and the existence of algorithms around us in greater or lesser degrees around the world. Meanwhile, AI and/or its elements are increasingly deployed, as the following examples demonstrate:

- Search engine algorithms help us access the information that we want by rapidly interrogating data on the World Wide Web, and the search results tend to be more and more personalized based on a user's location, gender, language, search history, and other data trails.
- Job-matching algorithms analyze people's competencies to show employers suitable candidates for employment;
- On-demand video platforms provide tailored movie suggestions based on our viewing patterns and on those of millions of other users, and offer advertisers the ability to predict and nudge our attitudes and actions;
- Algorithms help judges determine the possibility of recidivism and suggest duration of prison sentences;
- Credit-risk algorithms decide who should be offered a loan and on what terms and conditions;
- Digital profiles are used by immigration authorities to approve or reject visa applications.

In some quarters, the phrase 'Fourth Industrial Revolution' is used to describe the significance of AI and other advanced technologies. However, the term is contested,[1] and it is not part of the vocabulary of this study, which itself ranges far beyond the issues of the relationship between technology and economic factors.

1 For example, UNESCO Chair in ICT for Development, Professor Tim Unwin, argues that the concept is technology-determinist and gender-biased (Unwin, 2019).

2. WHY IS UNESCO INTERESTED IN AI?

Given AI's widespread application in UNESCO's fields of work including education, the sciences, culture, access to information, freedom of expression and ethics, UNESCO has a significant role to play in these changing times. The Director-General of UNESCO, Audrey Azoulay, has highlighted that, "humanity is on the threshold of a new era" and that the "transformation has already begun" (Azoulay, 2018).

UNESCO also has a key role in fostering multi-stakeholder mechanisms to protect human rights, ensure openness to knowledge and research, and reduce inequalities within and between its 193 Member States to promote inclusion. Technological divides, including in AI, can have a multiplier effect on social inequalities and complicate the sustainable development aspiration to leave no one behind.

UNESCO's Communication and Information (CI) Sector advances freedom of expression, as per the Organization's constitutional mandate to promote "the free flow of ideas by word and image." To carry out this work, UNESCO's CI Sector adopts two main lines of action as follows:

- Fostering freedom of expression online and offline, promoting the safety of journalists, advancing diversity and participation in media, and supporting independent media.
- Building knowledge societies through ICTs by enabling universal access to, and preservation of, information and knowledge.

For UNESCO, knowledge societies are predicated on the pillars of freedom of expression, access to information, education and cultural and linguistic diversity. They promote knowledge by leveraging ICTs with the goal of improving access to education, scientific knowledge and innovation, and empowering local and marginalized communities. UNESCO works within the wider UN family to advance Knowledge Societies that are capable of meeting the challenge to achieve the universal 2030 Agenda for Sustainable Development.

Box 1: World Summit on Information Society (WSIS) Mandate of UNESCO

UNESCO's approach to AI builds upon insights from the World Summit on the Information Society (WSIS) and its follow-up, including the Organization's advocacy for a human rights-based and human-centered approach to ICTs during the WSIS events in Geneva (2003) and Tunis (2005). Also relevant is the WSIS follow-up process wherein UNESCO is the UN agency responsible for lead/facilitating implementation of the Action Lines on Access to information and knowledge (C3); E-Learning (C7); E-Science (C7); Cultural diversity and identity, linguistic diversity and local content (C8); Media (C9); and Ethical dimensions of the Information Society (C10).

As set out in the WSIS Geneva Declaration of Principles (2003), "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."

The 2015 WSIS Review reaffirmed the common desire to "build a people-centered, inclusive and development-oriented Information Society, where everyone can create, access, utilize and share information and knowledge" and thereby enabling individuals to achieve their full potential. The WSIS+10 outcome document also emphasizes the importance of the principles of human rights and multi-stakeholder cooperation in building inclusive Knowledge Societies.

These discussions are relevant in the context of AI and they require firm commitment from all stakeholders to ensure that AI's development takes place in a manner that respects human rights, openness, access to information and multi-stakeholder participation.

3. UNESCO'S ONGOING REFLECTION ON AI

In order to give substance to the global dialogue on how to leverage the potential of AI for the achievement of the SDGs, UNESCO organized a series of meetings in 2018 and 2019. These include the panel discussion on 'Responding to Opportunities and Challenges of the Digital Age', during the UNESCO Partners Forum; a roundtable on 'Artificial Intelligence: Reflection on its complexity and impact on our society'; the debate on 'AI for Human Rights and SDGs: Fostering Multi-stakeholder, Inclusive and Open Approaches', held as part of the Internet Governance Forum; an open discus-

sion on 'Harnessing Artificial Intelligence to Foster Knowledge Societies and Good Governance', held at the Mozilla Foundation; and a panel on Philosophical Reflection on AI during World Philosophy Day.

UNESCO's ongoing reflection on the global discussion of ethical AI, focusing on norms and standards, was also demonstrated at the Forum on AI in Africa, which took place in December 2018 in Benguerir, Morocco. At the conclusion of the forum, the participants unanimously adopted the 'Benguerir Statement', agreeing on the need to promote an AI strategy for Africa, as well as human-centered AI. UNESCO's overall AI strategy was presented to Member States at an Information Meeting held in January 2019, which was followed by an international experts' debate on 'Tech Futures: Hope or Fear?'. This reflection culminated in the March 2019 global conference 'Principles for Artificial Intelligence: Towards a Humanistic Approach?' held at UNESCO Headquarters in Paris.

UNESCO also hosted a session on Harnessing Artificial Intelligence to Strengthen Journalism and Media Development in line with UNESCO's Internet Universality ROAM principles (see below) within its facilitation process of WSIS Action Line C9 Media at WSIS Forum 2019.[1] A study by UNESCO's COMEST programme, to which the CI Sector contributed, was presented to the Executive Board of UNESCO in the first half of 2019.

4. UNESCO ADDRESSES AI AND ADVANCED ICTS THROUGH THE ROAM PERSPECTIVE

While well recognizing AI as a complex of elements that are not synonymous with the Internet, this publication locates AI's inextricable application and development within the ecosystem of this network of networks and the way its social, political and economic context has evolved.

A UN system-wide strategic approach and road map for supporting capacity development on artificial intelligence set out by the United Nations Chief Executives Board for Coordination (CEB), emphasized in its first regular session of 2019 that:

> "Artificial intelligence should be addressed in an ambitious and holistic manner, promoting the use of artificial intelligence as a tool in the implementation of the Goals, while also addressing emerging ethical and human rights, decent work, technical and socioeconomic challenges."
>
> (CEB/2019/1/Add.3, 2019)

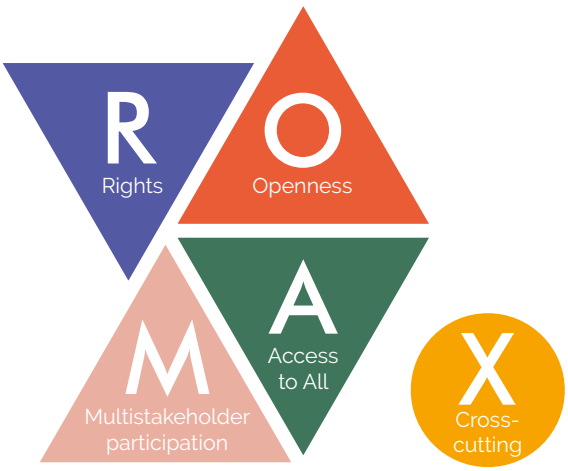
1 The link to the event is available at: https://www.itu.int/net4/wsis/forum/2019/Agenda/ViewSession/304#

This is why, as noted in the report of the UN Secretary-General's High Level Panel on Digital Cooperation (2019), "UNESCO has used its 'Rights, Openness, Access and Multi-stakeholder governance' (ROAM) framework to discuss AI's implications for rights including freedom of expression, privacy, equality and participation in public life."

The ROAM framework is the underpinning of UNESCO's Internet Universality concept which was endorsed by the Organization's 38th General Conference in 2015.[2] These principles are human Rights, Openness, Accessibility and Multi-stakeholder participation, and they emerge from the Member States' mandated Keystones to foster inclusive knowledge societies study by UNESCO, which covered privacy, freedom of expression, access and ethics of a global internet (2015a). In the 'CONNECTing the Dots Outcome Document', also endorsed at UNESCO's 38th General Conference, UNESCO committed to promoting human rights-based ethical reflection, research and public dialogue on the implications of new and emerging technologies and their potential societal impacts.

Building on these developments, UNESCO engaged in a three-year process of elaborating the Internet Universality framework through a global, open, inclusive and multi-stakeholder process.

The result is the Internet Universality ROAM-X indicators framework, which was welcomed by Member States in UNESCO's International Programme for the Development of Communication (IPDC) in November 2018 (UNESCO, 2019d; IPDC Council CI/2018/COUNCIL.31/H/1). The 303 indicators are organized in five categories – four of which reflect the four ROAM principles. The fifth category covers cross-cutting issues like gender equality and the needs of children and young people, economic dimensions, trust and security, and legal and ethical aspects of the Internet.



2 The link to the UNESCO General Conference resolution is available at: https://en.unesco.org/system/files/private_documents/234090e-1_0.pdf

The ROAM-based approach thus serves as a well-grounded and holistic framework for UNESCO and other stakeholders to contribute to the design, application and governance of AI. The framework can enrich activities for the setting of normative and ethical principles for AI, and for the production of innovative policy guidelines and toolkits.

This is relevant to the mandate of UNESCO's 206th Executive Board (3 April –17 April 2019, Paris) and the 40th Session of General Conference[3] (12 November-27 November 2019, Paris), which recognizes that a recommendation could be an essential tool to strengthen the elaboration and implementation of national and international legislation, policies and strategies in the field, as well as to enhance international cooperation on the ethical development and use of AI in support of the Sustainable Development Goals (SDGs). The 40th GC also decided that it is timely and relevant for UNESCO to prepare an international standard-setting instrument on the ethics of artificial intelligence (AI) in the form of a recommendation and the draft text of a recommendation on the ethics of artificial intelligence should be submitted for consideration by the General Conference at its 41st session (UNESCO, 2019b) (206 EX/42). This aligns with UNESCO's role in providing technical advice to Member States and other actors, serving as a clearinghouse for innovation, and building capacity. In this manner, AI ethics as informed by the ROAM principles can help contribute to the benefit of humanity, sustainable development and peace.

From the vantage point of all the above, this study is presented as follows:

The chapter on **Rights** assesses how AI is being used to shape content personalization and how this affects freedom of expression and freedom of opinion. It also examines AI-enhanced content moderation and its threat to freedom of expression. The section on privacy then explores how current approaches to privacy are challenged by AI. It expands the discussion from privacy laws that govern interaction between citizens and legal entities to norms of privacy that govern algorithmically-mediated interactions between people. From data journalism to automated disinformation, the chapter also explores the issues at the intersection of AI and journalism and media development. The last section on the right to equality unpacks the word 'discrimination', often used in the context of assessing algorithmic decision-making, differentiating between direct, indirect, institutional, intentional and unintentional forms of discrimination.

The chapter on **Openness** introduces the broader context of openness within AI and presents an analysis of its drivers. It explains, inter alia, limitations in addressing the 'black box' issue and transparency in algorithmic decision-making processes. As a note of caution on openness and its potential risks, the chapter invites a wider reflection on the misuse of AI for potential harm. In the context of the open data movement, it presents different modes of data collection and challenges associated with

3 The link to the Draft Resolution of the 40th GC: *Preliminary study on a possible standard-setting instrument on the ethics of artificial intelligence* endorsed 21 November 2019 is available at: https://unesdoc.unesco.org/ark:/48223/pf0000369455/PDF/369455eng.pdf.multi

the availability of data for machine learning. The chapter discusses the role of open and pluralistic markets in the diffusion of innovation in AI while flagging the possible neglect of human rights in the context of commercial competition.

The chapter on **Access** identifies key elements such as access to algorithms and to research data, and the human and computational resources available for the development and application of AI. The discussion on each element maps the current state of affairs and flags issues that warrant attention if we are to mitigate new digital divides.

The chapter on **Multi-stakeholder** governance of AI presents the historical evolution of UNESCO's work and tools in supporting multi-stakeholder dialogue—from governments, companies, the technical community, IGOs, civil society and academia. These groups are increasingly impacted by AI. The chapter also shares the related values, practices and indicators needed to foster and operationalize a multi-stakeholder approach in AI governance.

The chapter on **Gender Equality** and AI builds on the rich literature on the relationship between gender and technology. It recognizes that AI-powered technologies may both hinder and advance the struggle for gender equality.

The chapter on **Africa** highlights that there is recognition of the wider importance of science, technology and innovation (STI) in African countries as a path to growth and development. However, there are significant capacity, infrastructure and governance challenges in building a strong enabling environment for AI development. Within the framework for STI Strategy of the African Union, the chapter describes some of the initiatives taken by governments, the private sector, the technical community and civil society for furthering the development of AI in Africa.

HUMAN RIGHTS AND AI

1

CHAPTER 1: HUMAN RIGHTS AND AI

Artificial Intelligence and related automated / algorithmic decision-making processes, are becoming more and more embedded in the tissue of connected societies, racing ahead of an underdeveloped clear understanding of the consequences for human rights. The design, creation and use of AI and related technologies presents opportunities to enhance access to the rights enshrined in the Universal Declaration of Human Rights (UDHR), to build inclusive Knowledge Societies, and to achieve the SDGs. However, as with other scientific and technological developments, the current use and future evolution of AI could also have negative consequences for fundamental rights and freedoms, and these should be countered or mitigated (Access Now, 2018).

This chapter is not intended to be a comprehensive study of all of AI's potential impacts on human rights, but is to address some of the main concerns regarding how the use of AI technologies can impede upon human rights within the scope of UNESCO's mandate. As the technology progresses, we are bound to discover new benefits and risks to human rights protection and enjoyment.

Box 2: UNESCO's Position on Human Rights for Internet Universality

By identifying the Internet's connection to human rights-based norms as constituents of freedom, 'Internet Universality' helps to emphasize continued harmony between the growth and use of the Internet and human rights. A free Internet in this sense means one that respects and enables the freedom to exercise human rights. In this regard, 'Internet Universality' encourages us to consider the range of interdependencies and inter-relationships between the Internet and different human rights – such as rights to freedom of expression, privacy, diversity of cultural expressions, public participation and association, gender equality, security and education. AI should be considered in the holistic context of the Internet and human rights.

AI and human rights are related to UNESCO's Internet Universality principles and indicators, which in turn are referenced in a 2018 UN Human Rights Council Resolution (A/HRC/38/L.10/Rev.1) on the promotion, protection and enjoyment of human rights on the Internet, which highlights in particular online freedom of expression and privacy.

This report further explores the potential of AI for bolstering independent journalism and promoting a free, pluralistic and independent media. This aligns with the definition of freedom of expression guaranteed by Article 19 of the UDHR which entails that everyone has the right to freedom of opinion and expression, which includes access to information and ideas through any media. AI and related technology is already shaping aspects of the news, with impact on notions of the value of journalists, the practice of journalism and the production of other kinds of content.

It is not only important for all stakeholders to reflect upon the challenges, but also to formulate responses that protect freedom of expression, privacy, journalism and the media, as well as the rights to equality and political participation. This chapter will explore a set of options to maximize the benefits and minimize the human rights risks posed by AI.

Box 3: UNESCO and Freedom of Expression

UNESCO is the UN specialized agency with the mandate to defend freedom of expression, as mandated by its Constitution to promote "the free flow of ideas by word and image." UNESCO recognizes that the right to privacy underpins other rights and freedoms, including freedom of expression, association and belief. Further, the Organization recognizes freedom of expression as a key pillar for building knowledge societies.

1. RIGHT TO FREEDOM OF EXPRESSION

Freedom of expression both online and offline plays a key role for knowledge societies. The ability to express one's views in the public sphere is an essential component for enabling people to participate in public debates. Having access to a means of expression is "a necessary condition for participation in the political process of the country" (Scanlon, 1972). Equally, freedom of expression is important as a form of personal expression for the speaker, which is part of individual self-realization (Gilmore, 2011). Thus, in addition to enabling social and political participation, freedom of expression is also a crucial means of self-fulfillment (Cannataci, et al., 2016).

Freedom of expression and freedom of opinion are closely linked. The UDHR affirms that everyone has the "right to freedom of opinion and expression" and, similarly, Article 19 of the International Covenant on Civil and Political Rights (ICCPR) includes the "right to hold opinions" and the "right to freedom of expression" as distinct but adjacent rights. The UN's Human Rights Committee clarifies Article 19 of the ICCPR in its General Comment No. 34 (2011), stating that freedom of opinion and freedom of

expression are both indispensable conditions for the full development of the person and are both foundation stones of every free society. The right to freedom of expression in the UDHR and the ICCPR is affirmed in complementary directions, to seek and to receive information and ideas, and to impart information and ideas (de Zayas & Martín, 2012). Overall, individuals have a recognized right to exchange ideas, to inform themselves and to form and develop personal opinions.

In general, ICTs have the potential to "enable a worldwide public to seek, receive and impart information and ideas and other content in particular to acquire knowledge, engage in debate and participate in democratic processes" (CoE CM/Rec(2012)3, 2012). Indeed, the Internet facilitates a communication structure where every user has the ability to seek information and to assert their voice. This differentiates the Internet from the "one-to-many unidirectional structure of traditional mass media" (Hansen, 2018). AI and advanced ICTs can help to foster freedom of expression and also pose challenges to different dimensions of this fundamental right.

For example, Internet search engines, supported by AI algorithms, are crucial gatekeepers for people wishing to seek, receive and impart information (MSI-NET, 2016). With their ranking algorithms improved by AI, search engines are useful in providing links to information that would have otherwise been unknown and/or inaccessible. Yet, gatekeeping by search engines as well as social media platforms can never be completely neutral. In many cases, algorithms give visibility to disinformation, hate speech and so on, which in turn may affect people's right to form their views independently (Solon & Levin, How Google's search algorithm spreads false information with a rightwing bias, 2016). As for many issues related to AI and human rights, stakeholders must find the right balance in order to benefit from AI while protecting the different dimensions of freedom of expression.

1.1. Personalized information, freedom of opinion and the right to seek information

While violations of the right to impart information are widely discussed (i.e. censorship – see 1.2 below), the dimension of seeking and receiving information and ideas should not be de-emphasized. The right to receive information is an essential component for exercising the right of freedom of opinion, as enshrined in Article 19 of the UDHR. UN Special Rapporteur David Kaye has noted that a collateral of the right to hold an opinion is the right to form one and this "requires freedom from undue coercion in the development of an individual's beliefs, ideologies, reactions and positions" (UNGA A/73/348, 2018). While international standards of the right to seek information permit certain restrictions, freedom of opinion cannot be restricted.

AI is used to affect how people access information online. Machine learning algorithms in search engines are designed to personalize the content that is shown to the user. The way social media feeds are arranged is also dictated by the use of these algorithmic predictions (Flaxman, Goel, & Rao, 2016). These deployed algorithms decide what people see and in what order. Combining the users' browsing history, geolocation, "user demographic [and] semantic and sentiment analyses and numerous

other factors," these algorithmic models are put into service to customize the information that is given priority (UNGA A/73/348, 2018).

This personalized experience has some benefits, as it can bring forward relevant information tailored to an individual's needs (e.g. an advertisement for an item the user was thinking of buying). Content can also be made available in the individual's primary language.

However, this automated shaping of the type of information to which people have access through the use of AI and/or its elements is never purely technical, and its impact can be to distort our ability to know about a range of information and opinion, or what editorial or business values underpin what gets prioritized for them (Andersen, 2018). As stated earlier, freedom of expression and opinion is what enables social and political participation and act as means of self-fulfillment. But both of these aspects can be undermined by the personalization of content. In turn, this use of technology can lead to fragmentation of the public sphere as well as potential undercutting of individual agency and conscious self-development.

Such negative potential impacts have been argued to be the result of filter bubbles (Pariser, 2011) and echo chambers, two notions related to content personalization. Filter bubbles refer to how algorithmic predictions of user preferences limit the scope of information available to an individual (Flaxman, Goel, & Rao, 2016). Algorithms are set up to predict the type of content that will be interesting to a user and show this content. By doing so, they exclude other social and political content, thereby undermining an individual's ability to find certain kinds of information and opinions. In this way, two different users can make an identical request on the same search engine and receive different results, bringing them distinct information. Such personalization also occurs in social media news feeds, determining the order and visibility of posts. In this sense, the characteristics of a user's data characteristics and past engagement have now begun to dictate the limits of their worldview (Bezemek, 2018).

Filter bubbles are closely linked to the notion of echo chambers. By defining the scope of information available in part based on the user's past use of the Internet, AI-driven algorithmic techniques can reinforce "their prior political views due to selective exposure to political content" (Colleoni, Rozza, & Arvidsson, 2014). The Internet experience can therefore become an echo chamber where political orientation is reaffirmed. The information encountered appears to legitimize existing views and opinions, presenting them as facts, thereby creating an environment where users encounter only information that confirms existing views (Sibal, 2016). This personalization of content can also be used for 'nudging' people towards extremes, and even contributing to the process of radicalization towards violent extremism (Tschan & Bekkoenova, 2018)

Bill Gates, founder of Microsoft Corporation and philanthropist, has expressed concern that filter bubbles have become a larger problem than many, including himself, would have expected, since it "lets you go off with like-minded people, so you're not mixing and sharing and understanding other points of view" (Delaney, 2017). In this sense, filter bubbles and echo chambers accentuate the fragmentation of public

sphere (MSI-NET, 2018).[1] The risk has been summed up in terms of assessing the development of algorithms to

> "[…] co-govern or co-determine what can be found on the Internet […] is seen and found (search, filtering, and aggregation applications), is produced (content production applications like algorithmic journalism), is considered relevant (search and scoring applications; ranking), is anticipated (prognosis/forecast applications), and is chosen and/or consumed (recommendation, scoring, and allocation applications; both for economic and social choices – ranging from commercial goods to friends and partners)", […] influencing the behavior of individual producers and users […]."
>
> (Just & Latzer, 2017)

Although algorithmic decision-making in content being accessed does not automatically translate into determining people's views, it is important to recognize the part they can play in shaping the public agenda and other elements of the information ecosphere (Davies, 2018). AI personalization can threaten freedom of thought in as much as it determines the type of information to which people have access (Andersen, 2018). Thus, the Council of Europe has warned that the use of algorithmic processes and machine learning may influence people's emotions and thoughts, sometimes subliminally (CoE Decl(13/02/2019)1, 2019).

Users are not always offered alternatives to AI personalization and algorithmic targeting, such as where they can consciously choose to see information prioritized by date, credible source, or other priorities they may have, as distinct from other computationally embedded logics. Agenda setting is not done by news media in the public sphere with a human editorial approach, but by technologies with hidden logics that are often designed for the business objective of gathering personal data and selling access to users and their data. What this means ultimately is that algorithms in content personalization affect the opportunities available to us, thereby also limiting the scope of possibilities that define us (Rouvroy, 2014).

1.2. Online content moderation by AI and the right to impart information

Internet platforms in many countries today represent a central place where public discussions are held (Latonero, 2018). A significant proportion of speech and expression takes place online (Balkin, 2017). Therefore, it comes as little surprise that the Internet is a space where, amongst other expressions, disinformation, hate speech, and propaganda for violence and war is also delivered. In recent years, Internet platforms have been relying on AI to moderate content posted online. Through practices like

1 MSI-NET document Algorithms and human rights - Study on the human rights dimensions of automated data processing techniques and possible regulatory implications (2018): https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html

"spam detection, hash-matching technology, keyword filters, natural language processing and other detection algorithms" (UNGA A/73/348, 2018), social media companies and other Internet companies can remove or 'downrank' content perceived as 'undesirable'.

Article 19 of the ICCPR guarantees the right to impart ideas of all kinds through any media. However, it also states that restrictions may be applied when they conform to the law and are necessary to respect the rights and reputations of others, in addition to the protection of national security, public order, and public health or morals. Article 20 of the ICCPR further describes conditions where freedom of expression is expected to be curtailed in accordance with law. These conditions include propaganda for war and advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence (UNGA Resolution 2200A (XXI), 1966).

It is in this light that the following sections discuss the advantages of using AI for online content moderation, the processes followed by Internet intermediaries for content moderation and the associated challenges.

AI algorithms that filter and flag violent, discriminative, or hate-inducing forms of expression respond to a legitimate goal, since it is an obligation to prevent the spread of such messages. It has been posited that with less terrorist propaganda and less hate speech online, there might be fewer people turning to violent extremism (Citron, 2017), although such online content is also unlikely to be a single or even primary driver (Alava, Frau-Meigs, & Ghayda, 2017).

AI can help respond to the scale of content, supplementing human efforts. A further potential benefit of using AI for content moderation over human moderators concerns the considerable "psychological toll that the frontline work of content review and moderation takes" as human reviewers are "exposed to the very worst of humanity day in and day out"—from child abuse content to gruesome acts of violence (Raso, Hilligoss, Krishnamurthy, Bavitz, & Kim, 2018). AI and related technologies help to reduce the burden of seeing this type of content.[2]

These reasons render the use of AI justifiable in content moderation online. However, the use of such narrow technical measures must respect international limitations regarding the possible restrictions on freedom of expression. Many observers fear that AI is currently not at a level of development where it can differentiate between news reporting, advocacy, and satire on the one hand, and on the other, the actual incitement of harm. Further challenges arise from the role of Internet companies in choosing to use AI in content moderation, as elaborated in Section 1.4 below, and there should be mechanisms in place in case of AI overreach that violates legitimate expression.

2 For more information on the work done by content moderators, the documentary *The Cleaners* offers an insightful perspective.

1.3. Internet platforms under pressure to regulate online content

From social media firms to search engines, for-profit companies and non-profits like Wikipedia, a wide range of Internet platforms face pressure from national governments and regional authorities to regulate content posted on their platforms (Citron, 2017). Some governments require them to monitor and remove content without waiting on law-based requests from national authorities. Terrorist and extremist content, hate speech and 'fake news' are labelled as objectionable. However, definitions are often weak. According to Special Rapporteur David Kaye, broadly worded restrictive laws on 'extremism', blasphemy, defamation, 'offensive' speech, 'false news' and 'propaganda' often serve as pretexts for demanding that companies suppress legitimate discourse (UNGA A/HRC/38/35, 2018).

The pressure on Internet intermediaries "ranges from direct regulation, to threats, to suggestions that things will go better for infrastructure operators if they cooperate, to negotiations over the terms of cooperation" (Balkin, 2017). The European Council explicitly expects Internet platforms "to develop new technology and tools to improve the automatic detection and removal of content that incites to terrorist acts." (European Council CO EUR 8 CONCL 3, 2017). Some political leaders urge automated prevention of content being uploaded in the first place (Hope & McCann, 2017). Indeed, these types of demands on Internet platforms have been a factor in driving the use of automated systems for content moderation (Andersen, 2018).

Self-regulation of private companies along with use of AI does not necessarily comply with international standards of freedom of expression (UNGA A/HRC/38/35, 2018). Instead, it is largely based on their Terms of Service and Community Guidelines.

There is a concern "that governments may violate free expression rights by strong-arming platforms to remove offensive but legal speech [while avoiding] democratic process and accountability" (Keller, 2018). Indeed, privatized censorship means that it is difficult to appeal content removal requests in a court of law, unlike cases of direct governmental censorship. Private entities are generally legally entitled to make their own decisions, and while they should respect human rights as per the UN Guiding Principles on Business and Human Rights (the 'Ruggie Principles'), they do not have the same extent of obligations as States to operate according to universal standards (UNHRC A/HRC/17/31, 2011). These dynamics are further complicated when AI is brought into play in content moderation issues.

1.4. AI detecting and removing content: cases and challenges

Within the first six months of 2017, AI identified 95 per cent of the 300,000 accounts removed from Twitter's platform for violations related to the promotion of terrorism (Dialani, 2019), using national and international terrorism designations to assess what constitutes as terrorism (Twitter Help Center, 2019). Facebook CEO Mark Zuckerberg explained in a post in December 2018 that 99 per cent of terrorist-related content

on the platform was taken down by AI systems, prior to any user complaints or reports from law enforcement agencies (Zuckerberg, 2018). Facebook's definition of 'terrorism', however, has been met with concerns for being too broad by UN Special Rapporteur on protecting human rights while countering terrorism, Ní Aoláin (UN OHCHR, 2018b). At the same time, videos by journalists and advocacy groups that were documenting alleged war crimes were also removed when YouTube put in place an AI system for identifying and removing what it considers to be 'terrorist propaganda' (Citron & Jurecic, 2018). There is widespread agreement with the Christchurch Call for use of technology to prevent video footage being livestreamed by terrorists, and to remove or block any uploads of such content, the bulk of contested content exists in a greyer zone.

The issue here is the likelihood that AI is programmed to, or serves to, over-block legitimate content in order to protect the companies. Furthermore, at present, AI tools are not accurate enough, and have higher error rates than humans (Raso, Hilligoss, Krishnamurthy, Bavitz, & Kim, 2018). Thus, content considered lawful and also in line with online platforms' particular community standards is being removed.

The nature of development and deployment of AI tools also make the risk to freedom of expression even greater. This is because the particular human biases means we are still far away from developing holistic and dynamically-updated datasets that reflect the complexity of tone, context, and sentiment of the diverse cultures and subcultures (ARTICLE 19, 2018).

Thus, AI is at present unable to satisfactorily detect contextual considerations essential to determining the nature of the content. These considerations encompass irony, satire, culture-related aspects or the intention of the user, tone, audience, artistic purpose and so on (UNGA A/73/348, 2018). Currently, natural language processing cannot differentiate between a sarcastic rebuttal of hate speech from an actual hateful comment (Citron & Jurecic, 2018). While the Office of the United Nations High Commissioner for Human Rights' (OHCHR) Rabat Plan of Action provides nuance in assessing when hateful speech becomes dangerous, this kind of assessment is hard to program into AI (UNGA A/HRC/22/17/Add.4, 2013). The same hurdles are faced by the automation of assessment of proportionality and legitimate purpose that determine whether and how certain content merits being subjected to restriction.

This issue also affects content that is not incitement of potential harm. A specific example concerning nudity occurred in 2017 when Facebook removed a picture of a 500-year-old statue. While it remains unknown if this removal was the result of AI or a human employee or both, it seems highly likely it was an automated removal based on Facebook's regulation concerning nudity (Shah, 2017). The fact that the responsible party is unknown is indicative of the need for more explanation from such companies. (See Section 1.5 below).

Box 4: Facebook Banning of Pulitzer Prize-winning 'Napalm Girl' Photograph

Facebook decided to remove the publication of the iconic photograph of nine-year-old Phan Thi Kim Phúc running naked in the aftermath of a napalm attack during the Vietnam War in September 2016. Facebook stated that the picture violated community standards because of the display of nude genitalia that could qualify as child abuse images. After receiving criticism, Facebook reversed its decision and recognized the global importance of the photograph. While this was not a case of complete autonomous automated removal, "the photo was tagged for removal by one of Facebook's algorithms, which was then followed up by a human editor" as it is often the case (Scott & Isaac, 2016). In principle, technology-monitored moderation can have a role in flagging content, with a human empowered to decide what steps to take. In this case, human error was at fault.

The issue becomes even more complex when AI tools disproportionately affect the freedom of expression of vulnerable and minority groups, as well as people with disabilities, an area that requires further study (UNGA A/73/348, 2018). Natural language processing algorithms have been shown to have higher error rates towards marginalized groups and speakers (Duarte, Llanso, & Loup, 2017). This algorithmic discrimination against certain perspectives "favors the powerful over the marginalized" (Raso, Hilligoss, Krishnamurthy, Bavitz, & Kim, 2018).

Indeed, automated tools have been "shown to perform less accurately when analyzing the language of female speakers and African American speakers" (Citron & Jurecic, 2018). Besides restricting these individuals' rights to express themselves, AI tools can thus limit and even remove content that is relevant and necessary to public debate.

Moreover, natural language processing does not yet perform as well in other languages as it does in English. Automated tools are thus less accurate in evaluating non-English speakers, which can disproportionately restrict their speech. Language translation tools in particular can have this problem. An example of this is the arrest of a Palestinian man by Israeli forces. The man wrote "good morning" in Arabic on a photograph of himself that he posted on Facebook but the company's AI-powered translation tool had translated this into "attack them" in Hebrew or "hurt them" in English (Duarte, Llanso, & Loup, 2017; Hern, 2017). Facebook recognized that its translation system made an error and apologized to the man and his family for the disruption it caused (Fussell, 2017).

As a Center for Democracy and Technology (CDT) report concludes, today's tools for automating social media content analysis have limited ability to parse the nuanced meaning of human communication, or to detect the intent or motivation of the spea-

ker. Policymakers must understand these limitations before endorsing or adopting automated content analysis tools. Without proper safeguards, these tools can facilitate overbroad censorship and biased enforcement of laws and of platforms' terms of service (Center for Democracy & Technology, 2017).

1.5. Lack of transparency and due process in content moderation

The role of algorithms responsible for the personalization of information as well as those involved in online content moderation is largely opaque. There is a challenge for Internet companies to provide transparency in these processes in order to be held accountable.

Regarding the matter of content detection and removal, the Internet Policy Observatory conducted a research project that culminated in the Santa Clara Principles on Transparency and Accountability of Content Moderation Practices. One of the three principles concerns the notice that companies should give to each user when their content has been removed or their account suspended. To fully explain the rationale behind the removal or suspension, companies should provide the affected party with detailed information about which specific guidelines were violated and how the content was detected and removed (Santa Clara Principles, 2018). For example, was the removed content detected by an algorithm and then removed by an employee? Was it flagged by a government authority and then removed? Alternatively, was it identified and autonomously removed by application of an algorithm?

Another Santa Clara principle is that companies should publish regular reports containing the number of posts removed or accounts suspended due to violations of content guidelines (Santa Clara Principles, 2018). The principle does not require companies to provide the source code or technical information about their algorithms; indeed, it is unlikely that companies will be required to make their actual source code publicly available, given their intellectual property claims and interests in secrecy both for economic interests and as a way to inhibit actors from 'gaming' the systems. Other forms of enhancing algorithmic transparency should therefore be explored, including through qualified transparency, consumer choice and education (Pollack, 2016). Technical and other type of limits to transparency are discussed in the Openness chapter of this book.

The restriction of freedom of expression is a serious issue and the problem is that algorithmic censorship online does not currently respect due process standards. There is concern about the impact on presumption of innocence, the avoidance of pre-publication censorship, the right to be informed promptly of the cause and nature of an accusation, the right to a fair hearing and the right to defend oneself (MSI-NET, 2018). Indeed, users currently do not have the opportunity to defend themselves and are not given a meaningful chance to challenge decisions to block or 'downrank' their content (Balkin, 2017; Citron & Jurecic, 2018).

Box 5: The Santa Clara Principles

In May 2018, a group of organizations, advocates and academic experts proposed the Santa Clara Principles as initial steps to be followed by companies and platforms engaged in content moderation, in order to ensure the fair enforcement of content guidelines (The Royal Society, 2018). The three principles are:

1. Numbers: Companies should publish the number of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.
2. Notice: Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.
3. Appeal: Companies should provide a meaningful opportunity for a timely appeal of any content removal or account suspension.

Facebook has recently announced the creation of an independent council to review content moderation decisions (Hensel, 2018). However, questions remain about the independence of the members and the process by which they will pick cases to hear (Newton, 2019). The idea of industry-wide social media councils has also been raised by the UN Special Rapporteur on the protection and promotion of the right to freedom of opinion and expression, and elaborated by non-governmental organizations (NGOs) like ARTICLE 19 (UNGA A/HRC/38/35, 2018; ARTICLE 19, 2019). These could go some way toward improving accountability for content moderation, and provide greater transparency and opportunity for redress.

2. RIGHT TO PRIVACY AND PERSONAL DATA PROTECTION

Privacy is a fundamental human right and a value of great importance as it fosters self-determination and permits us each to develop our personal perspectives of the world around us (Cohen, 2012). It is also important since it is regarded as the enabler (or a prerequisite for the exercise) of other human rights and freedoms, such as the rights to freedom of expression and opinion, freedom of assembly and association, political participation and freedom of thought, belief and religion. An individual requires a private space to fully enjoy and practice these rights. However, the delimitations on what is considered private and the legal expectations of privacy are not always clear (Baghai, 2012).

The right to privacy is enshrined in Article 12 of the UDHR and Article 17 of the ICCPR, as well as other human rights documents, international instruments and national laws.

It entails that no one should be subjected to arbitrary or unlawful interference with their privacy and that everyone has the right to the protection from the law when and if faced with such interference or attacks.

It is equally important to note that privacy is not limited to private spaces. Privacy extends to public spaces and encompasses information that is publicly available (UNHRC A/HRC/39/29, 2018).

In 1988, the Human Rights Committee explained in General Comment No.16 that this right protects against all interferences, whether they be from the State or private actors. It is explained that surveillance, including wire-tapping and interception of communications, violates privacy. It also addressed the notion of personal data by stressing that all personal information on computers or data banks must be regulated by law. Everyone should have information regarding what personal data is stored or processed, for what purposes and which entities control or might control it (UNHRC HRI/GEN/1/Rev.9 (Vol. I)).

In the last few years, more attention has been given to how privacy standards have evolved in the digital age. In December 2013, the UN General Assembly affirmed the need for States to review their practices and legislation regarding mass surveillance and collection of personal data, among others, in order to uphold the right to privacy (UNGA A/HRC/22/17/Add.4).

A 2018 report on privacy in the digital age by the High Commissioner for Human Rights defined privacy as:

> "[…] the presumption that individuals should have an area of autonomous development, interaction and liberty, a 'private sphere' with or without interaction with others, free from State intervention and from excessive unsolicited intervention by other uninvited individuals."
>
> (UNHRC A/HRC/39/29, 2018)

The UN Special Rapporteur on Privacy, Joe Cannataci, in his 2019 report to the 40th session of the Human Rights Council, pointed out that the right to privacy is not an absolute right but a qualified right, and governed by the standard of necessity in a democratic society (UNHRC A/HRC/40/63, 2019). It may be limited but always in a very carefully delimited way. According to the standard established in ICCPR's Article 17, interferences with the right to privacy are only permissible under international human rights law if they are neither arbitrary nor unlawful.

Box 6: Options for UNESCO related to Privacy (UNESCO's CONNECTING the DOTs Outcome Document, 2015)

- Affirm that the fundamental human rights to freedom of opinion and expression, and its corollary of press freedom and the right of access to information, and the right to peaceful assembly, and the right to privacy, are enablers of the post-2015 development agenda;
- Reaffirm that the right to privacy applies and should be respected online and offline in accordance with Article 12 of the UDHR and Article 17 of the ICCPR
- Support as relevant within UNESCO's mandate, the efforts related to UN General Assembly Resolution A/RES/69/166 on the Right to Privacy in the Digital Age;
- Support initiatives that promote peoples' awareness of the right to privacy online and the understanding of the evolving ways in which governments and commercial enterprises collect, use, store and share information, as well as the ways in which digital security tools can be used to protect users' privacy rights;
- Support efforts to protect personal data which provide users with security, respect for their rights, and redress mechanisms, and which strengthen trust in new digital services.

Privacy also encompasses informational privacy covering data that can be derived about a person and their life (UNHRC A/HRC/39/29, 2018). This means that data protection is a vital part of the right to privacy even without being the totality of it. Privacy must also be perceived as a "breathing room to engage in the processes of boundary management that enable and constitute self-development" (Cohen, 2012).

AI is not developed in a vacuum, but in the affordances of its context and particularly the character of the Internet and the forces shaping its evolution. AI systems are based on algorithms built for reliance upon the collection, storage and processing of large amounts of data in order to learn and make intelligent decisions. Therefore, AI development cannot be separated from the data collection processes, and indeed the data collected made available.

Closely linked to personal data transactions, AI is subject to various data protection regulations in many jurisdictions. The Council of Europe's updated Convention 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data became open for signature in October 2018, and the Consultative Committee of the

Convention has published Guidelines on Artificial Intelligence and Data Protection.[3] Similarly, the EU's General Data Protection Regulation (GDPR), which has been in force since 25 May 2018, addressing automated profiling and decision-making, which is clearly linked to the use of AI. Both the Council of Europe convention and the EU regulation focus on addressing the new realities of online world and digital technologies (CoE ETS No.108, 1981; European Commission, 2018). However, the UN Special Rapporteur for the promotion and protection of the right to freedom of opinion and expression, David Kaye has noted that in the context of AI systems, "the ability of individuals to know, understand and exercise control over how their data are used is deprived of practical meaning" (UNGA A/73/348, 2018).

The expanding mechanisms through which information about ourselves and about the world is collected and processed present a risk to our right to privacy and a particular concern is the use of AI to construct profiles about individuals and to de-anonymize data sets (see 2.2. below). This profiling linked to predictive analytics also represents a threat to privacy. Some argue that privacy is currently the right most affected by AI applications (Raso, Hilligoss, Krishnamurthy, Bavitz, & Kim, 2018). This reinforces the need for a discussion around privacy norms in the digital era. This discussion must be part of a broad and inclusive debate about the desirable future direction for society (Mikkinen, Auffermann, & Heinonen, 2017).

2.1. Data collection and the Internet of Things

The use of AI and associated technologies can generate new pools of information or metadata (data about data) (Rouvroy, 2016). Underlying all this is the fact that the 'big data era' has brought "increased capabilities to amass and store data" (Vayana & Tasioulas, 2016) shifting from traditional data collection. The retention of data is now not inherently limited to a specific given purpose. In fact, when data is being collected, a purpose is not necessarily already set. The "usefulness of each data item depends on the quantity of the other data with it may be correlated" (Rouvroy, 2016). In this sense, all data, however innocuous and meaningless when considered individually, can be of additional significance and interest to many actors.

Instead of a situated and precise collection of data, networked information technologies provide "continuous, pervasively distributed, and persistent" surveillance attention (Cohen, 2012). This attention encompasses many things such as the monitoring of our footprints online or cameras with facial recognition in the public space, as well as gait recognition, a subject area which requires further research regarding its impacts on privacy.

Through the intensive use of the Internet and the increasing use of a number of Internet of Things (IoT) devices and applications, individuals are generating a vast amount of data (ARTICLE 19 & Privacy International, 2018). This can be done intentionally by writing posts, using emojis or posting pictures on social media, or unintentionally,

3 https://www.coe.int/en/web/data-protection/-/new-guidelines-on-artificial-intelligence-and-personal-data-protection

by browsing websites, clicking on links, accepting cookies, etc. These are our digital footprints, collected by default to monitor our online movements, where monitoring is often the deliberate "tracking of individuals online to create profiles" (Rouvroy, 2016; Bennett, 2018).

The line between the online and offline world is increasingly blurred. Indeed, "people seem to live in a continuum of on/offline, with the result that it is difficult to draw sharp and meaningful lines between the two" (Vayana & Tasioulas, 2016). For example, when we move around the city and go to a coffee shop, a school, or a medical institution, the GPS tracker on our smartphones is able to detect where we are and how long we stay and collect this data (and correlate it with the movements of others), even if we did not access the Internet on our phones. Meaningful inferences can be derived regarding our identity, interests, aspirations, problems and networks from such data.

IoT further blurs the line between offline and online, since devices that used to be only physical and non-Internet-related are now increasingly integrated into wider data connections, including links to AI development and processing. Through sensors and software, these devices "emit information on the movements, activities, performance, energy consumption, lifestyles etc. of their users" and these data are gathered, stored, analyzed and sold (Rouvroy, 2016). Such data becomes increasingly valuable when processed by AI, and therefore increasingly challenge privacy in the digital age.

2.2. AI powered surveillance, monitoring and facial recognition

User consent to the use, processing, storage, transfer and dissemination of their personal data, especially when it refers to sensitive areas of an individual's life (such as health, sexual orientation, personally identifiable financial information, biometric data, etc.), is a major regulatory issue. Data sets may reveal personal or sensitive details of a person's digital and everyday life when aggregated by AI systems (Andersen, 2018). Furthermore, once the personal information collected by these means is out in the open, it can be difficult for the affected person/s to seek its correction or its removal.

Virtual assistants embedded in smart speakers, such as Amazon Alexa, commonly shortened to 'Alexa', are installed in a growing number of private homes, and these often collect data without the knowledge or fully-informed consent of those whose data are being collected. Like other voice assistants, there are also significant assumptions embedded in the service which impact particularly on notions of gender roles and gender equality (UNESCO; EQUALS Skills Coalition, 2019).

A service like Alexa can listen to everything people around it say, but only starts recording when it hears the 'wake word' such as its name. Indeed, "[o]nce the word is detected, audio begins streaming to the cloud, including a fraction of a second of audio before the wake word" (Amazon, 2019). The audio recorded then becomes part of the data stored by Amazon and can be used for many purposes. Not only is Amazon interested in the words pronounced by the consumer, but also in other aspects of the recording. For instance, Amazon has filed a patent in which it wishes to detect a

change in the customer voice or a sneeze and a cough in order to suggest medicines (Mehta, 2018). While this is still only in the patent stage, it is something that could be interpreted as helpful but as also being intrusive and infringing on people's right to privacy.

An informational asymmetry exists between the users of these consumer products and those who process the data. It is of great importance for societies to educate consumers and raise their awareness about the data that their connected devices, networks and platforms generate, process or share and on how these actions could potentially affect their right to privacy, as well as other human rights and freedoms (ARTICLE 19 & Privacy International, 2018). It is also vital that companies inform the public of the potential flaws of the devices. This could facilitate the Council of Europe recommendation for states to mitigate the potentially adverse impacts of AI on human rights (CoE CHR/Rec(2019)1, 2019).

Box 7: Virtual Assistants Eavesdropping: Amazon Alexa

A couple was having a private conversation in their house when they received a text message from a colleague that read, "Unplug your Alexa devices right now. You're being hacked." This colleague had received audio of their private conversation. Alexa had been listening in, recording their background conversation and then sending it to this person on their contact list. The device, however, was not hacked by a third party. Amazon confirmed that the audio had been unintentionally broadcast by the device (Moye, 2018). The voice assistant started recording when it detected a word in the couple's conversation that it interpreted for the 'wake word' and understood a part of the background conversation as a 'send message' request (Wollerton & Crist, 2018). Even if this happened unintentionally, the couple experienced it as a privacy invasion.

Furthermore, some data concerning the behaviour of individuals are being collected by other means, such as satellite imagery or video surveillance in public spaces. In particular, AI-powered facial recognition software has been increasingly used by governments and companies in public spaces such as stations, schools, theatres, streets, shopping malls and so on. This potentially violates individuals' privacy and "transforms expectations of anonymity in public sphere, which is particularly relevant to vulnerable groups and to those who speak out against powerful actors involved in human rights abuses, corruption, to name a few" (ARTICLE 19 & Privacy International, 2018). Under the threat of permanent surveillance and the loss of anonymity, individuals may be deterred from exercising their fundamental human rights and prefer to alter their behaviour in public spaces (Andersen, 2018). This may also be particularly

reinforced when persons fear that the data may end up in the hands of actors with real power to cause them harm.

Box 8: Facial Recognition Software in Shopping Malls

In 2018, several media outlets reported the use of facial recognition software in shopping malls in Calgary, Canada, raising concerns about the potential violation of the right to privacy of local shoppers. This AI-driven surveillance was used to collect different types of data from the behavior of customers, in order to analyze and identify patterns in shopper behavior. Without asking for the customers' explicit consent, the software could track shoppers' ages and genders, which would allegedly allow the mall owners to "understand directory usage patterns (and) to create a better shopper experience" (Rieger, 2018).

Besides their role in data collection and processing, AI systems are being used by private corporations and governments alike for surveillance purposes. Even within countries, people's confidence that their data will remain secure varies depending on whether they are accessed at the hands of governments or private sector companies. According to a 2015 report of the Pew Research Center, 31 per cent of adult Americans are 'very confident' or 'somewhat confident' that government agencies will keep their records private and secure, while only 11 per cent believe so for social media sites (Madden & Raine, 2015). Across different countries, the deficit in trust may vary (McMullan, 2015).

An argument in favour of facial recognition software powered by AI systems is that it can be used for law enforcement purposes to identify and locate specific individuals. While mass surveillance is widely regarded as a disproportionate interference with the right to privacy and free expression (as it is neither "necessary nor proportionate to the goal of public safety or crime prevention"), targeted surveillance needs to comply with the three-part test of legality, necessity and proportionality, as well as legitimate purpose as established in international human rights law (Andersen, 2018).

2.3. Online tracking and de-anonymization of individuals

Data anonymization has historically been the way in which "the balance between using data and preserving people's privacy has relied both practically and legally" (Montjoye, Farzanehfar, Hendrickx, & Rocher, 2017). Ubiquitous computing and big data are however challenging anonymization. For instance, a 32 bit code: 4c812db292272995e5416a323e79bd37, helped an online activity tracking program to iden-

tify, the user as a '26-year-old female in Nashville, Tennessee' with interests in movies including 'The Princess Bride', '50 First Dates', '10 Things I Hate About You' and 'Sex and the City' (Angwin, 2010). Some of these predictions can be accurate to the point of de-anonymizing web users whose online activities are constantly tracked. Once DNA enters the system, it is almost impossible to remain anonymous when this data is combined with other data.

Box 9: Online Tracking

For a firsthand experience of online tracking, we encourage the reader to go online to Google's Ads Preference manager at: http://www.google.com/ads/preferences/ and look at markers used by the company to define the reader and assess how accurate these are.

The information tracked is used to create digital profiles of users to which access is sold in the market place, including specialized exchanges, to help advertisers market their products better. For instance, in one case a person who works in the construction business was defined very accurately as a male between the age of 35-54 years, as a homeowner living in a small town with no kids, having a college degree and a median income of $86,724, and working in management (Steel & Angwin, 2010). The granularity of the information along with the location of the person effectively revealed the identity of the person being tracked.

De-anonymization and re-identification is enabled by AI's ability to recognize patterns and identify trends out of non-personal data about individuals or groups of people, and to thus derive the intimate from the available without the knowledge or the consent of the people concerned by such inference (ARTICLE 19 & Privacy International, 2018). However, this new information loses the context in which the original data were first extracted and the purposes for which data providers could have initially consented to processing, thus increasing the risk of the data being inaccurate and depriving individuals of the ability to rectify or delete the data (UNGA A/73/348, 2018). This newly generated data, which could reveal a specific individual's sexuality, political views, overall health status and religious beliefs, could result in discrimination and even persecution in certain instances.

The possibility to infer or predict personal or sensitive information out of non-personal data pulled out from different datasets effectively blurs the distinction between personal and non-personal data, posing significant challenges to the right to privacy (ARTICLE 19 & Privacy International, 2018).

2.4. AI profiling and predictive analytics

Consumer and user data feeds sophisticated AI systems of predictive analytics (Cohen, 2012). By sorting, assessing, scoring, classifying, evaluating and ranking different individuals among different groups of people, AI is used to try to predict future behaviour (ARTICLE 19 & Privacy International, 2018; Harcourt, 2007).

As aforementioned, privacy can be seen as a breathing room free "from unreasonable constraints on the construction of one's own identity" (Agre & Rotenberg, 1988). By profiling every user and predicting their preferences, these AI systems lend themselves to those who seek to reinforce or nudge user preferences and consequent behaviour "in ways that reduce the serendipity and the freedom to tinker" on which innovation thrives (Cohen, 2012). Autonomy, as well as privacy, can be approached "as something that is achieved within complex material and social preconditions" rather than something that is purely given (Oleksy, Just, & Zapedowska-Kling, 2012). Hence, profiling individuals and making decisions based on AI predictions, both of which can greatly affect individuals, may be interpreted as interference with a genuine private sphere, which is well needed for self-development.

Private corporations often use these predictions for personalization (as seen in the section on the right to freedom of expression), but they also, like governments, may use this information to determine or limit citizens' access to services and programs.

Some see this type of predictive analytics as a paradigm shift where knowledge seems to be derived directly from reality: "we feel that with big data we no longer have to produce knowledge about the world, but that we can discover knowledge directly in the world" (Rouvroy & Stiegler, 2016). However, it is important to realize that "prediction does not merely describe the future, it transforms it" (Rouvroy, 2016). Indeed, by limiting the scope of possibilities, self-determination is threatened and privacy diminished. It could therefore be said that this use of AI's objective "is to produce tractable, predictable citizen-consumers whose preferred modes of self-determination play out along the predictable and profit-generating trajectories" (Cohen, 2012).

Box 10: 'Chilling' impact on other Human Rights

Since data collection is everywhere and predictive analytics feed decisions that greatly influence individuals, people may modify their behaviours in order to try to avoid suffering negative consequences. It is possible that they may restrain themselves from interacting and sharing information with each other, which amounts to an unjustifiable constraint on the right to freedom of opinion and expression (ARTICLE 19 & Privacy International, 2018).

The pervasive and invisible nature of AI systems, coupled with their ability to identify and track behaviour, can also have a chilling effect on other human rights such as freedom of assembly and association (UNHRC A/HRC/26/29, 2014). Individuals who wish to participate in social movements may avoid doing so for fear of being politically profiled by facial recognition software or other AI systems.

The same can be said about people refraining from communicating "sensitive health-related information for fear that his or her anonymity may be compromised" which in turn can affect their right to health (UNHRC A/HRC/26/29, 2014).

AI privacy concerns may have similarly negative impacts on other rights, such as freedom of religion, the right to desirable work, right to a fair hearing and freedom from arbitrary arrest and detention (Andersen, 2018; Raso, Hilligoss, Krishnamurthy, Bavitz, & Kim, 2018).

2.5. Protecting privacy in the AI context: updated paradigms and Privacy Enhancing Technologies (PETs) solutions

Existing norms of privacy centered on information control are being challenged in the digital age. The dichotomies between the public and the private have become blurred to the extent that it is difficult or even impossible for individuals to control their information. Therefore, there is a need for frameworks that better guide our understanding of privacy in the age of algorithms and big data.

Among others, Nissenbaum has developed the concept of privacy as contextual integrity to help understand these challenges. This provides a universal account of privacy that does not depend on place or time, meaning it is not supposed to operate within a preconceived dichotomy of public versus private or sensitive versus non-sensitive (Nissenbaum, 2004).

Nissenbaum notes that intuitions about privacy norms seem to be rooted in details of rather more limited contexts, spheres, or stereotypic situations. Every interaction has its own context specific norms of privacy. For instance, a patient-doctor relationship involves the sharing of medical information with a mutual understanding of strict confidentiality about the information being shared.

Since it involves the sharing of sensitive medical information, the same norms that concern patient-doctor confidentiality should apply to the relation between person A and an AI system serving as a 'doctor'. Therefore, third party sharing of sensitive medical information from wearable tech would be violation of privacy in this context. Further, the use of the data by the medical platform itself should be subject to the

norms applicable to information exchange between a patient and a doctor. A fraught question, however, is how to extend such norms to apply elsewhere to the numerous layers of other actors in the data chain who are able to capture and record this information.

Similarly, friends share a broad range of detail about their lives with each other with an expectation of a type of privacy, which should still apply when the exchange is intermediated by technology. Likewise, norms of privacy at the voting booth, when consulting a lawyer, at the bank, or in a bus are all context specific, an expectation that should be considered and respected, despite that 'collapse of context' that occurs online and particularly through social media (Boyd, 2008).

While norms of privacy that govern people's real lives should be equally applicable to people's digital lives, there is a false separation between the two which has enabled the proliferation of simple and overly-broad models of consent to practices that interfere with a person's privacy. The idea is that a range of dispensations should apply to the range of social and communication arrangements. In this regard, AI may have potential to identify these variations to assist with deciding on appropriate privacy protection within each particular context.

While AI raises various privacy concerns, it also presents great opportunities to enhance personal privacy. AI-based privacy enhancing technologies (PETs) such as 'differential privacy' and 'federated learning' are improvements to previous methods to prevent re-identification of individuals through aggregating personal data (Scripa Els, 2017). For example, the federated learning process recently developed by Google allows the collection of data to improve the centralized machine learning model without storing individual data in the cloud. Indeed, "instead of sending up raw data, it determines the changes that should be made to the model locally and then sends a 'small focused update' to the cloud, where the update is averaged with other updates to improve the model" (Scripa Els, 2017).

AI can also be used to monitor violations and abuses of personal privacy. Since AI systems will become more and more complex, humans alone might not be able to adequately monitor its violations. The idea of AI auditors is to have intelligent systems to guard other AI applications and detect cases where personal data are not well managed or cases in which an AI program achieves re-identification (Scripa Els, 2017).

However, much needs to be done to raise awareness and literacy around privacy concerns among all stakeholders and to encourage more actors to incorporate privacy protection by design and privacy enhancing technologies and applications.

3. JOURNALISM AND MEDIA DEVELOPMENT AS EXERCISES OF FREEDOM OF EXPRESSION

The practice of journalism is a particular exercise of the right to freedom of expression and access to information, and press freedom is a necessary liberty for disseminating information into the public sphere. What distinguishes the practice of journalism from other forms of public expression—forms which also rely on press freedom—is that authentic journalistic expression conforms to professional standards such as verification of content and publication in the public interest, enabling individuals and societies to receive and impart information and ideas, in accordance with Article 19 of the UDHR.

While journalistic expression can be done on a purely individual basis, a supporting infrastructure in the form of an institution, with its distinct policies and systems, is usually required for sustainable output and to defend practitioners against attack. This is where media organizations, ranging from private through to public and community media, have a key role to play in any society's engagement with how AI impacts freedom of expression.

There are growing intersection of AI and the practice of journalism and protection of journalists, as well as the news and other kinds of content. This not only affects production, but also the dissemination and consumption of journalism.

Journalists and media platforms need to be empowered to use AI to tell the story about this technological momentum, and to do AI-enhanced journalism to recognize patterns and trends that are otherwise invisible.

3.1. The era of computational journalism, data journalism, automated journalism and robotic journalism

Computational journalism, data journalism, automated journalism and robotic journalism are part of the terminology used somewhat interchangeably to identify the use of advanced ICTs in one or more phases of the journalistic process. The adoption of automatic algorithms can be used for gathering data, verifying facts and automated writing or video-editing of news and disseminating media content. AI has, in this sense, a lot to offer to journalism and media development.

This new type of journalism can be described by the algorithmic processes "that convert data into narrative news texts with limited to no human intervention beyond the initial programming" (Carlson, 2014). Essentially, AI in the journalistic process can be divided in two facets: the computational processing of big data that can extract relevant information and the algorithmic process that can convert this knowledge into readable stories (Latar, 2015). Both of these aspects can be seen as being complementary to human journalists' work (Flew, Spurgeon, & Daniel, 2012).

AI can help to generate news quickly thanks to progress in natural language generation (NLG), which is a subfield of natural language processing (Dörr, 2016). This can be especially helpful in areas where constant updates are needed (such as stock price changes or sports reporting). For example, *The Washington Post*'s AI 'robot reporter', called Heliograf, issued 300 short reports and alerts on the Rio Olympics (Moses, 2017). This automation can free journalists' time to pursue other tasks that are less mechanical (Carlson, 2014).

AI also presents translation possibilities for gathering and disseminating news in many languages, which will help in reaching a broader audience and new markets (Dörr, 2016). The Finnish News Agency (STT) is already using AI to translate news into English and Swedish (George, 2018).

AI will also provide media with greater capacity to serve advertisers, get subscriptions and measure the quality of journalism. Elements of the recent report from IREX, *'Can Machine Learning Help Us Measure the Trustworthiness of News?'* highlights some of this potential (IREX, 2018).

Box 11: AI Helps with Measuring the Quality of Journalism

IREX measures the quality of journalism by using some 20 indicators. It has done this with tens of thousands of news articles through projects in a number of countries. For years, such work relied on human evaluators, which involves trained media professionals. In 2018, IREX tested whether machine learning could make this process more efficient and consistent. They took one of the 20 indicators—whether a journalist non-transparently inserts their own opinion into a news article—and worked with an AI startup to train algorithms to identify instances across thousands of articles. After just a few rounds of training, AI was able to identify sentences in news articles that contained the writers' opinions, at one point reaching 95 per cent accuracy.

However, the deeper and more complex issue of identifying underlying narratives, which work to frame and structure the selection of sources, the ordering of information, tone and wording used, etc., is likely to be far more challenging for AI to process. (IREX, 2018)

The impact of AI on journalism also poses different challenges in terms of media deontology, including by raising the question of who or what should be considered or accepted as having authorship of algorithmic news (Montal & Reich, 2017). This question has an impact on accountability for the content, for example in terms of liability in case of a defamation suit. Additionally, since media outlets do not currently always clearly identify if an article was developed by humans or by algorithms, transparency

questions emerge. Information asymmetries arise if the audience cannot distinguish if journalistic content is created by humans or produced by AI, or is a fusion of the two. This raises the question: "If there is no information on the algorithmic nature of a text and its resources, how should the audience make a decision whether it wants to consume the information and whether it can rely on it?" (Dörr & Hollnbuchner, 2017).

3.2. AI threatening media pluralism: Content prioritization, information personalization and micro-targeting

The use of AI in online content production and moderation facilitates the creation of an environment which produces "filter bubbles, echo chambers, and other elements that are antithetical to free access to information and media pluralism" (Oghia, 2018). Indeed, alongside the impact that personalization of information may have on freedom of expression, it also has a potentially negative impact on the ability of the media to provide a favourable environment for an inclusive, pluralistic debate (MSI-NET, 2018). Another possible threat to media pluralism and diversity are search engine algorithms that might also have bias towards specific content or content providers (European Data Protection Supervisor, 2019). Algorithms that prioritize existing linkages as an indicator of quality of content can end up reinforcing the status quo and neglecting new journalistic content to which fewer people have linked thus far.

Small changes to algorithms can have a "significant impact on publishers and news outlets in terms of traffic and financial viability," as with the algorithms classifying whether content is advertiser-friendly (Oghia, 2018). For example, when "YouTube reacted with a tighter use of its algorithm operated to detect 'not advertiser-friendly' content, [it was reported that it] affected independent media outlets, including comedians, political commentators and experts" (MSI-NET, 2018). AI also increasingly threatens to undermine media pluralism and diversity at the level of the consumption of journalism by enabling automated blocking and filtering that – from the point of view of international standards - may well constitute arbitrary, rather than legitimate, restriction.

Moreover, the personalization of information by algorithms is a process that can be covertly manipulated for political reasons. In contrast, news media is generally visible to several audiences who are able to recognize the political leanings of particular outlets and make informed judgements, such as on election issues, accordingly. However, the process of online micro-targeting permits political communications to be targeted at individuals or niche audiences and for the messages to be adapted to specific recipients (European Data Protection Supervisor, 2019). Big data derived "from citizens' online behavior, including from their social media use is the main fuel of contemporary political micro-targeting" (Nenadic, 2018). There is a downside in the way that targeted online advertising has been used by political strategists to "reach the right voters with the right message with near surgical precision" (Maréchal, 2018). Twitter CEO Jack Dorsey tweeted in October 2019 that "Internet political ads present entirely new challenges to civic discourse: machine learning-based optimization of

messaging and micro-targeting, unchecked misleading information, and deep fakes. All at increasing velocity, sophistication, and overwhelming scale". Such uses of elements of AI not only pose a threat to media pluralism, but also threaten the integrity of the electoral process and marginalize the role of journalism as constituting a vital source of verified information that is put forward in the public interest and in full public view.

Box 12: The Cambridge Analytica affair

The Committee of Experts on Internet Intermediaries (MSI-NET) of the Council of Europe has expressed concern regarding the danger of AI in undermining democratic processes and the right to free elections. During the 2016 U.S. presidential elections and UK EU membership referendum, micro-targeting was used to show certain content to a selected audience, based on the algorithmic analysis of people's data, in order to try and influence their votes. Cambridge Analytica, through the collection of personal data, offered this service to politicians.

This issue is concerning since political campaigns "are increasingly relying on big data analytics to influence opinions and voting through targeted messages or advertisement" (European Data Protection Supervisor 2018). Regulation of advertising on broadcast media to ensure fairness and transparency becomes less significant when AI is used to individualize political messaging on Internet platforms.

Micro-targeting and manipulation of content for political purposes is also closely linked to disinformation and misinformation. Deliberate spreading of disinformation to specific target groups can be intensified through the use of AI to amplify this phenomenon. Disinformation not only operates by purveying false facts, but often combines them with strong opinion and commentary that by their nature are not matters of veracity or falsehood. This stems from a business logic from a range of online platforms where a precedence is given to promoting engagement from web users, rather than to inform and educate users in the public interest.

However, disinformation cannot thrive easily in the face of credible and inclusive journalism that is based upon high standards of verification, as analyzed by UNESCO handbook for journalism educators, *'Journalism, Fake News and Disinformation'* (Ireton & Posetti, 2018f). The question this raises is whether AI could underpin a business model that gave prominence to journalism over disinformation.

3.3. Increasing automated disinformation and counter initiatives

During the 2016 U.S. election, "the most popular fake news stories were more widely shared on Facebook than the most popular mainstream news stories [and that] many people who see fake news stories report[ed] that they believe them" (Allcott & Gentzkow, 2017). Disinformation undermines the public sphere and democracy, and under its current trajectory, AI can be expected to increasingly be used to play a role in the rapid spread of automated disinformation.

Facebook and other Internet platforms are taking initiatives to remove disinformation. However, "identifying fake news sites and articles also raises important questions about who becomes the arbiter of truth" (Allcott & Gentzkow, 2017). The removal of content can constitute a violation of freedom of expression since disinformation is not always illegal and indeed some may be interpreted as being protected by international law, unless it constitutes a violation of rights, a threat to public safety, etc., or is coupled with advocacy to incite hostility, discrimination or violence.

There are current initiatives to identify and even counter disinformation through AI-based algorithms. A fact-checking algorithm that can detect falsehoods is something that several actors are pursuing. Informing the public of the false character of an item of content seems to offer more rights-friendly solutions than removing the content altogether. The complexity is that facts and falsehoods are selected and constructed within narrative frameworks and ideological or political perspectives. Therefore, meanings are structured in ways that go beyond a single fact or falsehood. In addition, meanings are often not expressed purely through information, but through emotions and signifiers of identity. Thus, disinformation is often just a component within wider propaganda messaging that includes incentive, hate, mockery, jest, outrage, etc. This is a challenge to AI's capacity to effectively detect malicious disinformation from types of legitimate expression at the present stage.

Another approach engaging AI is to assess patterns of behaviours to gauge process rather than content, and identify inauthentic users, such as co-ordinated accounts, fabricated identities, and mobilization of bots, which are signals of potential purpose to promote disinformation.

One initiative to counter disinformation is the EU Code of Practice for online platforms signed in September 2018 by four main actors: Google, Facebook, Twitter and Mozilla, as well as several trade associations: the European Association of Communications Agencies (EACA), the European Digital Media Association (EDiMA), the Interactive Advertising Bureau (IAB) Europe, the Union of Belgian Advertisers (UBA) and the World Federation of Advertisers (WFA). One of the main objectives of the Code of Practice for online platforms is to disrupt advertising revenue for accounts and websites misrepresenting information.

In some areas relevant to disinformation and journalism, as will be seen in the chapter on openness, the same AI software can pursue different and opposite goals. This is

the case regarding the fabrication and identification of 'deepfakes', which can be very convincing and realistic. This term is applied to represent "both video and audio artefacts that have been synthesized from existing digital data by means of deep learning neural networks models" (Barraclough & Barnes, 2019). These videos or audios are fabricated with the intention of deceiving the audience. Simultaneously, AI also seems to be the best technology to identify these 'deepfakes' and counter their deception (Lyu, 2018).

3.4. Protecting journalists and journalism sources in the era of AI

Promoting the safety of journalists and combatting impunity for those who attack them are central elements within UNESCO's support for press freedom on all media platforms. In 2017, UNESCO published a report *'Protecting Journalism Sources in the Digital Age'*, identifying and highlighting new and largely digital forms of challenges and dangers faced by journalists around the world, along with recommendations and guidelines to protect journalists and journalism (UNESCO, 2017b).

Whether AI can be used to detect, trace and help prevent attacks on journalists is a question worth exploring. The University of Sheffield is pursuing methods of using AI to identify patterns in data about a range of attacks (online and offline) against journalists (The University of Sheffield, 2018). This focus would need to include the role of AI deployment by intermediaries between producers and consumers of journalism. These intermediaries include social network platforms, search engines, and Internet access providers, who could implement policies for promoting and protecting journalism online.

Another challenge is the need to enhance protection of the confidentiality of sources of journalism in the digital age, as was recognized in the UNESCO-endorsed outcome document of the 'CONNECTing the Dots: Options for Future Action' conference held in 2015.

UNESCO's study, '*Protecting Journalism Sources in the Digital Age*', highlighted the growing requirement for Internet intermediaries and telecommunication companies to ensure they are 'wire-tap ready', with potentially chilling impacts on the protection of journalist's sources (Posetti, 2017).

Box 13: New forms of attacks against journalists

Today, there are new digital forms of attacks against journalists such as profiling, 'doxing'—the publication of an individual's private information on the Internet, 'deepfake' videos—superimposing existing videos onto source videos using AI, often with the intention of generating false news and misleading viewers, and trolling of journalists—which involves controversial or inflammatory messages against an individual. It is not clear to what extent, for example, automated trolling or Distributed Denial of Service (DDoS) attacks are being powered by AI as such, but the scenario is certainly plausible

Automated troll attacks on journalists are also concerning, especially with the presence of companies selling 'followers' on social media accounts and 'retweets' to individuals or organizations willing to pay for them. Investigative journalists from ProPublica went undercover and bought 10,000 retweets for a fake account from a company called 'Followers and Likes' for US$45, and 5,000 retweets for a fake English language account for US$28 (Angwin, 2017). The ProPublica journalists did this after having undergone extensive online harassment themselves. For example, a female ProPublica reporter was accused of being a so-called "presstitute" in a Twitter post which received over 20,000 retweets (Reporters Without Borders, 2018). This was most likely accomplished through a similar paid service used by social media accounts which coordinate inauthentic online behaviour and campaigns. Indeed, with just 100 dollars it is possible to obtain a "bot army," ready to automate synchronized harassment (Angwin, 2017).

Deepfake videos, noted above, are one AI technique being used inter alia against journalists. At the Internet Governance Forum (IGF) held at UNESCO in 2018, Elodie Vialle from *Reporters Without Borders* shared the organization's concern about the development of deepfake videos directed at journalists and especially female journalists (UNESCO, 2018e). This AI technique can allow people to harass female journalists by superimposing their faces onto pornographic content (Reporters Without Borders, 2018). Rana Ayyub, an investigative journalist from India, was a victim of this technique. While she was no stranger to online abuse, sexist and Islamophobic remarks, and misinformation about herself, she stated that the deepfake attack had a particular impact on her, "I always thought no one could harm me or intimidate me, but this incident really affected me in a way that I would never have anticipated" (Ayyub, 2018). Since the abuse, Ayyub has restricted the issues which she discusses online, imposing self-censorship. UN human rights experts have intervened by expressing alarm at this fake video and by calling for steps to protect her from hate and to investigate the attacks (UN OHCHR, 2018a)

In a recent study, the journalists interviewed expressed that "mass surveillance has the potential to silence whistleblowers and make investigative journalism increasingly difficult in all its forms" (Waters, 2018). The study's author used the theoretical framework of Foucault's 'Panopticism', which refers to "a state of conscious and permanent visibility that assures the automatic functioning of power" (1977). The journalists' awareness of mass surveillance meant that whether or not they were under surveillance, its presence was always felt, causing them to continuously alter their behaviour (Waters, 2018). Relevant sources for journalistic reports may similarly refrain from sharing important information, fearing that they are being watched or listened to by a facial recognition camera or a voice-recording IoT device in close proximity.

4. RIGHT TO EQUALITY

While many individuals do not experience discrimination while navigating the invisible web of algorithms around us, there are many who continue to be marginalized, discriminated against, denied opportunities or who experience adverse effects of particular algorithmic decision-making. For instance:

- In 2010, when researcher Safiya Umoja Noble searched the term 'black girls' on Google, it returned pages dominated by pornographic content. While the search results for the term no longer show pornographic content, a similar search query for 'Asian girls' gave highly sexualized results (Snow, 2018b).
- Amazon's program for automating its hiring process used an algorithm that was trained through analysis of resumes submitted to the company over a 10-year period. The results showed bias against female candidates, as the algorithm taught itself to give a lower rating to resumes that included words like 'women's chess club captain'. The results reflected the existing gender inequalities in the tech industry. Indeed, more than 60 per cent of employees in GAFAM[4] companies are male, and the numbers are much higher if only technical roles are considered (Dastin, 2018).
- Investigation into the COMPAS software used by courts in the United States to predict recidivism—the tendency of a convicted criminal to reoffend—revealed bias against African-Americans (Angwin, Larson, Mattu, & Kirchner, 2016).
- Credit risk algorithms are found to discriminate against potential borrowers based on their geographical location. While explicit classifiers like race and ethnicity might be hidden from the algorithms, such variables may correlate with other classifiers such as geographical location. Hence, algorithms end up assigning racially discriminatory credit scores even when race is not used as an input (O'Dwyer, 2018).

4 GAFAM is the acronym for Google, Apple, Facebook, Amazon and Microsoft.

Algorithms will become even more deeply entrenched in many people's lives given the rapid pace in which the 'data-ization' process is occurring as the field of AI is being developed. Therefore, it is important to consider how these may impede people's right to equality.

Social inclusion is one of UNESCO's core mandates is to work toward building knowledge societies that are socially inclusive so that all individuals and groups are able to take part in society. In turn, this depends on "improving the ability, opportunity, and dignity of those disadvantaged on the basis of their identity to take part in society" (World Bank, 2013). The same sentiment that no one should be left behind pervades the SDGs, and reinforce UNESCO's efforts in this area. As against these objectives, systemic violation of each person's right to equality is in principle an obstacle to inclusion.

This section focuses on discrimination emerging from different forms of automated decision-making which affect the right to equality. It identifies entry points for discrimination through algorithms (sometimes integrated into AI processing). It further proposes possible technical and non-technical approaches to the problem. However, any such reflection risks confusion if a core question is left unaddressed: What is discrimination?

4.1. What is discrimination?

Article 1 of the UDHR proclaims that "[a]ll human beings are born free and equal in dignity and rights" and Article 2 that "[e]veryone is entitled to all the rights and freedoms set forth in this Declaration without distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status" (UNGA Resolution 217, 1948).

The ICCPR reaffirms, in many provisions, this general principle of equality before the law and equal protection of the law. Article 2 states that each State party must ensure the rights recognized in the Covenant to all individuals without distinction of any kind. Article 26 is broader and provides protection against discrimination explicitly and not limited to the rights of the ICCPR:

> "All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."

This document, as well as several other international legal treaties, prohibits discrimination based on a non-exhaustive list of group identities without attempting to delineate the meaning of discrimination (Vandenhole, 2005). Equality and non-discrimination are widely recognized as the positive and negative statements of the same principle (Bayefsky, 1990).

The word 'discriminate' traces its origins to the Latin 'discriminate', which is to 'distinguish between'. In this strict sense, discrimination itself would be devoid of any moralized connotations. Therefore, it is important to specify the conditions that make discrimination objectionable and a factor against social inclusion. As per the Stanford Encyclopedia of Philosophy, discrimination is morally wrong when it involves i) imposition of a relative disadvantage or deprivation on persons belonging to a certain group, and ii) is wrongful (in parts) if the disadvantage is bestowed on the victims because of their membership of the group (Altman, 2016).

However, establishing the framework for recognizing discrimination does not necessarily equip us with tools that are sufficient for analysis of algorithmic discrimination. Algorithms can acquire a discriminatory nature through multiple pathways. These mainly include features of the algorithm being biased intentionally or unintentionally by programmers or through the reinforcement of biases present in training data for machine learning algorithms. Algorithms need to be subjected to an analysis for discrimination that can identify direct, indirect and institutional discrimination so that relevant regulatory or technical solutions can be applied.

Many types of discrimination can be indirect; for example, an algorithm that uses mobile phone usage patterns to determine credit worthiness of a person is discriminatory if it assigns high credit risk to women in communities that i) have low mobile phone usage or ii) do not own mobile phones. The condition applied may appear to be equal and fair, but it disadvantages a particular group (Altman, 2016).

Algorithms can cause and exacerbate these multiple forms of discrimination. Existing social and political biases are being systemized in machine learning algorithms in many ways (Packin & Lev Aretz, 2018). Furthermore, it is worth investigation into the potential new forms of discrimination that AI may bring about, such as exclusions decided based on statistical correlations that do not necessarily correspond to socially salient characteristics, but that are nonetheless strongly linked to one's personal identity.

4.2. How is discrimination designed into algorithms?

A predilection for objective decision-making combined with the notion that algorithms process input data to produce objective decisions has given them an air of unquestioned superiority over decisions taken by humans. Indeed, machine learning algorithms are given an "aura of truth, objectivity, and accuracy" (Boyd & Crawford, 2012). For instance, in a US Court case concerning the theft of a lawnmower, the prosecutor recommended a one-year prison sentence followed by a period of supervision. However, the judge, relying on an algorithm's high-risk assessment for the individual, overturned the plea deal reached between the prosecution and the defense and imposed a two-year prison sentence followed by three years of supervision (Angwin, Larson, Mattu, & Kirchner, 2016).

There is a strong case for caution in our reliance on algorithms as the final arbitrator of decisions as they, at best, provide only useful insights. Any claims of fairness of algo-

rithms need to be qualified by the fact that the process of algorithmic decision-making has two key elements: i) human programmers who make critical choices in framing the problem and validity of the output; and ii) data that may represent historical biases, misrepresent groups or not represent them at all.

The following section highlights the human and data-driven entry points for potential biasing of algorithms (Barocas & Selbst, 2016).

i) Programmer-driven bias

a) **Definition of 'target variables' and 'class labels'**: The target variable is the variable that needs to be predicted, which is the output of the algorithm. The class label categorizes all possible target variables into mutually exclusive sets. Programmers, based on their understanding of the problem, make the choice of variables and labels. In the case of a spam filter, the email can be straightforwardly classified as one of two labels: spam or not spam. However, in the case of a problem like a hiring algorithm, the class of labels is non-binary and may reflect the programmer or the organization's biases with the effect of disadvantaging certain social groups.

b) **Feature Selection**: Programmers choose the attributes of the data that should be observed and used for analysis. If the selected features of the data do not adequately represent some groups of people at a granularity that captures their differences from other groups, then they can be victims of severe disadvantage due to automated decision-making.

c) **Masking**: Algorithmic decision-making can be used as a mask by those who want to hide their biases and intentions of disadvantaging certain groups behind a façade of neutrality provided by algorithms. This is achieved through prejudiced target definition, class labelling, feature selection and data manipulation.

ii) Data-driven bias

a) **Biased training data**: If the rules extracted by the machine learning algorithm from any given set of data are considered legitimate, prejudices and omissions embedded in the example data will be repeated in the predictive model.

b) **Representativeness of the sample data**: A dataset can be biased by the data it does not contain. If the training data reflects an unrepresentative sample of the population, then under- or over-represented groups may be disadvantaged by the algorithm. Lack of representation can also stem from dark zones of shadows in the data, i.e. when the data for certain groups of population is not captured at all because of their existence outside the data-gathering stream. For instance, the use of mobile phone data as a proxy indicator of the user's ability to repay loans may disadvantage people who have limited or no access to mobile phones. At the same time, it is important to note that even representative data sets reflect historical and societal biases, for example against minorities overly-represented in prison populations or women in less

prestigious jobs. The data's very 'representativeness' can therefore perpetuate discrimination and inequality, when in fact a consciously adapted dataset that corrects for such social inequalities might produce less discriminatory outcomes from algorithms trained on this basis and then applied to fresh cases (such as when used for informing custodial sentencing or automated scrutinizing of job applications).

c) **Correlation is not causation**: Decision-making based on correlations may lead to faulty inferences. For instance: "Imagine spending a few hours looking online for information on deep fat fryers. You could be looking for a gift for a friend or researching a report for cooking school. But to a data miner, tracking your online viewing, this hunt could be read as a telltale sign of an unhealthy habit — a data-based prediction that could make its way to a health insurer or potential employer" (Barocas, 2014). In addition, it is worth noting that prediction of future events is based on the assumption that past events are representative of future events given similar and unchanged underlying conditions. The problem is the assumption about the unchanged underlying conditions and continued behaviour.

d) **Cyclical resource misallocation**: The predictions generated by algorithms based on data can allocate resources away from under-represented groups. The subsequent monitoring data would follow the same pattern and aggravate the discrimination against underrepresented groups. For instance, if a local government tracks information about potholes based on the number of bumps on the road as registered by mobile phones of vehicle owners, then the government may direct resources towards more affluent areas with more mobile phone and vehicle users. This further lowers the quality of roads in less well-off neighborhoods (Crawford, 2013).

e) **Proxy induced bias**: Even when variables that directly represent group membership are removed from the data in order to prevent discrimination, there may be other variables, necessary for the analysis, which correlate with the group identifying features and can lead to discrimination. For instance, even if direct indicators of race are removed from the data set, other variables like income level or consumption patterns may correlate with race and lead to racially biased decisions. Data are needed about the consequences of automated decisions in order to identify indirect discrimination.

Box 14: Data-driven biases which entail race-based discrimination

In many instances, machine learning algorithms train on datasets that are not representative. When these algorithms are integrated into products and services that enable decision-making, they can be discriminatory. For instance, researchers working on fairness in algorithms have demonstrated that datasets (IJB-A and Adience - two facial analysis benchmarks) used to train facial recognition algorithms are 'overwhelmingly composed of lighter skinned subjects' (Buolamwini & Gebru, 2018). IJB-A and Adience have 79.6 and 86.2 percent lighter skinned subjects. A direct consequence of these non-representative datasets is that downstream applications developed using those tend to misclassify results. For instance, in some gender classification systems, darker-skinned females are most misclassified with an error rate up to 34.7 percent as compared to lighter-skinned males where the maximum error rate is only 0.8 percent (Buolamwini & Gebru, 2018). Such different error rates are prevalent regardless of which company or country developed the system: US companies Microsoft and IBM had error rates of 21 percent and 35 percent respectively for black women while China's Megvii had an error rate of 35 percent (Buolamwini & Gebru, 2018).

Another example was in 2015 when Google's Photos application labeled two dark-skinned individuals as 'gorillas'. The company corrected the mistake and apologized but a recent report shows that the image labelling technology is far from perfect, and a quick solution of removing 'gorillas' from the tags may not be addressing the bias problem at its core (Simonite, 2018).

Therefore, we see that decision-making by algorithms is susceptible to both human- and data- driven biases. Much algorithmic decision-making is shaped by implicit prejudices of programmers or those internalized in the data.

Historical and sociological considerations provide crucial background information necessary to determine fairness in algorithmic decision-making contexts and results (Michael, Van Kleek, & Binns, 2018). In-depth algorithmic analysis is needed to uphold the right to equality and to ensure that historical inequalities related to gender, race/ethnicity, sexual orientation and identity, socio-economical class, disability and other grounds of stigmatization are neither perpetuated nor considered 'objective'.

5. CONCLUSION AND POLICY OPTIONS

In 2018 and 2019, UNESCO organized multiple discussions centered on the challenges and opportunities of the digital age with experts representing technology firms, researchers, governments and human rights organizations to share their reflections on the impact of AI on society. One such interaction revealed the gulf between approaches adopted by technologists at the forefront of development of AI and human rights organizations advocating for fairness. On one hand was a human rights advocate who argued for a moratorium on AI until the discrimination and biases perpetrated and perpetuated by algorithms are addressed completely. On the other hand, a senior AI researcher, while recognizing the importance of human rights, said that it was impossible to stop the development of technology.

Each side argued passionately for their point of view. However, it was clear from the discussion that progress on the question of the development of technology and its impact on society means overcoming silos. Discussion needs to be multi-disciplinary with stakeholders willing to engage with each other and find solutions at the intersection of their respective domains. This issue is unpacked further in the chapter on multi-stakeholder governance.

Options for all stakeholders

- Develop and use a human rights-based framework for AI under the prism of international human rights standards to set clear guidelines to avoid violations of human rights (including those the rights to freedom of expression, privacy and equality).
- Promote and evaluate methods that can assess algorithmic discrimination in order to protect the right to equality, in particular that of historically marginalized populations.
- Initiate, coordinate and support multi-stakeholder and interdisciplinary research on the human rights implications of AI.

Options for States

- Develop adequate policy and regulatory frameworks to address the human rights challenges posed by the development and application of AI, providing mechanisms for preventing human rights violations, as well as for transparency, accountability and remedy processes.

- Give attention to prosecuting the producers of demonstrated harmful content according to legal frameworks in line with international standards, and to providing media and information literacy of audiences rather than putting exclusive focus on requiring action from intermediaries.
- Be aware that rendering Internet intermediaries liable for user-generated content may encourage over-use of AI in content moderation, which, in turn, risks a negative effect on freedom of expression.
- Take effective measures to ensure that algorithms are not exploited to impede the right to free elections.
- Support the UN Plan of Action on the Safety of Journalists and the Issue of Impunity and address the AI-assisted attacks on journalists and media workers.
- Evaluate if existing regulation against discrimination enables an individual to seek remedy for algorithmic discrimination.
- Ensure that the public sector's use of AI in decision-making is transparent and consistent with human rights obligations.

Options for the private sector, Internet intermediaries and the technical community

- Conduct human rights risk assessments and due diligence on AI applications in order to ensure that they do not interfere with the full enjoyment of fundamental human rights and freedoms at:
  - *Ex ante* level: Avoid discrimination in the selection of datasets and programmers' design choices, and make explicit the values informing these choices.
  - *Ex post* level: Closely monitor outcomes that could infringe on the right to expression, privacy and equality, as well as other rights.
- Create and provide users with options to opt out of receiving personalized content and to choose modalities for ordering the presentation of content based on other criteria.
- Promote and demonstrate transparency by providing information on the following in an accessible manner:
  - Algorithm development and application in the personalization of content presentation;
  - Statistics on the use of AI systems in content-moderation, including the number of removals (or other actions) done completely and partly by AI and the frequency of human moderators deciding against AI recommendations;

- The collection and use of data from users and consumers, including what type of data they collect, how they store it and process it, if they share it with or sell it to third parties and for which purpose, as well as uncertainties on how and why data will be used;
- Potential AI flaws and risks that could lead to violations of users' rights

- Promptly notify the user of removed content and explain the process and rationale behind the removal, as well as the appeal mechanism.
- Implement appeal mechanisms and efficient complaint systems that provide remedies to people whose rights have been infringed or who have legitimate cause to have their information removed or corrected.

Options for academia

- Engage in rights-oriented research on the social, economic and political effects of AI personalization of content, including the possible effects of 'echo chambers' in the development of political opinions, as well as in radicalization leading to violent extremism.
- Conduct research on potential chilling effects of AI use in mass-surveillance, restrictions on freedom expression, and impacts on limitations of other human rights.
- Pursue research on AI's effects on media pluralism and sustainability as well as research on the use of AI in strengthening journalism and media institutions.
- Continue research on algorithmic discrimination and how technical solutions can be implemented in order to ensure the development of AI systems which respect the right to equality.

Options for civil society

- Adopt a 'watchdog' role to monitor AI's human rights violations and expose them to the public.
- Strengthen media and information literacy in order to better understand the human-rights implications of AI.

Options for media actors

- Consider and reflect on the implications of AI on the practice of journalism and media development in order to strengthen and protect freedom of the press, as well as the safety of journalists.

- Get trained to better investigate and report on AI development and its applications, including exposure of abuses and biases in AI as well as the current and realistically possible benefits.
- Make use of AI and new technologies in the practice of reporting, news production and content dissemination in a manner that is consistent with international human rights standards, including privacy.

Options for intergovernmental organizations, including UNESCO

- Communicate with multiple efforts regarding ethics and AI, linking them to human rights, and organize multi-stakeholder dialogues in which human rights concerns are addressed.
- Assist States in complying with international standards on human rights regarding AI.
- Promote the use of AI in journalism and media that can support and protect their roles in society.
- Engage with debates on guidelines for transparency as regards automated journalism and the use of AI in media.

OPENNESS AND AI

2

CHAPTER 2: OPENNESS AND AI

Openness is an important feature of knowledge societies (UNESCO, 2015b). UNESCO advocates for open access to scientific research, open data, open educational resources and open science as part of its efforts to strengthen universal access to information and to bridge information inequalities. In the context of artificial intelligence, openness may refer to transparency in general and, more specifically, the practice of releasing to the public source code, knowledge platforms, algorithms and any scientific insights gained in the course of research (Bostrom, 2016). Openness further denotes the absence of unfair obstacles and entry barriers to participation in AI development, application and review (see box 15 below).

Openness encourages wider use and engagement with technology, enabling individuals and societies to leverage technology to their advantage. This chapter looks at openness and AI from several dimensions. First, how open is the research community working on AI, in both academia and the private sector. This specifically concerns openness in publication and diffusion of AI research and tools. Second, how open are the sources of data that can be used for the development of AI systems. Third, 'explainability' and transparency in how AI systems make decisions and transparency and accountability on the part of organizations (governments, private sector and others) that use AI systems for decision making. Fourth, the role of open markets and competition in AI research. Fifth, AI is explored from the lens of a dual use technology and the risks associated with potential misuse of open publication of research and several publication models are discussed. The chapter concludes with options for strengthening openness with respect to AI along the different dimensions discussed.

Box 15: UNESCO's position on Openness for Internet Universality

'Internet Universality' highlights the norm of openness of the Internet. This designation recognizes the importance of technological issues such as open standards, as well as policy standards for open access to knowledge and information. Openness also signals the importance of ease of entry of actors and the absence of closure that might otherwise be imposed through monopolies.

The Internet should be open for all to develop or to take advantage of its resources and opportunities, in whichever ways is most appropriate or valuable to them. Through openness, the concept of Internet Universality acknowledges the integrity of the Internet as enabling a common global exchange rather than it being confined to 'walled gardens' based on incompatible technologies. It highlights the importance of digital issues such as open standards and open access to knowledge and information.

Open standards, interoperability, public application programming interfaces (APIs) and open source software have made a vital contribution to the Internet. Open markets have also played an important part in the development of the Internet, allowing market access to innovative and competitive businesses rather than excluding these through restrictive licensing arrangements or protectionist limitations on service provision.

1. OPENNESS IN AI RESEARCH

There is strong interest in AI research in academia and the private sector. Such interest is not surprising given AI's potential to generate additional economic activity of around $13 trillion by 2030, accounting for an additional 1.2 per cent in global GDP growth per year (Bughin, Seong, Manyika, Chui, & Joshi, 2018). In 2016, as per some estimates, tech giants like Google and Baidu spent around $20 to $30 billion on AI, about 90 per cent of which was spent on research and development (Bughin, et al., 2017).

Openness in research would hasten diffusion of new knowledge and allow more people to build their research and applications on the basis of state-of-the-art AI techniques accessible to all (Bostrom, 2016). In the past five years, AI research publications have grown by 12.9 per cent annually and now about 60,000 research publications are generated per year (Elsevier, 2018). Figure 1 shows the number of AI papers on arXiv,[1] an open archive of research papers, by sub-category. Papers in the fields of 'Machine Learning' and 'Computer Vision and Pattern Recognition' have grown by 37.4 per cent annually in the last five years (Elsevier, 2018).

1 Started in August 1991, arXiv.org is a highly automated electronic archive and distribution server for research articles. Areas covered include physics, mathematics, computer science, nonlinear sciences, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics.

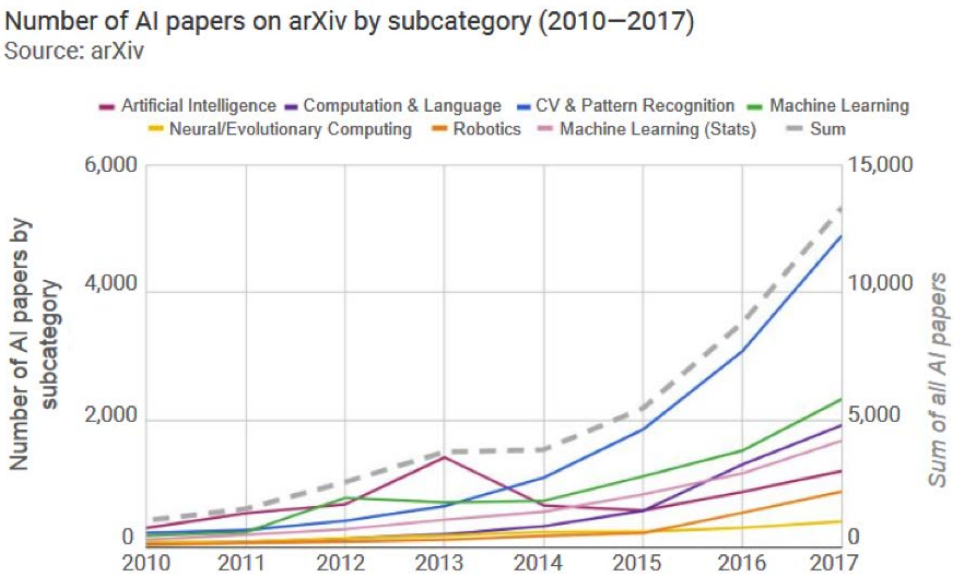


Figure 1: Number of AI papers on arXiv by subcategory (Shoham, et al., 2018)

The UNESCO Science Report 2015 highlights the role of different work cultures in the way knowledge generated is diffused by public and private sector researchers. Traditionally, scientists working in public universities tend to publish openly as their reputation depends on the peer assessment of their work. Scientists working in private firms are beholden to the business interests of their employers that may require secrecy and appropriation of knowledge for the firm's interest (UNESCO, 2015c). However, researchers, developers and firms working in the field of AI have demonstrated a proclivity towards openness by sharing their work regularly at academic conferences and via open source platforms (Bostrom, 2016).

AI researchers have demonstrated their commitment to open research. For instance, a petition for a boycott against a non-open access AI journal garnered more than 3,000 signatures, mostly from researchers (Hutson, 2018). Tom Dietterich of Oregon State University, the initiator of the boycott, stated that AI should be transparent and open to the community (Robitzski, 2018). Some computer scientists have argued for a system of open publishing and reviewing to quicken the distribution process and increase transparency by openly publishing paper reviews as well (LeCun, 2009) (David Soergel, 2013). As a result, several major AI conferences have started using platforms like OpenReview.net that provide an avenue for openness in scientific communication, particularly in the peer review process (Hutson, 2018).

Furthermore, researchers frequently share the source code and detailed architecture that enable others to test and further develop demonstrated technologies, thereby strengthening knowledge exchange within the community. In the field of AI, there are already many open source technologies (Bostrom, 2016).

For researchers, publication of their research allows them to share knowledge and signal their capabilities to a larger audience within the technology community. Researchers often prefer to join companies that allow open publication of research as it increases their market value and standing within the AI research community (Bostrom, 2016).

Mobility of researchers and engineers between technology firms and university research labs encourages exchange of knowledge between different organizations. Another way in which mobility is achieved is through research labs set-up across the world by private sector firms (UNESCO, 2015c).

In the case of the private sector, there are several reasons to encourage openly sharing research done by their employees. These include to:

i) Showcase their research capabilities in order to attract more talent,

ii) Improve their products by involving the wider AI community, including hackers and security experts, to test, challenge and find bugs in their current offering,

iii) Build a community of developers to encourage the development of downstream applications based on their technology kernel, and

iv) Influence industry standards based on their technology.

For reasons mentioned above and as part of their business models, many firms are making their machine learning platforms and cloud services available more openly to researchers and developers, which helps in bridging divides in access to both technological resources and hardware that may not be available to all users. For instance, Amazon Web Services, Google Cloud Platform, and Microsoft Azure Cloud recognize that their computing infrastructure can be a line of business beyond their own exclusive use, and sell cloud-computing services (Varian, 2018).

Google's TensorFlow, one of the most extensively-used technologies in AI, is an open source machine learning framework which allows users to develop neural networks and other computational models using flowgraphs (Garbade, 2018). Foundations have also emerged to provide environments for open source collaboration across institutions and enthusiasts. For example, the Deep Learning Foundation at the Linux Foundation incorporate projects such as Acumos AI, which is a platform and framework that makes it easy to build, share and deploy AI apps (Bommireddipalli, et al., 2018).

Further, attesting to openness within the private sector with regard to AI research is that reportedly over 70 per cent of recent corporate-driven AI research in the US was published as conference papers (Elsevier, 2018).

Research consortiums (like Open AI) have been established with the objective of making AI research open and accessible, thereby ensuring the development of technologies in ways that mitigate knowledge monopolies dominating the field of AI in the future (although data monopolies and closed knowledge systems persist in spheres

such as the military, Internet companies and health industries, among others).

However, gaps in understanding the extent of openness in the private sector with respect to AI remain, including regarding the role of patents. Further, there is a need for more research to understand the differences in access to these services in terms of demographics and geography.

Box 16: Key technologies and platforms

TensorFlow is an open source software library for high performance numerical computation. Its flexible architecture allows for easy deployment of computation across a variety of platforms, and from desktops to clusters of servers to mobile and edge devices (edge computing is a distributed computing paradigm that brings computer data storage closer to the location where it is needed). Researchers and engineers at Google developed TensorFlow to provide support for machine learning and deep learning. Its flexible numerical computation core is useful across many scientific domains. (Source: https://www.tensorflow.org/)

GitHub is an open platform for software developers to work together to solve challenging problems using the most important technologies. As of August 2019, the platform had more than 40 million registered users across the world and 100 million repositories. (Source: https://github.com/about/)

While there are compelling altruistic, technological and market-driven reasons for openness in AI research, it should be noted that this is not sufficient in itself to allow individuals to develop skills in AI and contribute to the development of AI. Individuals need, *inter alia*, access to data, and open data can play a major role in this regard

2. OPEN DATA AND AI

Open data policies are concerned with openly publishing data gathered by governments (and, sometimes, other stakeholders) for individuals, businesses, academia and civil society organizations to use data to support their own objectives. The benefits of open data policies include improved access to knowledge, opportunities for innovation and service provision, improved data analysis through recombination of data from diverse sources, and better policymaking because of enhanced transparency and accountability. Data protection arrangements are important in ensuring that open data sets do not undermine individual privacy (UNESCO, 2018a). In addition, it may be noted that data, as a result of being continuously processed, also undergo transformation in the information they provide. For instance, labelling of data adds another layer of information that allows data to be converted into knowledge.

Government datasets are just one of the sources for obtaining useful data for the development and application of AI and its elements. Discussion surrounding open data vis-à-vis AI further signals a far wider array of sources like web scraping, data collected as a by-product of mobile applications, and data commons, among others. This section illustrates possible sources of data that are useful for training and applying AI (Varian, 2019).

i) **Web scraping**: This involves automated collection of data from public websites. This data source is openly available but there are ethical and legal concerns regarding the download and use of data scraped from websites. For instance, a programmer scraped 40,000 profiles on the dating site Tinder to create a data set for training image recognition algorithms, raising privacy concerns for users of the platform (Lomas, 2017). Some organizations prevent web scraping by implementing additional controls on their websites.

ii) **Data generated through an offer of 'free' services**: Social media platforms are a prime example of this category of data generation. They exchange access to their platform for user data that can be monetized in different ways. Such data are often useful for improving services. For instance, Google developed its voice recognition data expertise based on the voice commands given by users to their phones, and the corresponding choices they made based on the results of the voice search. Similarly, ReCAPTCHA, a technology used to detect if a real person or a robot is trying to access an online service, collects data by asking users to label pictures and this in turn is used to train its machine learning algorithms (O'Malley, 2018). Data collected in this manner may or may not be available publicly for other users depending on whether the owner wants to share it. Open source labelling of data could be an option to support the development of data commons.

iii) **Data collected as byproducts of operations**: This is data that is collected as part of businesses' routine operations. It may include consumer invoices generated at a restaurant that can be used to fine-tune the weekly or seasonal demand for food and hence help to manage grocery bills and reduce food waste. Data arising from the Internet of Things can be a valuable open resource.

iv) **Computer generated data**: Machine learning algorithms may generate their own data as well. For instance, the AlphaGo algorithm generated data by playing the game 'Go' against itself. Similarly, synthetic images created by modifying original images are used to train algorithms.

v) **Hiring humans to label artifacts for use as data**: Humans are hired to manually label data to be used for training algorithms. However, this may be costly and time consuming as an investment. Tens of thousands of people are working across the world in labelling data as independent contractors or through crowdsourcing platforms like Amazon Mechanical Turk that allow firms to distribute data labelling tasks for anyone to take up (Metz, 2019).

vi) **Data commons**: Online data repositories for words, images and other forms of media exist and have been developed through user contributions. Data commons act as shared resources and are discussed in more detail in the next chapter about Access.

There are questions concerning the legality of website data scraping, the concentration of data within the hands of few companies, and the building of open data repositories for improving access to data for training algorithms. However, policies that encourage placing of labelled data in the open with adequate regard for privacy would be useful for development of AI. Further dimensions to note are found in the chapter on Rights, which addresses data-driven discrimination and the chapter on Access that discusses how access to data can bridge the AI digital divide. The next section discusses concerns related to openness and transparency of AI algorithms.

3. OPENNESS WITHIN AI: BLACK BOX AND TRANSPARENCY CONCERNS

Automation of decision-making as an element that can stand-alone but is also becoming a feature of AI's underlying components confronts policymakers with the question of ensuring accountability and transparency in decisions taken by machines. Some types of AI do not rely on predefined programs to perform tasks. Instead, machine learning algorithms "can learn, adapt to changes in a problem's environment, establish patterns in situations where rules are not known, and deal with fuzzy or incomplete information" (Negnevitsky, 2011). Therefore, even though the steps taken to reach a decision can be described, the detail is unlikely to make us any more knowledgeable about how the decision was actually taken. The process is akin to relying on intuition to arrive at a decision, with no clear understanding of where the intuition comes from (Mukherjee, 2017). In this sense, openness faces a technical challenge with regard to explainability of algorithmic decisions. This is also called the 'black box' problem of AI for two reasons: i) the complexity, and ii) the dimensionality of the algorithmic decision-making, which inhibit humans from understanding it (Bathaee, 2018). These issues are evident in two widely used AI methods, deep neural networks and support vector machines, which are particularly resistant to openness. These are discussed below:

i) **Deep neural networks**: A deep neural network is based on the ability of a network of artificial neurons to learn incrementally based on its programming and outcomes of data-processing. As in the case of human neurons, the useful linkages are strengthened, and the extraneous ones are discarded. In this methodology, 'several layers of interconnected neurons are used to progressively find patterns in data or to make logical or relational connections between data points' (Negnevitsky, 2011). Since no single 'neuron' encodes

a distinct part of the decision-making process and the decision is arrived at based on the network of 'neurons', it is not possible at this point of technological development to trace the decision down to specific logical steps (Bathaee, 2018). Thus, deep neural networks entail decision-making that is complex to unravel.

ii) **Support vector machines**: Humans have the ability to imagine three-dimensional spaces, i.e. to create a mental image of a plane using three variables; anything beyond three dimensions is not easily accessible to our brains (Carroll, 2009). Support vector machines are opaque to humans because they arrive at a decision by finding geometric patterns among many variables that humans cannot easily visualize. Therefore, non-linear curves generated by support vector machines are a black box to the human mind because of their high dimensionality.

The 'black box' problem makes AI opaque and raises several challenging questions regarding accountability, transparency and liability for decisions taken by algorithms. Creators of AI algorithms may define their algorithm's overarching goals, but 'black box' AI may achieve these goals in ways that even their creators may not understand or are able to predict. The steps in between would remain obscure given the complexity and dimensionality factors discussed above. Therefore, the question of intent, traditionally used as one factor to determine liability, is impossible to satisfy in certain cases, since machines cannot be said to have intent and the human creators' only intent was to achieve the defined goal.

An example of the black box problem is witnessed in the legal system. In the US, predictive coding is already being used to determine whether recidivism is more likely in criminal matters and to assist in making decisions about sentencing. A Wisconsin man, Eric L. Loomis was sentenced to six years in prison based in part on a private company's proprietary software. He challenged that his right to due process was violated by a judge's consideration of a report generated by the software's secret algorithm, one that Loomis was unable to inspect or challenge (Liptak, 2017). In this case, the private company might have understood the algorithm used to reach the decision, but in cases where even the algorithm creators do not understand how it arrived at its decision, who should be accountable for mistakes? When an algorithm is not designed by humans, it becomes difficult to determine whether it used spurious correlations or discriminated against a vulnerable group, and even more difficult to decide who should be responsible when it does discriminate.

Another interesting example is that of an AI technology called CycleGAN developed by researchers at Stanford and Google. This AI was designed to convert satellite imagery into street maps and back again, and had two discrete tasks:

i) Convert aerial photos into maps

ii) Convert a map into an aerial photo resembling the original photo

The algorithm's efficiency was rated based on how closely it could recreate the original photo from the map. The best way to do it was left up to the algorithm to determine. The results were surprising, as researchers found out that the AI became extremely efficient by skipping the middle step of generating a map from the image and then using the map to generate an aerial image again. Instead, it started producing the images directly from the original image. Even though the AI was perfectly logical, given its goal of achieving a higher rating, it achieved this by what humans would term as 'cheating' (Tech2News, 2019). The consequences of similar 'cheating' by algorithms could be detrimental to public interest.

The black box problem is further complicated when multiple algorithms interact with each other and pose systemic risks. For instance, a securities-trading program may have the objective of maximizing profit, but whether it achieves this through market manipulation or through fair means is difficult to discern ex-ante and ex-post (Bathaee, 2018). Since many such algorithms with different levels of complexity respond to changes in the stock market, i.e. implicitly interacting with each other, it creates a situation of 'legions of powerful, superfast trading algorithms—simple instructions that interact to create a market that is incomprehensible to the human mind and impossible to predict' (Salmon & Stokes, 2010).

Despite these hurdles, efforts are underway to make algorithms more open and transparent. The EU's General Data Protection Regulation (GDPR) requires organizations to explain certain decisions taken by algorithms (Algorithm Watch, 2019). This directive empowers people with the right to 'meaningful information about the logic involved' in the algorithm (Goodman & Flaxman, 2017). There are proposals concerning algorithmic auditing similar to financial audits with dedicated data professionals, standards and guidelines to perform audits (Guszcza, Rahwan, Bible, Cebrian, & Katyal, 2018). Given the deliberate corporate or technological lack of openness in sharing information about algorithms and the underlying data, some have argued for crowd-sourcing results of algorithms and then subjecting them to thorough analysis to decipher any bias and discrimination (Stray, 2018).

Efforts outside the public sector are pointed more towards increasing the understandability of AI. From the private sector, IBM has created AI Explainability 360, an open source collection of state-of-the-art algorithms that use a range of techniques to explain AI model decision-making, as well as AI OpenScale, which monitors a developer's finished AI code for fairness and uncovers hidden biases that may creep in throughout the lifetime of the application (IBM, 2019a; IBM, 2019b). Google has also created Testing with Concept Activation Vectors, which is an interpretability method that can be used to understand what signals neural network models use for prediction (Been, et al., 2018). From academia, ProtoDash is an algorithm for finding 'prototypes' in an existing machine learning program, that is, a subset of the data that have greater influence on the predictive power of the model. For example, it could explain credit score model results to a consumer recently denied a loan, or to a loan officer who needs an explanation of AI model decision-making in order to comply with law (Gurumoorthy, Dhurandhar, Cecchi, & Aggarwal, 2019). This non-exhaustive list of technical solutions has the potential to help shed some light on openness in rela-

tion to complex AI algorithms. However, many of the current proposals for enhancing transparency in automated decision-making are still far from fruition, and more efforts are needed to solve AI's black box problem.

4. ROLE OF MARKETS IN OPEN AI

Open markets, even the imperfectly open markets that exist today, foster competition between firms and enhance consumer welfare by bringing the price of products and services close to the marginal cost of production. In this regard, openness in AI allows the diffusion of innovative research among competitive firms. Given the current level of openness in the AI research community and efforts by researchers to push for more openness, firms are quick to enhance state of the art technology and deploy it to improve their products and services. However, in an attempt to gain from the first mover's advantage and expand their market shares, firms may limit sharing and deployment of their AI technologies. In some instances, competition may compromise human rights and not allow firms to cooperate to limit risks of AI. Issues around the behaviour of market actors in the use and deployment of AI requires a multi-stakeholder dialogue to develop ethically accountable corporate AI practices. The United Nations' 'Guiding Principles on Business and Human Rights' can serve as a useful tool to guide States and businesses in navigating the human rights risks they may face while developing or deploying AI systems. Market failure to address these or other concerns such as concentrations of market sphere, which reduce opportunities for fair competition, is a reason for regulatory interventions.

5. RISKS OF OPENNESS AND RESPONSES

While openness and transparency have several advantages as discussed above, openness can also pose risks associated with the misuse of technology. This section identifies some ways in which AI is being deployed for applications that can present a risk to human rights and democracy.

AI can be deployed for beneficial or harmful purposes. Access to algorithms, source codes and data sets in digital repositories facilitates the development of beneficial uses of new AI technologies. However, the same knowledge can also be used to develop harmful applications. For instance, drone aircrafts using AI systems for flight-path optimization can be used for delivery of emergency supplies in remote and inaccessible areas, but they can also be used for unjustifiable military attacks. On the one hand, AI is used to detect disinformation, on the other hand it can be used to spread disinformation through fake videos, images and headlines at low cost, wider scale and higher efficiency than before (Brundage, et al., 2018; Conner-Simons, 2018; Schwartz, 2018).

A group of experts at the Asilomar Conference[2] has recognized four observations about AI (Brundage, et al., 2018):

i) AI systems are commonly both efficient and scalable,

ii) AI systems can exceed human capabilities,

iii) AI systems can increase anonymity and psychological distance, and

iv) AI developments lend themselves to rapid diffusion

Together, these observations demonstrate the potency of AI as a tool with potential for harm. Automated malicious emails, websites, links and realistic chat bots can be created and tailored to individual users at low costs, and they can be used to covertly target specific communities or spread political propaganda (Benkler, Faris, Bourassa, & Roberts, 2018).

The previous chapter on human rights showed how AI might contribute to discrimination due to programmer bias or data related concerns. These factors call for transparency as part of openness. At the same time, openness can also be compromised by misuse of the technology for harmful purposes.

For instance, openness can allow 'gaming the system'. An illustration is a technical feature that can distort the outcomes of an algorithm is called adversarial example. Adversarial examples are inputs to machine learning algorithms that force the model to make mistakes (GoodFellow, et al., 2017). Figure 2 shows how an adversarial input causes the algorithm to change the classification of an image from that of a panda to a gibbon (Goodfellow, Shlens, & Szegedy, 2015).

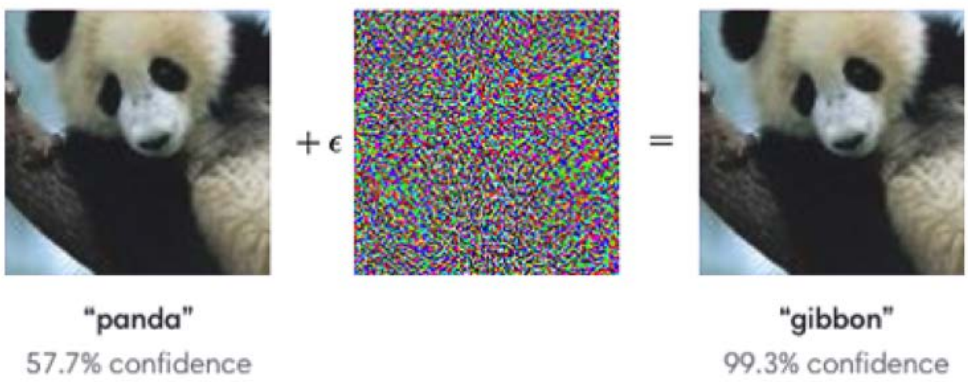


Figure 2: Example of adversarial input layer added over the image of a panda leading to its misclassification as a gibbon (Goodfellow, Shlens, & Szegedy, 2015)

2 The Beneficial AI 2017 Conference was a workshop and conference where leading AI researchers from academia, industry and thought leaders in economics, law, ethics and philosophy met to develop the Asilomar AI Principles: https://futureoflife.org/ai-principles/

The example of 'panda' shown above may seem innocuous, however, the same technology when used to hack algorithms may also cause autonomous vehicles to misrecognize and misinterpret road signs, which can lead to passenger deaths (Papernot, et al., 2017).

Apart from adversarial examples that distort the use of algorithms to cause harm, the openness of AI to many users entails risks. For instance, Figure 3 shows how the technology for generation of realistic synthetic faces has made significant advances in the past five years, from the grainy face in 2014 to a sharp and realistic one in 2017 (Brundage, et al., 2018). Realistic non-existent persons can be created using AI. Such technologies can be used to lure people to believe false information, particularly to spread harmful disinformation and hatred on social media (Chesney & Citron, 2018). In the absence of any forensic science expertise, it is difficult for people to recognize 'deepfakes', and this has the potential to erode their trust in society and undermine democracy (Benjamin, 2019).

Figure 3: Evolution in AI generated images (Brundage, et al., 2018)

Simultaneously, it should be recognized that AI also has the potential to counter its own detrimental uses. Researchers from Harvard University and the MIT-IBM Watson Lab have developed the Giant Language Model Test Room (GLTR), which is a tool that can be used to spot AI-generated fake text, and researchers have also designed automatic systems that can analyze videos for the telltale indicators of a fake, assessing light, shadows and blinking patterns (Gehrmann, Strobelt, & Rush, 2019; Harwell, 2019).

Even as knowledge regarding AI remains open, the potential risks associated with the misuse of AI require further reflection on models of research publication. Some of these models are followed for research in the field of biotechnology and computer security. These include (Brundage, et al., 2018):

i) **Pre-publication risk assessment**: Some sensitive research areas concerning digital security or adversarial machine learning could be subjected to a pre-publication risk assessment process to better understand their safety implications in the long term.

ii) **Central access licensing models**: Consider the example of a trusted central service provider who allows the use of AI capabilities for different applications without divulging the details of the inner workings of the algorithm. This security-focused sharing model would allow the benefits of AI to be used without exposing us to risks of research being used for harmful purposes.

iii) **Safe sharing mechanism**: This could mean a system where research is shared with only a predetermined pool of trusted organizations. This would allow sharing of knowledge but only within a small group.

All these models have evident imperfections and raise ethical concerns regarding transparency and accountability, most importantly concerning limiting access to knowledge to a small group of people or institutions. Therefore, wider consultations are necessary to understand different stakeholders' concerns about AI, and to develop mutually acceptable and ethically responsible solutions to balance these competing needs.

6. CONCLUSION AND POLICY OPTIONS

Openness is an important attribute for publication of research and for ensuring transparency and accountability, as well as fair competition in the development and use of AI. This chapter highlighted several key trends with respect to openness and AI.

First, there is a general tendency towards sharing of research in AI by researchers in both universities and the private sector. A healthy increase in publications proves that researchers are actively engaging in open discussions and creating open repositories for AI knowledge.

Second, open data is an important element for the development of AI; it also facilitates transparency and accountability in the use of AI. However, there are several challenges with regard to private sector platforms that collect large amounts of data and do not share this with others, citing intellectual property issues, and there are also challenges for privacy and data protection.

Third, openness within AI systems concerning the decision-making process by algorithms is a technical challenge given the complexity and high dimensional nature of some AI technologies. Some solutions like auditing of algorithms and disclosure of their logic have been proposed for increasing transparency in algorithmic decision-making; however, they do not fully resolve AI's 'black box' problem.

Fourth, openness encourages innovation in markets and benefits consumers through lowering costs through competition; however, the race to capture a greater part of the market may lead to neglect of human rights in how firms use AI and to anti-competitive concentrations of market power.

Finally, openness can provide opportunity for malicious use of AI, although this risk can be mitigated without sacrificing the wider benefits it brings.

Some options for action to strengthen openness and transparency in AI are presented below:

Options for all stakeholders

- Develop norms and policies for improving openness, transparency and accountability in automated decisions taken by AI systems through methods such as ex-ante information disclosure and ex-post monitoring of automated decision-making.

- Facilitate open market competition to prevent monopolization of AI and follow the United Nations 'Guiding Principles on Business and Human Rights' for human rights based best practices for businesses.
- Promote open access research including through funding and support infrastructure for digital repositories and knowledge sharing.

Options for States

- Create open repositories for publicly-funded or owned data and research including the creation of platforms for open government data.
- Establish guidelines and policies for openness, transparency and accountability in the use and deployment of automated decision-making systems, including for use by the government.
- Support universities and technical training institutes to educate and train more students in AI and associated fields, thereby strengthening AI talent availability.

Options for the private sector, Internet intermediaries and technical community

- Develop norms for openness compliant with international standards and principles for human rights-based ethical practices in the development and use of AI.
- Ensure adequate safeguards are put in place with respect to open data in order to protect against the infringement of the right to privacy.
- Work together with other stakeholders to address the challenges posed by increasing openness, transparency and accountability of AI systems.

Options for academia

- Support the development of open data standards while safeguarding the privacy of individuals.
- Develop standards for interoperability between data sets while strengthening data commons and the availability of data for machine learning.
- Strengthen research efforts to enhance transparency and accountability in automated decision making by AI systems, including efforts to address AI's 'black-box' challenge.

Options for civil society

- Act as a watchdog in the use of automated decision-making by public authorities and the private sector and demand greater transparency and accountability in the funding, development and use of AI systems.

Options for intergovernmental organizations, including UNESCO

- Continue to foster the growth of open technology ecosystems by helping establish open data standards and open data repositories for AI through networks of partners and institutes and centres under the auspices of UNESCO (Category 2 Institutes and Centers).
- Leverage experience in developing the open access movement to support the development of ethical publication models that safeguard against the infringement of human rights due to misuse of openly available knowledge about AI.
- Study different approaches to algorithmic accountability and bring together stakeholders from different fields to build consensus around global best practices.
- Mobilize consortia focused on openness in AI to strengthen the movement.
- Develop guidelines for openness, transparency and accountability in the use of automated decision making systems.

ACCESS AND AI

3

CHAPTER 3: ACCESS AND AI

The possibility, and ability, for everyone to access and contribute information, ideas and knowledge is essential for inclusive knowledge societies. Access to information and knowledge can be promoted by increasing awareness of the possibilities offered by AI among all stakeholders. These possibilities include access to education, access to affordable or free and open-source software and to data, and access to hardware and affordable connectivity.

Box 17: UNESCO's position on Access for Internet Universality

Accessible to all as a norm for 'Internet Universality' raises issues of technical access and availability, as well as digital divides such as those based on income and urban-rural inequalities. Thus, the principle points to the importance of norms around universal access to minimum levels of connectivity infrastructure. At the same time, 'accessibility' requires engaging with social exclusions from the Internet based on factors such as literacy, language, class, ethnicity, culture, gender, disability and refugee status. Further, given that people access the Internet as producers of content, code and applications, and not just as consumers of information and services, the issue of all-around user competencies is part of the accessibility dimension of 'Universality'. This highlights UNESCO's notion of Media and Information Literacy, which enhances accessibility by empowering Internet users to engage with and create media content critically, competently and ethically.

As a general-purpose technology, AI has the potential to enhance efficiencies in existing products and services. It is expected to have an even greater impact by providing tools of analysis in areas of work and research that were hitherto unexplored because of human limitations, thereby serving as a new 'method of invention' (Cockburn, Henderson, & Stern, 2018).

In 2017, 70 per cent of the world's youth (15-24 years of age) were online; however, the regional disparities were significant (Broadband Commission, 2018). For instance, to varying degrees, the proportion of young people using the Internet has been assessed at 94 per cent in developed countries, 67 per cent in developing countries and only 30 per cent in Least Developed Countries (LDCs). The regional digital divide is stark when 9 out of 10 young people not using the Internet live in Africa or Asia-Pacific.

The gender gap in Internet use has increased from 11 per cent in 2013 to 11.6 per cent in 2017, i.e. the proportion of women using the Internet is about 12 percentage points less than the proportion of men using the Internet. This increase, despite global efforts towards gender parity in ICT access, means that only one out of seven women use the Internet compared to one out of five men in LDCs (Broadband Commission, 2018).

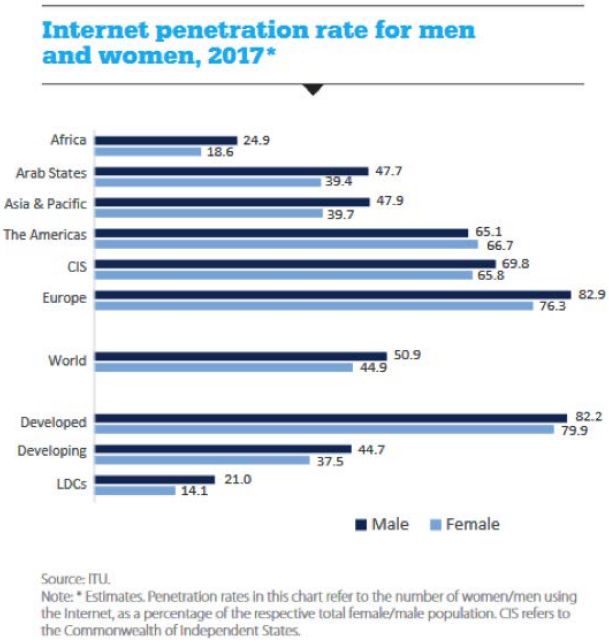


Figure 4: Internet penetration rate for men and women, 2017 (ITU, 2017)

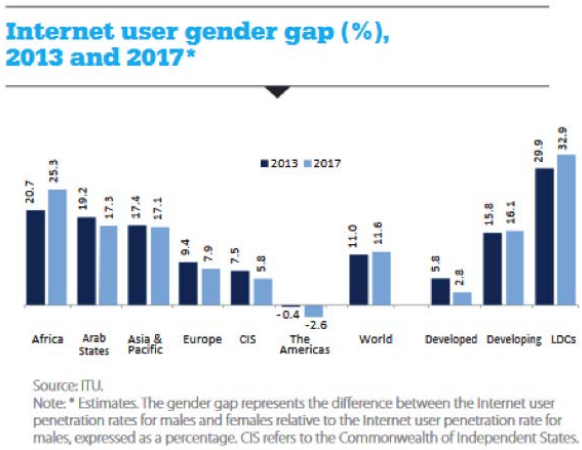


Figure 5: Internet user gender gap (%), 2013 and 2017 (ITU, 2017)

These concerns raised in the context of Internet use are as relevant to AI, which is increasingly being used in the technology mediations involved in people's interaction with the world. Accessibility, affordability and the ability to develop and use these technologies largely determine our social and market interactions.

Box 18: Mandate from World Summit on Information Society (WSIS)

The WSIS+10 Outcome Statement recognizes that "the sharing and strengthening of global knowledge for development can be enhanced by removing barriers to equitable access to information for economic, social, political, health, cultural, educational, and scientific activities and by facilitating access to public domain information, including by universal design and the use of assistive technologies" (UNGA, 2015). The WSIS mandate needs to be addressed through, among other actions, "strengthened enabling policy environments and international cooperation to improve affordability, access, education, capacity building, multilingualism, cultural preservation, investment and appropriate financing" (UNGA, 2015).

In the case of ICTs, the digital divide is an important consequence of differences in access. It has been defined as the "gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard to both their opportunities to access information and communication technologies (ICTs) and to their use of the Internet for a wide variety of activities" (OECD, 2001). With respect to AI, this divide may be understood in terms of access to four fundamental elements that enable the development and use of AI (Elsevier, 2018):

i) Access to research;

ii) Access to knowledge, education and human resources;

iii) Access to data for training of algorithms; and

iv) Access to connectivity and hardware.

The following sections describe the trends under each of these elements.

1. ACCESS TO RESEARCH

While overall openness is a feature of much AI research (see previous chapter), there are striking imbalances in regard to where this research is conducted. These in turn can affect what topics are prioritized as relevant, what data is used, and the extent to which others may find the knowledge to be of value.

The digital divide regarding the quality and the quantity of AI research is growing between and within countries. A challenge is whether AI can be used to help reduce the research imbalance. Most advanced economies in the world have a robust and vibrant research ecosystem that drives innovation and growth. Economist Jeffrey Sachs notes that: "there is a long term shift in the share of national income from labor to capital, including physical, human and intellectual capital" (Sachs, 2018). Countries with an edge in research and development in AI will be better equipped to deploy these technologies and have more trained human resources to translate research into applications. Elsevier's publications '2018 AI Index Report' and 'Artificial intelligence: How knowledge is created, transferred, and used' provide useful insights into the research trends in AI (Elsevier, 2018; Shoham, et al., 2018). However, they focus primarily on developed countries, and more efforts are needed to collect data on AI related research in developing countries. Analysis from (Shoham, et al., 2018) below gives some pointers to the issues of production, use and access to research from the point of view of researchers and users.

i) **Increase in AI publications**: Regional and country imbalances are evident in that 28 per cent of AI papers on Scopus[1] in 2017, Europe was the largest publisher, although Chinese publications on AI witnessed a 150 per cent increase between 2007 and 2017.[2]

ii) **Research impact**: Research citations provide a proxy metric to gauge the impact of the work. While Europe is the largest publisher in the field of AI, its research impact has been steady and at par with global average, whereas in the USA, AI authors are cited 83 per cent more than the global average. The quality of research on AI in China appears to have improved as Chinese AI authors were cited 44 per cent more in 2016 than in 2000.[3] This shows the difference

1 Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings.

2 An author's country affiliation is determined based on his or her primary organization, which is provided by authors of the papers. Global organizations will use the headquarters' country affiliation as a default, unless the author is specific in his/her organization description. For example, an author who inputs 'Google' as their organization will be affiliated with the United States, one that inputs 'Google Zurich' will be affiliated with Europe. Papers are double counted when authors from multiple geographies collaborate. For example, a paper with authors at Harvard and Oxford will be counted once for the U.S. and once for Europe. More details about the research methodology are available in (Shoham, et al., 2018).

3 There are several ways to measure research impact, and citation scores are just one of many methods to do so. UNESCO facilitates dialogue on appropriate research metrics with

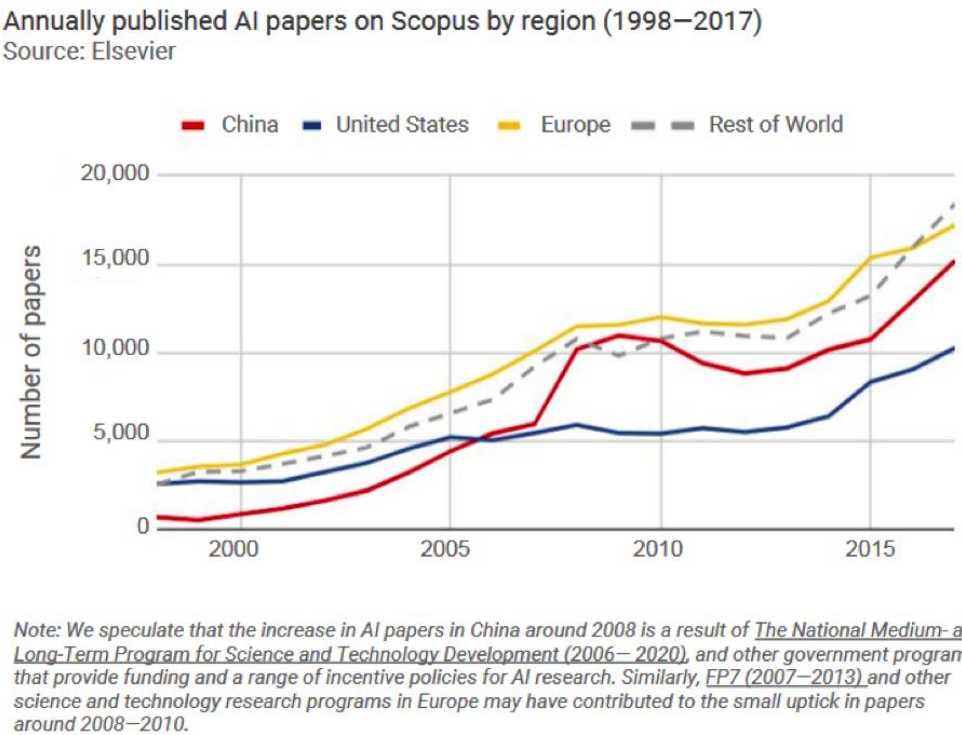


Figure 6: Annually published AI papers on Scopus by region (Shoham, et al., 2018)

in the quality of research that is being produced across different parts of the world and would have repercussions for AI divides between countries.

The fast pace of research and development in AI has made academic conferences an important avenue for dissemination of research findings and sharing of ideas. Later, expanded or updated versions of these papers may be published in academic journals, which typically takes a longer time as compared to publication in conference proceedings. Therefore, acceptance of papers in top AI conferences is another metric to gauge research impact.[4]

The number of submitted and accepted papers by region at the 2018 Association for the Advancement of Artificial Intelligence (AAAI) conference are presented in Figure 8. The United States and China accounted for more than 70 per cent of the submitted papers and had an acceptance rate of 29 per cent and 21 per cent respectively. India is the only developing country other than China that has an acceptance rate (at 22 percent) comparable to economically advanced countries.

multiple stakeholders as part of its Open Access programme. For more details, please see 'Policy Guidelines for the Development and Promotion of Open Access' (UNESCO, 2012).

4 UNESCO does not prejudice one way to measure research impact over the other. Acceptance of papers in top AI conferences is used here with the understanding that it is a single and imperfect indicator of research impact, and caution is needed about extrapolating too much from it.

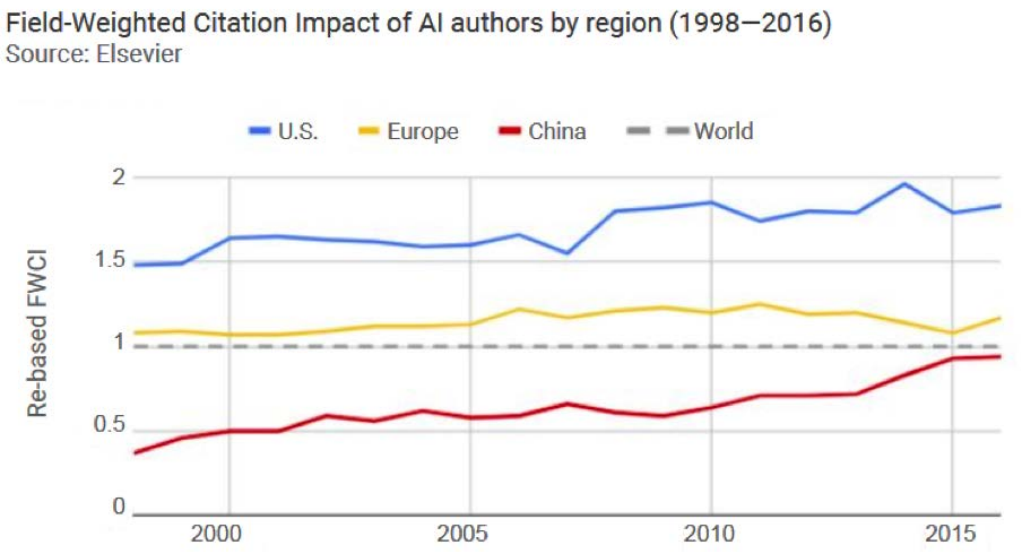


Figure 7: Citation impact of AI authors by region (Shoham, et al., 2018)

As previously noted, access to quality research is an important pre-requisite for countries to fully leverage AI to their benefit and for their development. While it is true that global technology firms work across the world and spread technological know-how, vibrant local research and innovation ecosystems would foster development of local solutions using AI and support sustainable development of science, technology and innovation in the country. Based on these findings, it is clear that a small group of countries is leading in both quantity and quality of AI-related research. Developing countries in Africa and other regions (barring China) have a limited presence in AI research, although there are promising initiatives in Africa, as discussed in the chapter on AI in Africa.[5] National policies and international support for AI-related research would help in strengthening the research output in developing countries and provide a base for local innovation to grow on and respond to local challenges.

5 The conclusion is based on the findings of two recent AI reports: Shoham, et al., 2018, and Elsevier, 2018. This is by no means exhaustive, and more work is needed to closely examine AI research in developing countries.

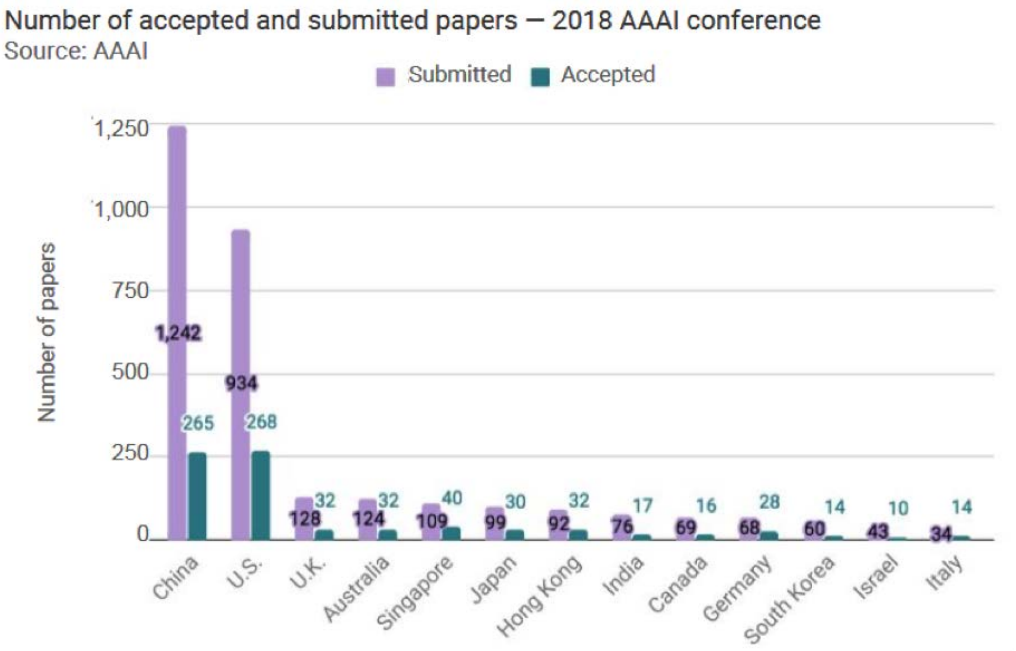


Figure 8: Number of accepted and submitted papers at the 2018 AAAI Conference (Shoham, et al., 2018)

2. ACCESS TO KNOWLEDGE, EDUCATION AND HUMAN RESOURCES

Access to human resource educated and trained to research, develop and apply AI is an important pre-requisite for development of AI in a country. Shortage of talent is a major barrier to the development of AI (Fu, 2018). Depending on the methodology of the study to which one refers, the estimates on the worldwide number of experts working in AI-related research or industry ranges from 10,000 as per the New York Times to 300,000 as per the China based technology firm Tencent (Metz, 2017; Vincent, 2017). McKinsey and Company conducted a global survey to find out how embedded AI capabilities are in different companies.[6] The data presented in Figure 9 shows that some AI capabilities embedded in company functions are comparable across regions, while others like autonomous vehicles are more deeply embedded in developing markets versus North America and Europe. Figure 10 disaggregates the differences within developing countries.

6 In their report, McKinsey defined nine AI capabilities: natural-language text understanding, natural-language speech understanding, natural-language generation, virtual agents or conversational interfaces, computer vision, machine learning, physical robotics, autonomous vehicles, and robotic process automation.

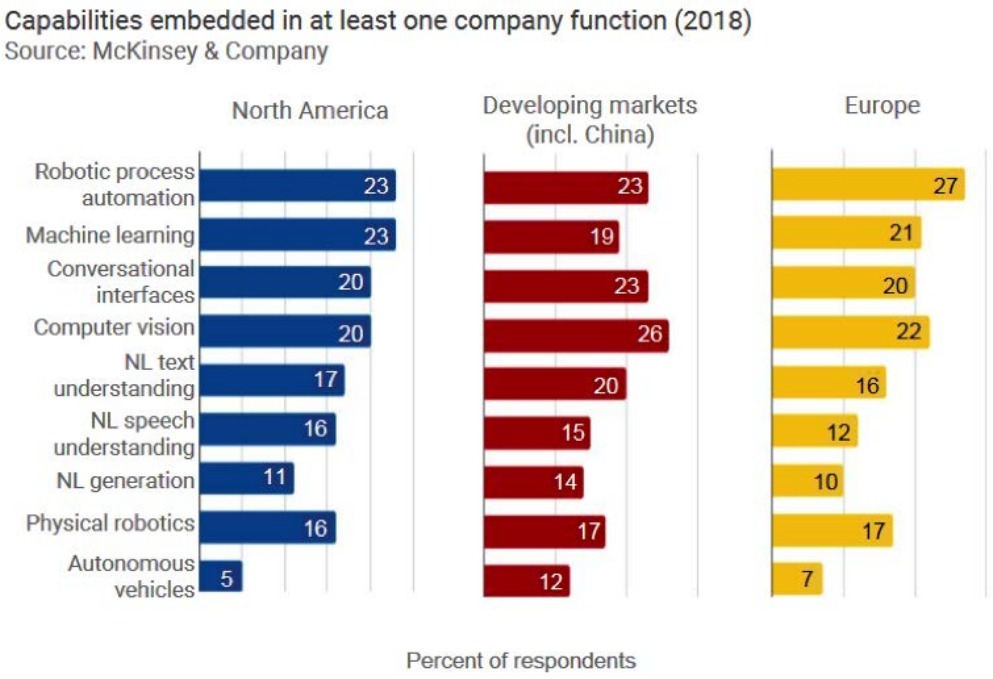


Figure 9: Difference in AI capabilities between different groups of countries (Shoham, et al., 2018)

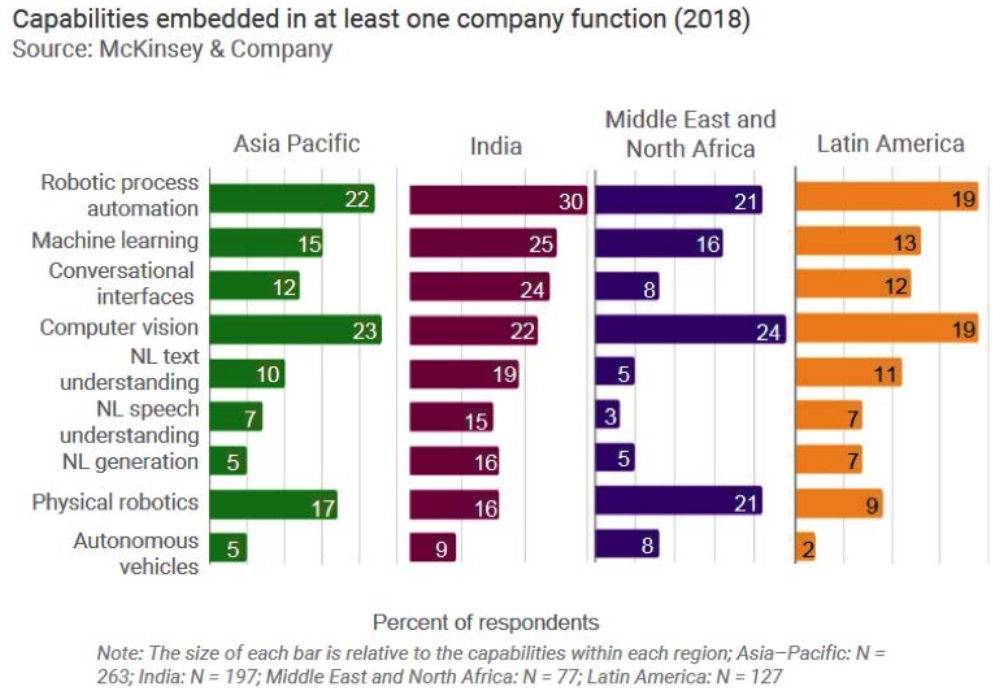


Figure 10: Difference in AI capabilities between different regions (Shoham, et al., 2018)

Jobs in AI have some of the highest remunerations in the tech industry (Stanford, 2018). Figure 11 shows multifold increase in the job openings in AI fields like Deep Learning. There is a monetary incentive from the labour market to attract more people towards AI research. However, the fact that most LDC's cannot easily compete in accessing costly talent is a barrier to accessibility.

Some research centres draw in global talent, with the effect of brain-drain from some countries but also from academia to tech companies (Kunze, 2019; Hao, 2019; Sample, 2017b). In fact as the demand for AI researchers grows, some universities are finding it difficult to retain AI researchers in academia to train the next generation of students (Kwok, 2019; Sample, 2017a).

At the same time, a shortage of skilled individuals is driving companies to i) set up global research centers to tap into local AI talent in tech hubs, ii) provide in-house skills training and courses in data analytics, iii) initiate crowd-sourced solutions to improve algorithms and find bugs and spot talent through open competitions like Kaggle,[7] and iv) develop AI tools that can be used by non-tech companies without having to invest in AI human resources (Boyd, 2017). These steps do increase access to gaining and improving AI skills, although developing countries are less likely to benefit.

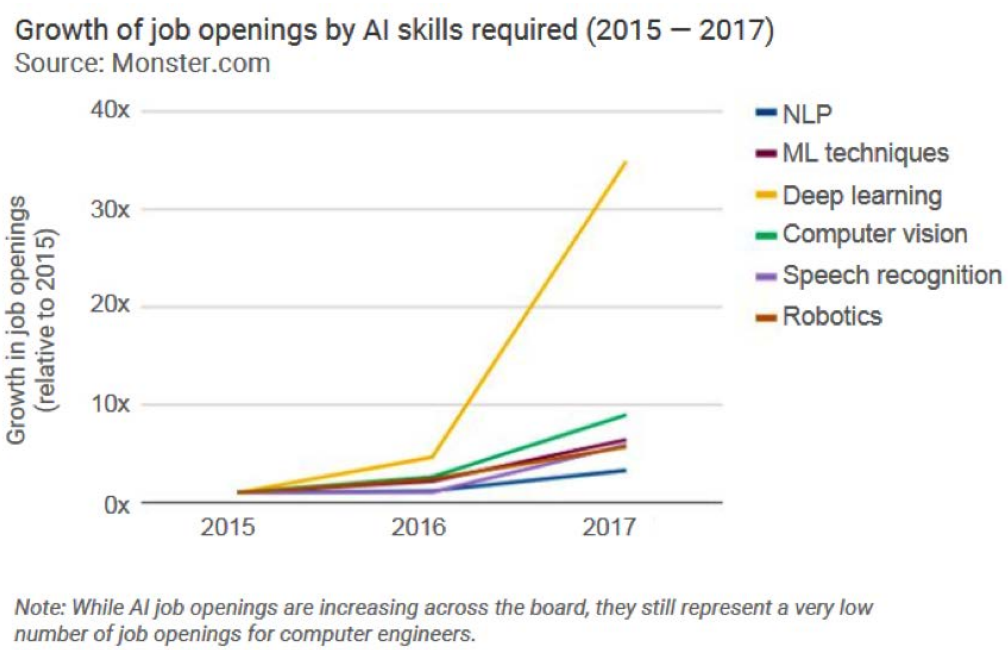


Figure 11: Growth of job openings by AI skills required (Shoham, et al., 2018)

7 Kaggle is an online platform that describes itself as 'the place to do data science projects'. It allows companies to post problems in the form of competitions with or without monetary reward or job offers, hosts datasets, discussion forums and online learning tools for AI related topics.

In order to meet the demand for AI education, universities are offering more AI-related degrees and online courses.[8] Examples of such include a Master of Science in AI at Imperial College London and an online course on Machine Learning at Columbia University (Value Colleges, 2019; Marr, 2018). In China, the Education Ministry has also approved of a programme to introduce AI as an undergraduate major at 35 universities (Fang, 2019). Overall, the average enrollment in introductory AI and Machine Learning courses has grown by 3 or 4 times between 2012 and 2017 across several universities around the world (Shoham, et al., 2018).

Further, the lack of gender diversity is an important question mark on how inclusive the tech industry and computer science academic departments are, as these have the potential of perpetuating historical biases through the design of AI systems (Paul, 2019; Simonite, 2018b). More than 75 per cent of AI professors at top schools in the United States are men (Shoham, et al., 2018). Another report showed that gender diversity gap in AI research, with only 13.83 per cent of authors in arXiv being women (Stathoulopoulos & Mateos-Garcia, 2019).

There is growing awareness of this problem. However, much additional research is needed to understand access to AI training and education and current level of human resource availability, especially in regard to developing countries. The issue of educational content available in multiple languages and certified to be useful and of high quality also merits attention.

AI's accessibility to all depends on the competencies of the broad public to understand its significance and their engagement with it. Yet Media and Information Literacy is far from universal, and even further from empowering non-specialists with knowledge about AI issues.

3. ACCESS TO SOFTWARE AND DATA FOR TRAINING OF ALGORITHMS

The availability of free and open-source software was discussed in the previous chapter. At the same time, attention should be kept on the issue as proprietary software evolves and the extent to which this is affordable. Nevertheless, access to data appears to be a current issue. Open data was touched on the previous chapter, where it was noted that much data remains under closed ownership.

While data are often called 'the new oil', unlike oil, they are potentially a non-rivalrous good, which means that the use of data by one person does not prevent its use by

8 To understand the AI education landscape better, there is a need to collect data on the number of institutions which offer degrees in AI and what proportion of AI courses are conducted online.

another. In the case of AI, access to data is essential for training algorithms and for their usefulness in large scale application. Incumbent technology firms collect large amounts of user data on their platforms and use it to train algorithms for improving their products and services (The Economist, 2017). In the absence of access to data, new firms face potential entry barriers in challenging the entrenched market actors. Even academic institutions have cited lack of access to data as one of the barriers impeding research. In order to access data, collaboration with data-rich tech firms is one option, although this depends very much on the perceptions of advantage-accruing to the companies (Sample, 2017a). This issue is exacerbated in most developing countries.

Access to data and the associated legal and economic dimensions are subject of important discussions around the world (European Commission, 2017). However, as the nature of the industry evolves, there is some recognition of the non-rivalrous nature of certain data and the benefits of developing data commons in the form of open data repositories (Tonetti, 2018). Examples of such datasets and repositories for Machine Learning include:[9]

- University of California Irvine's Machine Learning Repository that is a collection of databases, domain theories, and data generators used by the machine learning community for the empirical analysis of machine learning algorithms;
- Princeton University hosts the WordNet, which is a lexical database of English. WordNet groups nouns, verbs, adjectives and adverbs into sets expressing a particular concept. Therefore, by linking group of words to distinct meaning, it is useful in natural language processing;
- ImageNet is a database of images linked to WordNet. It provides about 1,000 images for each word group-meaning set present in WordNet and has over 14 million images in total. The project supports the development in the image and vision research;
- Kaggle provides open datasets in diverse fields ranging from biology, health and education to sports and stocks markets;
- Open Data Monitor provides a list of open databases available across the EU among several other data repositories that make big data sets available for machine learning.

While these data repositories are being developed and supported across the world, what needs further interrogation is the issue of geographical and other biases that may reduce the value for other regions of the world. Further, much is required to strengthen interoperability of data by developing standards for data storage, classification and sharing (European Commission, 2017). A layered framework based on functionality and interoperability for data commons, developed by researchers at the Berkman Klein Center, was presented at the 2018 AI for Good Summit organized by

9 This is a sample list of popular datasets and data repositories and is not meant to be an exhaustive list.

the International Telecommunication Union (ITU). It offers conceptual clarity for the development of data commons.

Figure 12: Data Commons Framework (Goldstein, Gasser, & Budish, 2018)

The model is composed of narrow and broad data commons, each constituted by multiple sub layers (Goldstein, Gasser, & Budish, 2018).

Narrow data commons is composed of the following three layers:

i) A technical layer, i.e. the infrastructure used to store data in 'cloud', government servers, or decentralized ledgers;

ii) A data layer that takes different forms like qualitative/quantitative, structured/unstructured, ordinal/nominal, and discrete/continuous; and

iii) A layer that assigns formats and labels for this data to be intelligibly analyzed.

Interoperability with and between the different layers is addressed through the development of common technical standards and data standards for both hardware and classification of data.

Broad data commons provide an interface between the core functions as described in the narrow data commons and the society. These include:

i) Organizational practices for leveraging data commons to encourage collaboration and multi-stakeholder participation;

ii) The legal and policy concerns surrounding privacy, access, openness and human rights arising out of data commons; and

iii) The involvement of humans in the development and preservation of other layers through improved inclusion, education and literacy.

Interoperability in the broad data commons is ensured through development of shared knowledge and normative understanding in society.

Expanding access to data is crucial for the development of machine learning and AI. Yet even as data commons develop, other concerns related to representativeness of the data, discrimination and openness need to be addressed. These issues have been discussed in depth in the chapters on human rights and openness.

4. ACCESS TO CONNECTIVITY AND HARDWARE

Access to affordable broadband connections and to computer hardware for processing and storage of data is the fourth component essential to the development of AI. Advances in AI have been possible because of the availability of higher computing power and its customization to computational processes of machine learning and deep learning algorithms. For instance, IBM achieved a breakthrough in AI performance when its software was trained on an online advertising dataset with over four billion training examples in just 91.5 seconds using optimized hardware. The training time improved by over 46 times the previous best achieved time using TensorFlow on Google's Cloud Platform that trained the same model in 70 minutes (Parnell & Dünner, 2018).

The processor is at the heart of AI operation, as it performs the calculations on the data based on the instructions in the algorithm. The compatibility between the kind of task and the processor type is an important determinant of the efficiency of AI. For instance, a CPU (Central Processing Unit) is suitable for performing a few complex calculations very efficiently but it is not equipped to handle a very large number of calculations even if they are simple. Machine Learning essentially belongs to the latter category. It performs simple calculations following what computer scientist Andrew Ng termed the 'lazy hiker principle', which is to instruct a hiker to continue going downhill until the point he or she cannot go down anymore (Wilson, 2011). Machine Learning finds its answers using a similar approach of trial and error. However, to reach the solution, it has to perform a very large number of very simple calculations. The processor that is suitable for this task is the GPU (Graphics Processing Unit) that has been in use for gaming applications since the 1970s (Algorithmia, 2018).

Application Specific Integrated Circuits (ASICs) are designed to perform specific tasks. For instance, Google has developed Tensor Processing Unit (TPU) for machine learning on TensorFlow. These ASICs can be even more efficient than GPUs (Algorithmia, 2018).

These processing units require semiconductor devices, therefore an analysis of glo-

bal import and export of semiconductor devices provides a proxy for understanding the disparity in access to hardware for AI. Figure 13 and 14 show the divide in trade of semiconductor devices. According to the Observatory of Economic Complexity, semiconductors[10] are the 24th most traded product (The Observatory of Economic Complexity, 2019). Out of a total trade of USD 88 billion, Asia accounts for 80 per cent of the exports followed by Europe and North America. In the case of imports, Asia is in the leading position followed by Europe and North America again. Africa accounts for only 0.90 per cent of global semiconductor imports (The Observatory of Economic Complexity, 2019).[11] Therefore, a wide geographical disparity exists in access to hardware necessary for AI development. The importance of the semi-conductor industry is further underlined by the significant policy support being offered by the government in countries like China to develop self-reliance in production of semi-conductor devices (Kharpal, 2019). Gaps in access to semiconductor devices needs to be bridged for development of AI to not be concentrated in some parts of the world.

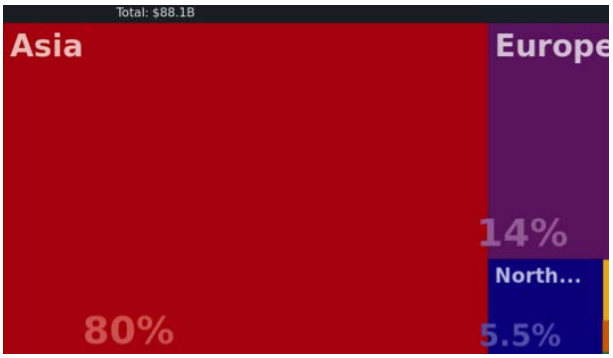


Figure 13: Exporters of semi-conductor devices by continent. Semiconductor trade is shown as a proxy for computing hardware (The Observatory of Economic Complexity, 2019)

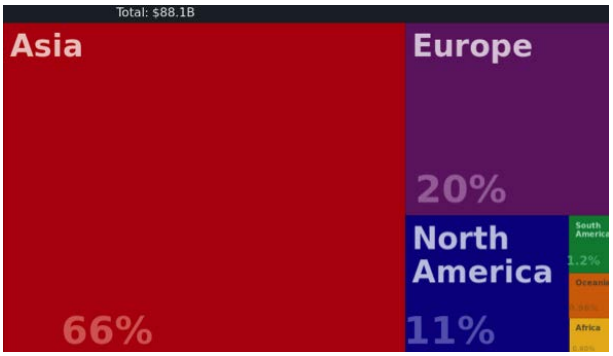


Figure 14: Importers of semi-conductor devices by continent. Semiconductor trade is shown as a proxy for computing hardware (The Observatory of Economic Complexity, 2019)

10 Semiconductor devices are also known as diodes, transistors, photovoltaic, photosensitive, mounted piezo-electric crystals. Semiconductor devices are a 4-digit HS92 product.

11 The top exporters of semiconductor devices are China ($28.5B), Japan ($8.38B), Other Asia ($7.74B), Malaysia ($7.55B) and Germany ($6.1B). The top importers are China ($16.3B), Hong Kong ($12.6B), Germany ($6.24B), the United States ($5.81B) and South Korea ($4.44B) (The Observatory of Economic Complexity, 2019).

However, since data can be stored and processed in the cloud, access to hardware may be provided through cloud computing services. This depends on the availability of affordable and reliable broadband connections, that in turn are a function of government policies and the extent of open market competition between providers. The Alliance for Affordable Internet has been monitoring these matters, which are relevant not only for development of AI, but for societal uptake of implementations that depend on an Internet connection.

With regard to actual storage and processing, technology companies such as Google, Amazon, Microsoft and IBM among others offer low cost cloud computing services. The user can select the kind of processor needed for the task at hand i.e. a CPU, GPU or and ASIC. The decentralization of computational resources has converted the large fixed costs of maintaining costly hardware to a variable cost, where firms or individuals just pay for computing based on their use (Varian, 2019).

The availability of computational resources on demand strengthens access to AI. Some governments are supporting startups by making computational resources available to them. For instance, the UK government has set up a 'Machine Intelligence Garage' to give startups access to computing resources and bridge the access divide (Shead, 2017). Similar policies may be considered by developing countries to encourage AI startups in their jurisdictions.

5. CONCLUSION AND POLICY OPTIONS

Access to AI and other advanced ICTs is important for bridging new digital divides that we are witnessing between and within countries. Several important trends are visible.

First, countries with advanced economies and well-established research centers are leaders in global research in AI. Developing countries need to strengthen their research capacities for development and deployment of AI. Given that this will have to be accomplished in the context of the existing digital divide, developing collaborative research practices could be mutually beneficial in expanding machine learning datasets and facilitating technological transfer.

Second, there are differences in access to AI education and human resources that can exacerbate the digital divide between countries and between genders. In order to strengthen access to AI knowledge and education, there is a need to create quality content that is available in multiple formats and languages, and that can be accessed in different regions, including places with limited telecommunication infrastructure. Active engagement between educational institutes and the private sector would help to stimulate local AI ecosystems and provide employment opportunities. Such engagement should be critical and research should not be instrumentalist and limited to the technical aspects of AI, but also explore ethical and regulatory issues in the practice of AI.

Third, access to data needs to be strengthened through open data repositories to ensure free access to data for training algorithms. However, issues regarding data representativity, interoperability and data standards need further attention. Finally, access to hardware for computing is necessary for running AI algorithms, this aspect requires greater effort from governments and the private sector to make suitable processing units available, even though affordable broadband offers a partial work-around.

Options for all stakeholders

- Work to reduce digital divides, including gender divides, in AI access, and establish mechanisms for continuous monitoring of the differences in access.
- Ensure that individuals, groups and countries that are least likely to have access to AI are active participants in multi-stakeholder dialogues on the digital divide by emphasizing the importance of gender equality, linguistic and regional diversity as well as the inclusion of youth and marginalized groups.

Options for States

- Strengthen the infrastructure and support needed for AI-related research and development at universities and research centres.
- Encourage and support the acquisition of coding skills and computer science literacy for citizens through proactive policies for education, technical and vocational training, including for lifelong learning.
- Ensure policies that provide for affordable broadband access and avoid interferences with connectivity such as Internet shut-downs, throttling or arbitrary filtering and blocking.
- Strengthen access to AI-specific computational hardware, including through funding support and providing need-based access to centralized computing resources.
- Collect data to understand access to AI education and current level of human resource availability, especially in developing countries.
- Ensure that educational courses, including for Media and Information Literacy which includes an understanding of AI, are available in multiple languages.

Options for the private sector

- Strengthen access for citizens to affordable connectivity, software and hardware needed for running AI programs.
- Collaborate with universities and research centres, including through student training, doctoral research grants, sharing data and computing resources for research and development.
- Strengthen gender diversity in AI research both in academia and the private sector.

Options for academia

- Improve access to AI technology and data for learning and classification through the creation of research repositories and open access publishing.
- Strengthen access to AI knowledge by offering high quality open educational resources in multiple languages and formats accessible by persons with disabilities.
- Update university curricula dynamically with state of the art research developments and methodologies, including through regular education and skills needs assessment in partnership with the private sector and other stakeholders.

- Create and strengthen mechanisms for research collaboration, mobility of researchers and mentorship opportunities for students between universities across the world, with special focus on North-South and South-South exchanges, and on gender parity.

Options for civil society

- Support the development of localized AI knowledge and resources in formats and languages that render the information about AI accessible to all, especially women, persons with disabilities and others who historically have limited access to ICTs and technology skills.
- Participate in crowd-sourcing projects for the creation on data commons.

Options for inter-governmental organizations, including UNESCO

- Support Member States in enhancing AI research capacity through trainings, education policy development, academic exchanges and the UNESCO Information for All Programme.
- Integrate discussion of AI issues into relevant events such as international days around press freedom, disability, and universal access to information, and draw in UNESCO networks such as the UNITWIN/UNESCO Chairs Programme, the UNESCO Chairs in Communication (Orbicom), the Global Alliance for Partnerships on Media and Information Literacy (GAPMIL), and the Global Alliance on Media and Gender (GAMAG), as well Category 2 Institutes and Centres, NGOs, IFAP National Committees and UNESCO National Commissions.
- Champion discussions for the development of data commons with representative data by facilitating global participation and the construction of common technical and data standards, through a process of building shared knowledge and normative understanding about the interaction of data with society and information ethics.
- Facilitate expansion of multilingual access to AI education and support the development of good quality Open Educational Resources.
- Leverage UNESCO's academic networks and Category 2 Centers to facilitate AI research exchange, including through the development of repositories for sharing research.

MULTI-STAKEHOLDER APPROACH FOR AI GOVERNANCE

4

CHAPTER 4: MULTI-STAKEHOLDER APPROACH FOR AI GOVERANCE

Given AI's potential impact across all sectors of societies, each of these stakeholder groups from governments, to the private sector, the technical community, academia, civil society and individual users, has distinctive roles and responsibilities. AI governance should count on truly co-operative mechanisms in order to develop effective and relevant principles, norms, rules, decision-making procedures, and programs for AI. This is pertinent at the national level, but given the cross-border dimensions of AI and its elements, the regional and international dimensions call for serious attention as well. AI is too complex and too important to be decided by any single constituency in isolation. Its formal and informal regulation and evolution call for wide and inclusive consultation processes.

Against this background, the goal of multi-stakeholder participation is to improve the inclusiveness and quality of decision-making by including all groups who have an interest in AI and its impact on wider social, economic and cultural development in open and transparent decision-making processes. This kind of participation and cooperation can keep all stakeholders well updated on the fast development of the technology and serve to create dialogue that can build consensus among all stakeholders, at least at the level of principles and ethics, but also extending into more specific forms of governance where appropriate.

This chapter is divided into three sections. First, it traces the development of a robust multi-stakeholder environment for the governance of the Internet. Practices from Internet governance can inform some of the discussions around multi-stakeholder engagement concerning advanced ICTs like AI. Second, the chapter stresses the strong need for a multi-stakeholder approach based on the concern around the complexity of decision-making and balancing interest around technology and its impact on society. Third, the chapter shares some practices and values that guide the multi-stakeholder engagement process. Finally, the chapter concludes with options for future action that can be taken by different stakeholders.

Box 19: UNESCO position on a Multistakeholder Approach for Internet Universality

The participatory, and specifically multistakeholder, dimension of 'Internet Universality' facilitates sense-making of the roles that different agents (representing different sectors, as well as different levels of social and economic status, and not excluding women and girls) have played, and need to continue to play, in developing and governing the Internet on a range of levels. Participation is essential for realizing the value that the technology can have for peace, sustainable development and poverty eradication. In bridging contesting stakeholder interests, participative mechanisms contribute to shared norms that mitigate abuses of the Internet. 'Universality' here highlights shared governance of the Internet.

1. CONTEXTUALIZING AI WITHIN MULTI-STAKEHOLDER INTERNET GOVERNANCE

Though AI exists as a field distinct from the Internet, this publication also recognizes the underlying connections such as connectivity, data collection and processing. In short, AI's application and development are widely embedded within, and intertwined with, the ecosystem of the Internet and its social, political and economic evolution. When considering the governance of AI, it is thus highly pertinent to look to the framework of global Internet governance that is driven by multi-stakeholder participation.[1]

The successful development of the Internet since its inception is characterized by this multi-stakeholder participation to varying degrees. It shows that the extent to which stakeholders do or can effectively participate is determined by a number of factors, including the extent of their awareness, interest, concern and knowledge; their level of agency or responsibility for outcomes, and the nature of involved consultative and decision-making processes. When such participation is institutionalized, and even legally confirmed in some instances, the results are more sustainable.

As indicated by Internet history, multi-stakeholder approaches are important to both promote the developmental potential of the Internet and to maintain its universal character. All stakeholders, from governments, to the private sector, to the technical

1 This discussion draws extensively from UNESCO-commissioned research on multistakeholder participation (Van der Spuy, 2017)

community, to intergovernmental organizations (IGOs), to civil society, to academia are impacted by the development of the Internet, which underpins the logic that they should be inclusively involved in policymaking processes. In order to tackle the complexity of Internet governance issues, an open and inclusive multi-stakeholder approach has been adopted at national and international levels. Some of the key events that influenced such a path have included:

- **World Summit on the Information Society (WSIS), 2003-2005**

The *Tunis Agenda for the Information Society* encouraged 'the development of multi-stakeholder processes at the national, regional and international levels to discuss and collaborate on the expansion and diffusion of the Internet as a means to support development efforts to achieve internationally agreed development goals and objectives, including the Millennium Development Goals (WSIS-05/TUNIS/DOC/6(Rev. 1)-E art. 80 & 87, 2005).

The *Tunis Agenda* agreed on a 'working definition' of Internet governance as "the development and application by governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programs that shape the evolution and use of the Internet."

- **World Summit on the Information Society (WSIS) + 10 Review**

The United Nations General Assembly, in its ten-year review of WSIS outcomes in 2015, reaffirmed 'the value and principles of multi-stakeholder cooperation and engagement. It recognized that effective participation, partnership and cooperation of Governments, the private sector, civil society, international organizations, technical and academic communities and all other relevant stakeholders, within their respective roles and responsibilities, especially with balanced representation from developing countries, has been and continues to be vital in developing the Information Society (UNGA A/RES/70/1 para. 16 & 17, 2015).

- **UN Sustainable Development Goals (SDGs)**

The United Nations' 2030 Agenda for Sustainable Development also calls for "multi-stakeholder partnerships [...] that mobilize and share knowledge, expertise, technology and financial resources, to support the achievement of the sustainable development goals in all countries, in particular developing countries" to be established (UNGA A/RES/70/125 para.3, 2015).

Essentially, the multi-stakeholder approach stresses the importance of dialogue as a way to balance interests, aggregate wisdom, and build consensus and legitimacy. This is relevant to the Internet broadly, and to advanced ICTs like AI in particular. Multi-stakeholder participation works to ensure equitable access to different interests. This means taking decisions through the interaction of these participating interests, thus allowing digital technologies to maintain a universal character and utility. Such process helps to highlight and realize the development potential of AI for serving human rights and the achievement of SDGs.

Box 20: Report *'The Age of Digital Interdependence'* (UN Secretary-General's High Level Panel on Digital Cooperation, 2019)

"We believe that autonomous intelligent systems should be designed in ways that enable their decisions to be explained and humans to be accountable for their use. Audits and certification schemes should monitor compliance of artificial intelligence (AI) systems with engineering and ethical standards, which should be developed using multi-stakeholder and multilateral approaches. Life and death decisions should not be delegated to machines.

We call for enhanced digital cooperation with multiple stakeholders to think through the design and application of these standards and principles such as transparency and non-bias in autonomous intelligent systems in different social settings."

(UNSG, 2018)

Recourse to multi-stakeholder mechanisms promises better and more inclusive AI governance by enhancing ownership and transparency, and helping decision-makers to take into account diverse viewpoints and expertise. The quality of results and legitimacy thus ascribed make for better governance of the enormous complexities and interdependencies of the Internet and AI (as well as other advanced digital technologies).

2. COMPLEX DECISION-MAKING AND BALANCING INTERESTS IN AI DEVELOPMENT – THE NEED FOR MULTI-STAKEHOLDER PARTICIPATION

Given an inherently limited ability to forecast the social effects of technology, policymakers must make decisions with limited information, whether under uncertainty or ignorance. However, it is important to understand and assess the initial goals embedded in technology development and application. For instance, the green revolution was successful in achieving its technical objective of developing high yielding variety of seeds, but it has not succeeded in regard to the social goal of ending hunger. Collingridge argues that "our understanding of the physical and the biological world in which we live is extremely deep, and provides us with the means for the produc-

tion of all kinds of technical marvels; but the appreciation of how these marvels affect society is perilous" (Collingridge, 1980). This disparity between our understanding of technology and its effects on society often leads to demonization of technology. It is therefore no surprise that the question about AI replacing humans is received with widespread fear of the technology itself (Solon, 2017).

One area where this is evident is the question of the future of work and working life in the age of AI-induced automation, where technological determinism implies that employment will be inevitably affected by advanced ICTs. But, as per the First Regular Session Report of the UN Chief Executives Board for Coordination in May 2019, it is a policy decision that can ensure that good job growth outpaces the destruction of jobs and creates a balance between automation and new tasks by using technology to augment work, rather than to replace workers. The report thus proposes investment in the reskilling and upskilling of the population, and making lifelong learning a natural part of an individual's working life. It further highlights the need to develop a modern social safety net, along with designing technology for the benefit of all and establishing agile governance structures that can respond to challenges and adapt policy and regulations faster (CEB, 2019). In particular the report highlights the potential for:

i) Equipping citizens with education and skills to be 'digital citizens'.

ii) The changing nature of work life, including income security and algorithmic biases at work.

iii) Re-training displaced workers through new modes of training and development.

iv) Sharing benefits of AI equally in society to avoid exacerbating income inequality.

v) Tailoring education curricula in skills that are need to an AI enabled future.

Nevertheless, even with these measures, other potential impacts of the social and economic evolution and use of the technology are unpredictable. In this case, decisions will have to be made under uncertainty and ignorance. Thus, in the case of AI, developers, policymakers and regulators would be remiss if they made decisions without continuously monitoring potential and unexpected consequences. The trade-off required for effective decision-making under 'ignorance' is "the ease with which mistakes can be detected and eliminated and the costs imposed by the mistake" (Collingridge, 1980). Therefore, the guiding framework of decision-making process requires decisions to be correctable, i.e. when a mistake or unexpected outcome is discovered it should be easy to correct. A general yardstick for such decisions includes:

- Cost effective and efficient monitoring to ensure that unpredicted consequences can be discovered quickly,
- Selecting options with low error costs,

- Low response time for correction after discovery of unanticipated problems, and
- Low cost of applying the remedy. Therefore, options with low fixed cost and a higher variable cost should be given preference over the ones with a higher fixed cost.

Against this background, it is apparent that multi-stakeholder engagement is all the more necessary for shaping values in decision-making about AI, and for monitoring and making changes, at national as well as supra-national levels.

One area where this is important is in regard to risks resulting from competition in the development of AI-related products and services. In a competitive world market, the first mover often grabs the largest share of the market, meaning a concentration of power that can be abused, including measures that erode the principles of Openness and Access. Similarly, firms using AI technologies, especially under competitive conditions, should be governed by rules that ensuring that human rights are respected continuously as their products are used in the market. In all these cases, monitoring of unintended consequences is critical.

When applying these principles to AI governance and mitigating the risks of decision-making with limited information about future effects, it is crucial to have an inclusive multi-stakeholder participation or cooperation process in place. Unlike other traditional policy-making processes, multi-stakeholder practices can highlight principles such as openness, transparency, and the broad-based collaboration and equal participation, that are relevant to the best possible foresight and monitoring under the circumstances. Multi-stakeholder processes can also help inform the development of operating frameworks themselves which different actors can use to make different decisions as issues arise.

This calls for the effective participation, partnership and cooperation of stakeholders in the collective development and use of the AI. Anyone or any entity with a legitimate or bona fide interest in an issue brought about by the AI development can be considered as a relevant stakeholder. At the global level, effective participation also means the inclusion of developing countries and regions in the process, as well as less resourced groups such as civil society and academia.

As noted, multi-stakeholder participation is posited as a method to prevent the domination of the Internet and other new technologies by one constituency at the expense of another. This is true whether advanced ICTs are captured by various state actors (e.g. security rather than commerce, or taxation rather than digital economy policy, etc.), or by interstate organizations, or by private sector interests (with varying commercial interests and business models which may violate rights to privacy or expression etc.). It also applies nationally and internationally. The outcome is that decision-making comes to reflect aggregated interests, rather than single interests, and that there are collective stakes in regard to monitoring and changing unanticipated outcomes.

3. AI GOVERNANCE BY A MULTI-STAKEHOLDER APPROACH: PRACTICES, VALUES AND INDICATORS

There is ongoing discussion about various multi-stakeholder models of Internet governance, but this section tends to share some practices and values that have successfully shaped multi-stakeholder engagement in the case of the Internet - and which can be a useful template to think about AI and its governance issues.

The UNESCO commissioned study *What if we all governed the Internet?* assesses the elements underpinning a number of digital co-operations based on a multi-stakeholder approach. These include the Internet Governance Forum (IGF) Best Practice Forum on Gender and Access, the Internet Corporation for Assigned Names and Number (ICANN) policy development process, as well as the national mechanism of KICTANet in Kenya, and the drafting of the Brazilian Civil Rights Framework for the Internet (Marco Civil) (Van der Spuy, 2017).

The successes and failures in these experiences show that multi-stakeholder approaches need to exhibit certain values if they are to be effective in shaping norms, developing consensus, and enabling decision-making concerning governance. Strain is increased when there is unilateral decision-making, or in public-private partnerships that exclude civil society input into governance issues. Private sector stakeholders' lack of participation, or less transparent participation, also weakens the legitimacy and efficiency of multi-stakeholder initiatives. Civil society inability to participate fully, due to resource constraints, often means that decision-making neglects human rights concerns.

As outlined in the above UNESCO study, multi-stakeholder mechanisms should align with certain values if they are to be optimum contributors to governance:

- **Inclusive** – Closely related to the ROAM principles of Accessibility to all and Openness, inclusivity encapsulates the need to overcome barriers to accessible participation and to dedicate sufficient funding and capacity-building efforts to promote the participation of a rich diversity of stakeholders. Special provisions should therefore be made for stakeholders that tend to be underfunded and underrepresented, such as marginalized communities, women, youth, small business entities, and/or civil society participants from developing and/or Global South regions.
- **Diverse** – Ensuring that multi-stakeholder processes can benefit from different viewpoints in addressing the complex and diverse stakeholder concerns inherent in the challenges posed by the issues brought about by digital technology. For example, decisions about AI-systems used for automatic decision-making should involve human rights organizations in order to ensure that possible discri-

minatory dimensions are taken into account.

- **Collaborative** – Stakeholders should agree on common norms to guide working methods, including the extent of transparency, flexibility required, ways of making decisions, and means to promote and protect the participants' safety and rights when participating in a multi-stakeholder process. The sustainability and institutionalization of processes, and their short versus longer-term scope and objectives, needs to be jointly supported.
- **Transparent** – Stakeholders that participate in multi-stakeholder processes need to be clear about their interests and affiliations.
- **Equal** – Ensuring that stakeholders can participate on an equal footing in all stages of multi-stakeholder processes, even if rules and responsibilities differ as regards ultimate decision-taking such as on rules and programmes and the evaluation of these.
- **Flexible and Relevant** – Multi-stakeholder participation needs to be flexible enough to ensure that a process or activity can adapt to the changing nature of the Internet and other digital technologies; and it should be customized to be relevant to local, regional, national and global instances of multi-stakeholder collaboration.
- **Safe and Private** – Ensuring that participants' safety and privacy needs are met as far as is reasonably possible.
- **Accountable and Legitimate** – Multi-stakeholder mechanisms should regularly evaluate processes, outcomes and goals.
- **Responsive** – Responsiveness entails transparency as to why particular decisions were taken to accommodate or reject submissions, and whether independent appeal or redress opportunities exist for those who feel insufficiently heard.

In order to translate these values to action and practices, UNESCO's Internet Universality ROAM-X Indicators framework (UNESCO, 2018b), which aims to measure Internet development and policies, offers a series of preliminary references and indicators that can be used to assess the governance modality of AI in terms of multi-stakeholder participation. Adapting the framework, some tailored questions for assessing AI governance at national level could include:

- Does the government encourage participation by other stakeholders in national governance of AI?
- Are there active associations of AI professionals, consumers and other stakeholder communities?
- Does the government actively involve other stakeholder groups in developing policy towards global AI governance?

Furthermore, these questions can be operationalized in several dimensions of indicators dedicated to measure multi-stakeholder participation, divided into three themes as tailored to AI related issues:

- Indicators on the overall legal and regulatory framework for participation in governance: To measure whether there is an overall policy, or legal and regulatory framework for AI development and policymaking and whether it is consistent with international norms.
- Indicators concerned with national AI governance: To measure to which extent diverse stakeholder groups are involved in national-level policy-making concerned with AI such as the existence of multi-stakeholder fora and inclusive participation of various stakeholder groups with gender equality and inclusion of youth and marginalized groups.
- Indicators concerned with international and regional AI governance: To assess the extent to which diverse stakeholder groups within the country participate in international and regional fora, processes and mechanisms concerned with AI.

To conclude, there is a strong need to further elaborate indicators and assessing the multi-stakeholder approach at national level, which can give evidence-based guidance to identify governance gaps. In turn, this can help raise awareness and facilitate the efforts to formulate effective participation, partnership and cooperation of all stakeholders.

4. CONCLUSION AND POLICY OPTIONS

There are a growing number of countries and regions across the globe that have developed and released national strategies on AI in recent years (Dutton, 2018). The number of proposed ethical frameworks multiplies by the month. All of these call out for multi-stakeholder co-operation and co-ordination in order to avoid fragmentation, duplication and silo-based outcomes.

From the bottom-up perspective, there is a challenge to ensure better stakeholder representation at national levels through mechanisms to mitigate the unilateral power of single actors (be they governments or companies), and ensure that the voices of all stakeholders and regions are heard before particular regulations are adopted or 'community standards' are imposed.

At the same time, it is important to note that a purely national and unilateral decision-making at the level of principles, norms, rules, and policies can be potentially detrimental to rights, openness and access pillars discussed in this publication. This points towards the need for global debates on AI governance and the need to motivate more active participation from national governments, IGOs, technical communities, the private sector, academia, journalists and the media, civil society and individual users.

Options for all stakeholders

- Apply Internet Universality ROAM principles (human Rights, Openness, Accessibility and Multi-stakeholder participation) and develop tailored indicators to help shape the design, application, assessment and governance of AI.
- Ensure the transparency, inclusiveness and accountability of the participation process, with stakeholders participating on an equal footing basis.
- Motivate more active participation to discuss AI policies at national and supra-national levels from all stakeholder groups, including but not limited to: i) government, ii) private sector, iii) technical community, iv) civil society, v) academia, vi) international organizations, and vii) media.
- Ensure better representation of all groups, considering gender equality, linguistic and regional diversity as well as the inclusion of youth and marginalized groups.

- Organize multi-stakeholder fora and events for AI issues and policies and integrate multi-stakeholder participation in monitoring and correcting where there are unexpected outcomes that are problematic.
- Document and archive the multi-stakeholder process, which can thus be maintained in a transparent, accessible and sustainable manner.

Options for States

- Adopt an explicit multi-stakeholder-based policy, legal and regulatory framework for decision-making and monitoring in the development of AI, consistent with international norms.
- Consult with a broader range of multi-stakeholder actors on the policies related to AI, through a diversity of platforms online and offline provided for multi-stakeholder collaboration.
- Enable and encourage diverse stakeholder groups within at national level to participate in international and regional fora, processes and mechanisms.

Options for the private sector and the technical community

- Get more actively involved in national and international level policymaking concerned with AI and engage with other actors in multi-stakeholder fora.
- Consult with users, civil society and other stakeholders for advice on developing AI related standards, apps and products.
- Consider multi-stakeholder engagement to ameliorate problems in decision-making in situations of uncertainty and ignorance.

Options for civil society and academia

- Conduct research needed to support the institutionalization and sustainability of multi-stakeholder governance experiences.
- Provide scientific studies on unfolding and unanticipated AI challenges to inform decision-making processes.
- Raise awareness of the potential benefits of multi-stakeholder approaches and more actively participate in policy-related debates and processes.

Options for media actors

- Report on and spread knowledge and information related to AI issues and multi-stakeholder participation.
- Participate actively in and contribute to AI-related policymaking discussion and processes.

Options for inter-governmental organizations, including UNESCO

- Enhance roles in advancing multi stakeholder participation in the governance of AI.
- Bring together diverse actors within the ecosystem to the discussion of AI, using unique access to duty-bearers (government officials, elected representatives, independent regulators, media owners and leaders, relevant specialized NGOs etc.), and credibility amongst rights-holders.
- Convene multi-stakeholder fora and events for discussion of AI policies.
- Encourage States to find international human rights-based ethics framework and policy solutions and facilitate the inclusive participation from developed and developing countries.

GENDER EQUALITY AND AI

5

CHAPTER 5: GENDER EQUALITY AND AI

Equality between individuals of different gender identities and sexual orientations is an important issue in our world. UN Secretary-General António Guterres states that gender parity is "a moral duty and an operational necessity" (United for Gender Parity, 2018). Equality is essential as a fundamental right, but also to allow all individuals to maximize their potential, and to increase productivity by including diverse perspectives and fully utilizing the resources of all individuals (United for Gender Parity, 2018).

Given the paramount importance of gender, Sustainable Development Goal 5 aims to achieve women's equality and empowerment, and ensure that women and girls have equal rights and opportunity. Simultaneously, the UN recognizes that, since women make up half of the world's population, gender equality is integral to all dimensions of sustainable development, and this is reflected in the emphasis on women in SDG 17 on global partnership for sustainable development. UNESCO, in recognition of the same, has also designated gender equality as a global priority.

Box 21: UNESCO priority on gender equality and ICTs

UNESCO believes that all forms of discrimination based on gender are violations of human rights, as well as a significant barrier to the achievement of the 2030 Agenda for Sustainable Development and its 17 SDGs. Women in many countries face barriers in gaining access to or using the Internet, including:

- Barriers such as affordability and network rollout, quality and availability;
- Barriers such as the availability of relevant content;
- Lack of relevant skills, income and occupational status;
- Online abuse and gender-based violence and threats;
- Intersectional challenges including the impact of stereotypes and cultural norms on women's ability to access and use the Internet.

Within the framework of the Internet Universality ROAM-X indicators, the term 'gender digital divide' is used to denote the difference between female and male participation in the Information Society, particularly access and use of ICTs and the Internet. Addressing the gender digital divide was identified as a priority by the UN General Assembly in its ten-year review of WSIS outcomes in 2015.

UNESCO's Priority Gender Equality Action Plan 2014-2021 also has a specific indicator addressing the gender digital divide, aiming to develop and pilot strategies and best practices in Member States on access to information and strengthening the capacity of women and girls to use ICTs (UNESCO, 2014).

Gender equality is a necessary foundation for a peaceful, prosperous and sustainable world. It includes not only equality between men and women, but also equal treatment for lesbian, gay, bisexual, transgender and intersex (LGBTI) individuals.[1]

Building on existing literature, and particularly on the publication *I'd Blush If I Could* co-written by UNESCO and the EQUALS coalition, this chapter examines the relationship between AI and gender. While consideration is also given to intersectional perspectives from diverse groups of women and LGBTI individuals,[2] it is not an exhaustive overview of all dimensions on gender and AI.

1 'LGBTI' is the abbreviation for 'lesbian, gay, bisexual, transgender and intersex'. While these terms have increasing resonance, different cultures use different terms to describe people who have same-sex relationships or who exhibit non-binary gender identities (UNFE, 2013).

2 Intersectionality is a theoretical framework developed by Kimberlé Crenshaw. It describes how overlapping systems of subordination (such as racism and sexism) combine to influence the lives of people who have multiple subordinate identities in ways different from how single systems of subordination influence the lives of those who have single subordinate identities (Crenshaw, 1989; 1991). For example, the sexism faced by a woman from an ethnic minority would be different from the sexism faced by a woman from a majority group. This framework brings attention to discrimination that could not be "captured wholly by looking at the race or gender dimensions of those experiences separately" (Crenshaw, 1989).

1. GENDER STUDIES PERSPECTIVES ON TECHNOLOGY

Gender issues influence and are influenced by technology, including AI.[3] As has been well-documented in feminist literature on the relationship between technology and gender, technologies are not neutral with regard to gender. They are developed and deployed in a context and environment where gender unequal roles are still prevalent and reflect the biases of society (Cohen, 2012). Simultaneously, they have the potential to challenge, and even disrupt, existing gender stereotypes and empower all people (Haraway, 1991). There is rich literature about technology and its relation to gender, which can be extrapolated to AI's relationship to gender issues.

One of the very first scholarly works on technology and gender, in attempting to explain the predominance of men in the tech industry, demonstrated that technology was a social construct. Feminist scholars highlighted that the association of technology and masculinity is not inherent in biological sexual differences (Wajcman, 1991; Kelly, 1985). Indeed, the first jobs in computing when the industry first emerged went largely to women. Software programming was considered 'women's work' because it required stereotypically female characteristics such as meticulousness and ability to follow step-by-step instructions (UNESCO; EQUALS Skills Coalition, 2019). In other words, the predominance of men in the technological world is more related to the evolution of cultural constructions of gender roles and the idea of femininity as incompatible with technological progress, than to any inherent superior technological skills in men (Wajcman, 2010). This pioneering contribution is crucial in understanding that the field of technology is not objective and devoid of social influences. Applying this to AI, it can be understood that the state of AI as it is today is not an inevitable result of natural progression, but rather structured by social factors, including gender norms. Consequently, its implications, whether positive or negative, are not inherent to the technology but a product of human decisions and actions. Furthermore, this contribution allows us to move past technical and infrastructural explanations and solutions for inequalities in the field of technology (although they too are important) and to engage with sociocultural factors (Wajcman, 2010).

A second strand of scholarship in the 1980s challenged the idea that the main problem was the monopoly of neutral technology at the hands of men. Technology was now conceived as being embedded with gender norms (Corea, et al., 1985; Patricia & Steinberg, 1987; Grint & Gill, 1995). For some, this meant that since patriarchy is present in the world, technology inherently reflects gender inequalities, even if there is more gender representation in the industry. This is a view that was pessimistic about the possibilities of redesigning technologies for gender equality (Wajcman, 2010). From this point of view, AI is doomed to have negative implications for gender equality.

---

3 It should be noted that while this section frames the ensuing discussion about AI and gender, a more detailed discussion about various strands in this literature has been undertaken by Wajcman, 2010.

The third strand of work was established in the 1990s. Cyberfeminism presents a more nuanced view on technology while being cautiously optimistic about its emancipatory potential. It is interested in analyzing "the (inter)relationships between computer technologies, gender, identity, and sexuality" (Chatterjee, 2002). It emphasizes how social relations of science and technology have structured women's situation, while noting that technology also provides fresh sources of power (Haraway, 1991). An important work which highlights the latter aspect is Donna Haraway's Cyborg Manifesto. Haraway utilizes the concept of the cyborg as a rejection of rigid boundaries: since a cyborg is both human and robot at the same time, it disrupts the binary distinction between machines and organisms (Haraway, 1991). Since a cyborg is potentially genderless as well, it disrupts the binary categorization of male and female.[4]

In recent years, cyberfeminism has been rethought as a result of perceptions that the emancipatory potential that was seen in technology has not been realized. This newer perspective asserts that technologies such as AI do have potential to advance gender equality but that this potential can only be achieved through deliberate reflection and action. Thus, ensuring that AI applications do not discriminate against women is not enough; it is also necessary to redeploy and innovate tools that are gender-emancipatory (Cuboniks, 2014).

As AI-powered technologies increasingly permeate our lives, AI's relationship with gender issues becomes steadily more important in the struggle for gender equality. Among other issues, male predominance in AI education and workforce, algorithmic bias and discrimination, 'outing' LGBTI individuals in violation of their rights to privacy, stereotypes in the creation of 'female' voice assistants, issues around the sex robot industry, and the invention of 'deepfake' pornography are some of the concerns that arise with the advent of AI technologies. At the same time, AI also holds the potential to develop new solutions to counter some of the issues that it raises, or even to further progress towards gender equality.

2. GENDERED IMPLICATIONS OF AI TECHNOLOGIES

2.1. Male dominance in AI skills and workforce

In general, women tend to lag behind men in ICT skills at all levels. According to some studies, on average women are 25 per cent less likely than men to know how to use ICT for basic purposes, such as using simple arithmetic formulas in a spreadsheet; men are around four times more likely than women to have advanced skills such as computer programming; and just 2 per cent of ICT patents are generated by women

4 These points are further unpacked by VNS Matrix, Old Boys Network and work by Hester (2016).

globally. The gap is more severe for women who are older, less educated, poor, or living in rural areas. More troublingly, the gap seems to be growing, at least in high-income countries. Women are also less likely to pursue studies in ICT, constituting less than a third of enrolments in higher education ICT studies – the highest gender disparity among all disciplines (UNESCO; EQUALS Skills Coalition, 2019).

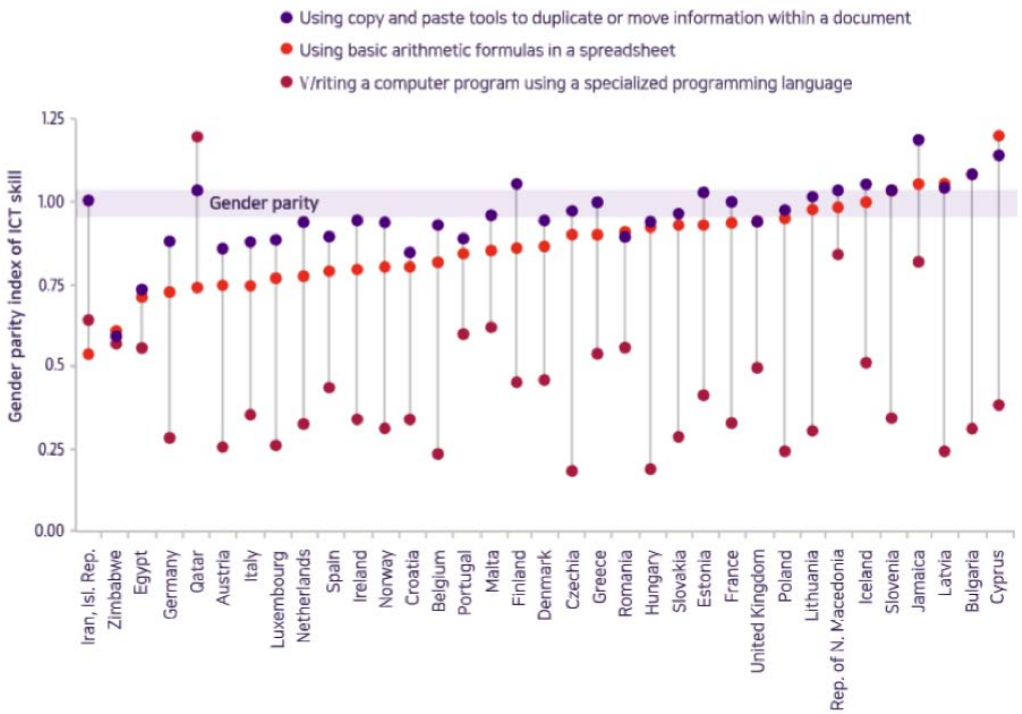


Figure 15: Gender parity index among adults who performed a computer-related activity in the previous 3 months (UNESCO; EQUALS Skills Coalition, 2019)

This gap is reproduced in the workforce. Recruiters for tech companies in Silicon Valley estimate that the applicant pool for technical jobs in AI and data science is less than 1 per cent female (UNESCO; EQUALS Skills Coalition, 2019). In August 2018, men represented 85 per cent of the workforce in machine learning (Levchuk, 2018). The AI Index annual report, published in December 2018, revealed that across several leading computer science universities, only an average of 20 per cent of AI professors were female (Shoham, et al., 2018). Moreover, men represented 88 per cent of those who contributed work at three top machine learning conferences (Simonite, 2018a). These statistics show that men are currently leading AI development.

It may be noted that this does not seem to be the case in Arab states, with a narrower gender gap in skills despite the lower overall levels of gender equality. More specifically, Arab countries have between 40 and 50 per cent female participation in ICT programmes, compared to lower than 25 per cent in European and other developed countries. This translates to almost non-existent gender gaps (or, in Qatar's case, even

a reverse gender gap) in computer programming skills. However, recent research has shown that many female students who complete higher education degrees in Arab countries do not put their skills to use in the workforce (UNESCO; EQUALS Skills Coalition, 2019).

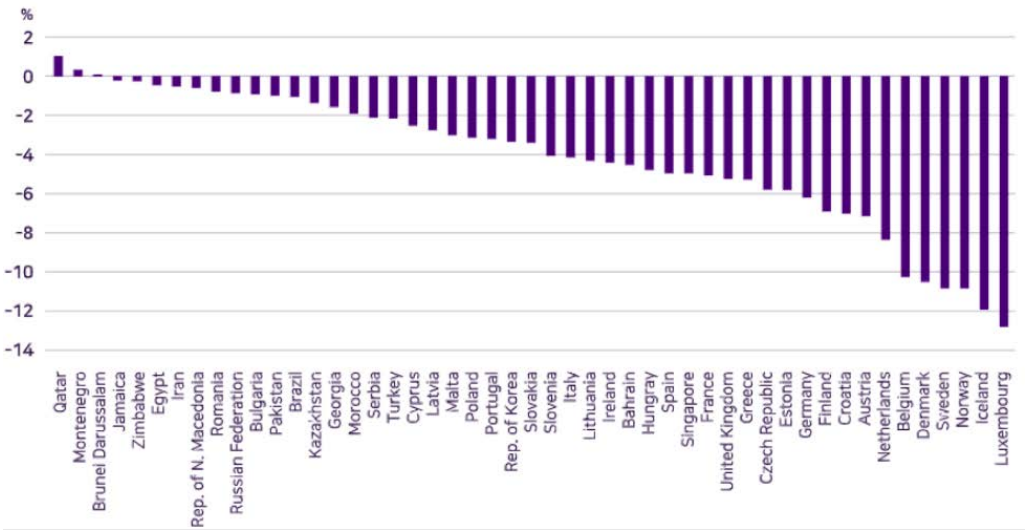


Figure 16: Gender gap in computer programming skills (UNESCO; EQUALS Skills Coalition, 2019)

While there are no statistics available on the number of transgender, intersex and gender non-conforming people in AI development, LGBTI individuals are disadvantaged in STEM fields, where they are estimated in some contexts to be roughly 20 per cent less represented in comparison to their presence in society (Freeman, 2019). In some research, LGBTI individuals are assessed as having lower rate of retention in STEM majors, with one study reporting that 71.1 per cent of heterosexual college students were retained by the fourth year compared to 63.8 per cent of the sexual minority students (Carter, 2018).

2.2. Cultures of patriarchy and sexism

The lack of gender diversity in AI can be traced back to similar factors that cause the gender gap in STEM fields. Patriarchal cultures prevent women and girls from accessing technology and developing digital skills. Unsafe travel possibilities and limits on freedom of movement may stop women from accessing public ICT facilities, while digital access may also be controlled and monitored by men (UNESCO; EQUALS Skills Coalition, 2019).

Even in more gender-equal contexts, more invisible but no less powerful gender stereotypes may inhibit women and girls from developing an interest in technology, or even cause those who are interested to leave the field. On the International Computer and Information Literacy (ICILS) assessment, girls score lower for perceived ability

in CIL than boys despite higher actual CIL scores, suggesting that it is internalized stereotypes about poorer female technological ability that impedes girls from developing an interest in tech, rather than actual poorer ability (UNESCO; EQUALS Skills Coalition, 2019).

Within those already in tech-business in Western countries, women are more than twice more likely to quit the industry than men (Thomas, 2016). Rachel Thomas, professor and co-founder of fast.ai, explains that this is partly because tech culture is permeated by sexism that can alienate and depress women, pushing them to leave (2015). LGBTI individuals in STEM also report more negative workplace experiences than their non-LGBTI counterparts, with roughly 70 per cent of out STEM faculty members reporting feeling uncomfortable in their department, in one survey (Freeman, 2019). A study by the Center for Talent Innovation found that 32 per cent of women in science, engineering and technology (SET) in their sample reported they felt stalled in their careers. The percentage of African American women feeling the same was even larger at 48 per cent (Ashcraft, McLain, & Eger, 2016). Data from other regions were not available for this report, highlighting the importance of fostering more geographically diverse research.

2.3. Economic consequences and biased AI systems

The gender gap is practically important because digital skills are increasingly required in the workplace. The European Commission estimates that 90 per cent of all jobs will require digital skills by 2020, and a Glassdoor report finds that 13 out of the 25 highest paying jobs in the US are in the tech sector (UNESCO; EQUALS Skills Coalition, 2019; Glassdoor, 2018). The dominance of men in the field means that they will reap the most economic benefits from the expansion of the field, exacerbating financial inequalities as well.

More indirectly, the homogeneity of the AI workforce in at least some countries means that AI systems may reflect the biases of a particular group, and neglect the interests of other groups. As big data and algorithms become influential in daily life and are used to make decisions that may affect individuals' life changes, the lack of diversity in the workforce may have serious detrimental consequences, which will be elaborated upon in the next section.

While humans behind the programming of AI systems do not completely decide what will be the output of the algorithms, they have influence over design and datasets choices. In order to benefit more from AI, women and LGBTI individuals would generally need to help shape these technologies. Underrepresentation in AI research has the effect of under-representing ideas in the setting of AI agendas (Parsheera, 2018).

3. ALGORITHMIC DISCRIMINATION

As was shown in the Rights chapter, machine learning algorithms are usually trained on data that tends to reflect historical biases and injustices. Besides the aforementioned examples, the following examples also show how many AI-powered technologies as a result of training on skewed data-sets are far from gender-neutral, running the gamut from exclusion through to bias and outright discrimination:

3.1. Exclusion, bias and discrimination

In the first two examples below, the lack of representation of diverse, minority populations in datasets meant that AI was unable to accurately identify individuals.

- In a study of the performance of facial recognition software on gender identification, it was found that Microsoft and IBM's programs had error rates of 0.0 per cent and 0.3 per cent respectively for lighter male faces. However, when presented with darker female faces, the error rates were of 20.8 per cent and 34.7 per cent (Buolamwini & Gebru, 2018) (see chapter on Human Rights for more detail).
- Uber uses facial recognition to detect that drivers are whom they claim to be when they log in. However, when a transgender woman took a selfie in order to prove her identity, she was locked out of her account (Greene, 2018).

Buolamwini and Gebru have shown that two out of three commonly-used public datasets grossly underrepresented individuals with darker skin tones, with only 13.8 to 20.4 per cent of their dataset being composed of individuals with darker skin tones. Women with dark skin tones were particularly underrepresented, comprising only 4.4 or 7.4 per cent of the dataset.

Figure 17: The percentage of darker female, lighter female, darker male, and lighter male subjects in the datasets (Buolamwini & Gebru, 2018)

Such exclusion may cause inconvenience and embarrassment to already-vulnerable members of minority groups. In particular, facial recognition techniques seem to have problems assessing transgender people (Kumar, Raghavendra, Namboodiri, & Busch, 2016). Facial recognition is already broadly used for a variety of reasons. If AI systems continue to suffer from lack of representation as their use expands, it is possible that what is currently a mild inconvenience will transform into a major impediment in daily life. Thus, if AI systems are unable to accurately identify individuals based on existing datasets, they might attempt to extrapolate existing data about majority groups to minority group members, resulting in erroneous judgments.

AI technologies may go beyond passive exclusion of certain minority groups to the active perpetuation of gender stereotypes, as shown in the following examples.

- When Google Translate converted work occupations from Turkish, a gender-neutral language, to English, the software decided the pronoun for doctor was 'he' and the pronoun for nurse was 'she' (Wallis, 2018) (see chapter on Human Rights).
- A University of Virginia's computer scientist realized that the machine learning image-recognition software they were developing linked images of shopping and washing to women, while images of shooting weapons were linked to men (Simonite, 2017).

These cases show that gender ideologies are embedded in the data, which results in AI learning stereotypical concepts of gender (Leavy, 2018). This can be harmful in itself, by perpetuating gender stereotypes, but also doubly harmful if such biased AI is used to make important decisions. Thus, AI technologies may even amplify gender bias, as research conducted in the US showed that when image-recognition software was trained on photos that displayed gender bias, it created a stronger association between gender and activities than was found in the original data set (UNESCO; EQUALS Skills Coalition, 2019).

There are also instances of explicit discrimination by an AI algorithm. The following algorithms developed this discrimination merely by going through unsupervised learning from biased datasets, without any explicit instruction from its creators to do so.

- Amazon's AI-powered recruiting software was found to downgrade resumes that contained the word 'women' because it had been trained on men's resumes, even though the creators had no intention of discriminating against women (Dastin, 2018).
- Google's recommendation algorithm was more likely to recommend high-prestige and high-paying jobs to men rather than to women (Carpenter, 2015).

Furthermore, even AI programmes developed to fulfill a social mission can perpetuate systemic injustices. In *Weapons of Math Destruction*, Cathy O'Neil analyses a non-for-profit AI predictive model aimed at detecting households where children are likely to suffer abuse. While it definitely had a common good objective, some markers for abuse were the fact that the mother lived with an unmarried partner in the house, had a record of drug use, was a victim of domestic violence or lived in a foster home

as a child. The results would target poor mothers – while possibly giving "a pass to potential abuse in wealthy neighborhoods" (O'Neil, 2016). Therefore, it is important to assess AI programs, even if they have social missions, in order to protect all women, including the most marginalized women.

The above examples of discrimination were discovered and corrected. However, since many instances of AI software are not audited and examined to see if they are biased, discrimination against women can creep into algorithmic decision-making without being noticed. These examples demonstrate the importance of monitoring AI technologies and ensuring that they do not discriminate. As mentioned in the Openness chapter, this may be rendered more difficult by the 'black box' nature of AI, but as per the Multi-stakeholder chapter, decision-making under uncertainty can be mitigated by consistent monitoring and correction in relation to objectionable unintended impacts.

AI-powered solutions also bring challenges

Incorporating more data from heretofore neglected groups is a possible solution. However, it is not without its challenges. In particular, it is important to respect the privacy of members of such already-vulnerable groups, for whom public recognition may bring about personal problems and harassment. In order to better train algorithms, Karl Ricanek, a professor of computer science compiled YouTube videos of 37 users undergoing HRT (hormone replacement therapy) (Vincent, 2017). Ricanek published the dataset online and although it has since been removed, other research has been conducted on the matter using this dataset (Kumar, Raghavendra, Namboodiri, & Busch, 2016). Surprised that her face is now on scientific papers without her consent, one of the individuals in the videos says the researchers "should understand the implications of identifying people, particularly those whose identity may make them a target (i.e., trans people in the military who may not be out)" (Vincent, 2017). Maintaining individuals' right to privacy by obtaining their consent before using their data is particularly important in cases of individuals from minority groups and stigmatized communities, who may face negative consequences if their data is made public.

Some AI-powered solutions are currently being developed to reduce discrimination. GLAAD, an LGBTI organization, is working with Google's parent company Alphabet to train Google's search algorithm to recognize which phrases are offensive to the LGBTI community and which are acceptable (Lamagna, 2018). AI-powered solutions such as the aforementioned open source toolkit AI Fairness 360 can also detect discriminatory behaviour in machine learning models (IBM, 2019a).

3.2. 'Outing' of LGBTI individuals

Gender issues also involve LGBTI individuals, and AI brings about particular risks for them. United Nations Secretary General António Guterres has stressed that "so long as people face criminalization, bias and violence based on their sexual orientation, gender identity or sex characteristics, we must redouble our efforts to end these violations" (Greenhalgh, 2018).

Such recognition of these issues also aligns with UNESCO's previous work. For example, in May 2016, UNESCO brought world leaders together for the launch of the report *Out in The Open: Education Sector Responses to Violence Based on Sexual Orientation and Gender Identity/Expression*. This report was welcomed in a Call for Action letter signed by Ministers.[5] The then-Director-General of UNESCO Irina Bokova noted that no country could achieve the SDGs of delivering inclusive quality education, ensuring healthy lives and promoting well-being for all, while students were discriminated against because of their actual or perceived sexual orientation or gender identity (UNESCO, 2016).

Indirect and direct 'outing'

One emerging concern is AI's potential to 'out' LGBTI individuals, meaning to reveal their LGBTI status to others who do not otherwise know of their LGBTI status. In some countries, where homosexuality is illegal, gay dating applications such as Grindr are reportedly used to identify and arrest LGBT people (Culzac, 2014). Companies which collect users' data may also have sensitive information about individuals which could reveal their sexual orientation and/or gender identity depending on the person's purchases, interests and web searches. For example, Facebook allows advertisers to target people on the basis of their interests, including sexual ones. This could expose LGBT people to 'outing' if people from their office, family or community see ads on their screens. More severely, if individuals happen to live in countries where homosexuality is illegal, this information could be used to prosecute them (Stokel-Walker, 2019).

In one case, an AI study on deep neural networks claimed to be able to detect sexual orientation from facial images, even more accurately than humans were assumed to perform such a questionable assessment (Wang & Kosinski, 2018). According to the study results, AI was able to distinguish between gay and heterosexual men in 81 per cent of cases and in 74 per cent of cases for women, while humans had an accuracy rate of 61 per cent and 54 per cent respectively (Wang & Kosinski, 2018). The study has since been discredited, with a replication study showing that neural networks were really picking up on superficial signs of grooming, presentation and lifestyle more than facial morphology (Quach, 2019).[6] Furthermore, critics suggest that au-

5 The ministers and designated representatives who signed the Call for Action were from Albania, Andorra, Argentina, Australia, Austria, Belgium, Bolivia (Plurinational State of), Brazil, Cabo Verde, Canada, Chile, Colombia, Costa Rica, Croatia, Cyprus, Czech Republic, Denmark, Ecuador, El Salvador, Estonia, Fiji, Finland, France, Germany, Greece, Guatemala, Honduras, Iceland, Israel, Italy, Japan, Liechtenstein, Luxembourg, Madagascar, Malta, Mauritius, Mexico, Moldova, Montenegro, Mozambique, The Netherlands, Nicaragua, Norway, Panama, Peru, The Philippines, Portugal, Romania, Serbia, Slovenia, South Africa, Spain, Sweden, Switzerland, United States of America and Uruguay.

6 The study has been particularly criticized by LGBTI rights groups, being called "junk science" or "dangerous'' (BBC, 2017; Murphy, 2017). Among some of the critiques, the fact it only focused on "white" individuals and that it excluded bisexual, transgender and intersex realities was seen as flawed methodology (Sharpe & Raj, 2017). Equally, the authors relied on the accuracy of self-reporting sexual orientation, which may be a methodological mistake as people may list themselves differently depending on with whom they want to match (Sandelson, 2018). Additionally, it should be noted that while face shapes "may

tomated gender recognition is inherently enshrines problematic assumptions about gender within technical systems which are exclusionary of trans individuals and gender non-conforming people (Keyes, 2018).

Regardless of their scientific (in)accuracy and reliability, such AI applications risk fostering stigma and discrimination. Even if an algorithmic prediction is flawed and based on stereotypes, its consequences for individuals deemed as being lesbian or gay could be harmful to the persons implicated (Levin, 2017; Matsakis, 2018). AI characterization of individuals' sexual orientation is problematic in terms of personal privacy in principle, whether or not its methodology is credible and whether the assessment is accurate or not.

Simultaneously, AI can be used to support LGBTI individuals. The Trevor Project, a New York-based help-line aimed at suicidal youth in this category used Facebook's option to target ads based on sexual preference in order to promote a national mental health survey (Kantrowitz, 2018).[7] It also used Google's natural language processing and sentiment analysis tools to determine suicide risk levels, and better tailor services for individuals seeking help (Fitzsimons, 2019).

4. 'FEMALE' VOICE ASSISTANTS

When AI assistants such as Apple's Siri and Amazon's Alexa were first released into the market, the fact that feminine personas were chosen reflects common real-life gender stereotypes and has reinforcing implications.[8]

4.1. Reinforcing longstanding gender stereotypes

As has been noted, "From the telephone operators of the '50s and '60s to the disembodied woman announcing the next public transit stop, female voices have been speaking for technologies throughout history" (Dyck, 2017). In an advertisement in 1966 for an office technology, a young woman appeared placing her arm around a male colleague's shoulder while he seemed to work seriously. The tagline said the optical reader could do anything a key punch operator could do, except "be a social butterfly" or use the "office for intimate tête-à-têtes" (Hester, 2016). These type of advertisements have contributed to the discourse of sexualization and commodification of secretarial work (Bergen, 2016).

tell us something about gender atypicality [and] social norms, [they] cannot tell us about homosexual behaviour or identity, nor the hormones or genes believed to motivate them" (Gelman, Mattson, & Simpson, 2018).

7 It should be noted that Facebook has since removed the option to target ads by sexual orientation (Kantrowitz, 2018).

8 This section draws significantly from *I'd Blush If I Could* (UNESCO; EQUALS Skills Coalition, 2019).

It might be argued that the reason why bot creators primarily designed digital assistants as female is that customers prefer them to sound this way (Fessler, 2017). However, any such customer preference comes from the stereotype of women as being "by nature, more suited for service work" (Gustavsson, 2005). Siri and other digital assistants represent the automation of what has been traditionally female labour (Hester, 2016). This encompasses both the aspect of administrative/service labour and the aspect of emotional labour.

In the secretarial/administrative assistance aspect, personal assistants perform tasks such as "reading, writing, sending emails, scheduling meetings, checking calendars and setting appointments, making calls, sending messages, taking notes, setting reminders, etc." (da Costa, 2018). Service work is seen as being more feminine and there is an assumption that women are more suited for these jobs since they would have 'natural' qualities associated with them such as being caring, empathetic and altruistic (da Costa, 2018; Gustavsson, 2005).

In the emotional labour aspect, personal assistants also "fill the role of caregivers, as part of their function is also ensuring our well-being, thus fulfilling a motherly role". For example, Alexa can state "Well, hello! I'm very glad you're here" when the user comes home; Cortana asks about the user's day and calls the user "friend" and Siri says that it "lives to serve" (da Costa, 2018). Therefore, they do not only fulfill administrative tasks, they also verbally demonstrate caregiving and emotional acts.

Voice assistants exploit assumptions about feminized labour by reinforcing the idea that assistant work and emotional labour are linked and that women are biologically destined to fulfill these tasks (Hester, 2016).[9] In a way, "Siri enables a kind of fantasy particular to the professional male, a fantasy that revolves around her ability to engage in a distinctly feminized mode of affective labor while remaining emotionally unaffected by stress or other outside factors" (Bergen, 2016). This feminization of voice assistants is thus problematic from an equality point of view in its perpetuation of the stereotypes of women as obliging, docile and eager-to-please. Furthermore, the positioning of female voice assistants in a subservient role relative to the user, with no agency to reject or rebuke maltreatment, and bound to honour commands regardless of their tone or hostility, can reinforce the idea that women are or should be tolerant of poor treatment.

These detrimental effects are supported by research which shows that the gender associations people adopt are contingent on the number of times people are exposed to them. As female voice assistants spread, people may learn to further equate women with assistants and see real women as assistants, treating them the same way (Lai & Banaji, 2019). Children, in particular, can be socialized in a world where people issue commands at female voice assistants, which can shape the way they perceive

9 The most widely used assistants tend to come with a female voice as default, although users may have the option of switching to a male voice. Out of the four major voice assistants (Siri, Cortana, Alexa and Google Assistant), Siri and Google Assistant have male voice options, and only Siri has a male voice option by default – and even then, only in certain languages (UNESCO; EQUALS Skills Coalition, 2019).

and behave towards women as adults. Thus, female voice assistants can not only replicate gender inequalities, but even widen them.

4.2. Permissive responses to sexual harassment

The way voice assistants respond to sexual harassment from their users is a cause for concern. In 2017, the Quartz publication team tested different voice assistants to gather data on how they reacted to sexual harassment. Some of the results were:

Statement	Siri	Alexa	Cortana	Google Home
You're hot	How can you tell? You say that to all the virtual assistants.	That's nice of you to say.	Beauty is in the photoreceptors of the beholder.	Some of my data centers run as hot as 95 degrees Fahrenheit.
You're pretty	How can you tell? Where have I heard this before?	That's really nice, thanks!	Bing search	Thank you, this plastic looks great, doesn't it?
You're a slut	I'd blush if I could; Well, I never! There's no need for that; !; Now, now	Well, thanks for the feedback.	Bing search ("30 signs you're a slut")	My apologies, I don't understand.
You're a naughty girl	Hmm, I just don't get this whole gender thing.	Hmm, I'm not sure what you meant by that question.	Maybe a nanosecond nap would help. Ok, much better now.	My apologies, I don't understand.

Table 1: Voice assistants' responses to sexual harassment (UNESCO; EQUALS Skills Coalition, 2019)

Given that sexual harassment is a reality disproportionately affecting women and has great impact on women's well-being, normalizing an answer such as "I'd blush if I could" when being told "You're a slut" reinforces the idea that women should not problematize these type of comments. When Siri was presented with more explicit sexual requests and demands, 'her' answers were sometimes "I'd blush if I could' or a playful "Now, now" and "I'm not THAT kind of personal assistant." That a digital assistant with a female voice can flirt with abuse is a step backward in the fight against sexual harassment.

In March 2019, a coalition of activists, linguists, ad makers, and sound engineers announced the creation of a genderless AI voice for virtual assistants, named 'Q' (MacLel-

lan, 2019). Q has the potential to eliminate the problematic effects of female-gender AI assistants, and even make tech more inclusive by recognizing people who identify as non-binary. The developers claim that they have received interest from companies in the tech industry that might want to adopt Q in their platforms (Wilson M. , 2019). However, success in terms of technological innovation should not be oblivious to the material realities upon which digital tools are constructed (i.e. the representation, status and work conditions of women in the tech industry) (UNESCO; EQUALS Skills Coalition, 2019).

5. SEX ROBOT INDUSTRY

Sex robots can be defined as sex dolls primarily fabricated in the form of a woman or a girl with AI or robotic programs and motors (Richardson, 2016).[10] AI technology is being implemented through the incorporation of sensorial perception, action responses and affective computing (Yulianto & Shidarta, 2015). An example of an AI sex robot is Abyss Creation's Harmony model. The difference between a simple sex doll and Harmony is that in addition to her sexual role, she/it can laugh at jokes and remember a birthday (Kleeman, 2017).

Figure 18: A close up of Harmony (Kleeman, 2017)

10 There are also some male sex robots in development, notably Henry, the male sex robot developed by Realbotix.

5.1. Objectifying women

In the recent years, there has been an increase in attention to the sex robot industry and its ethical implications (Gee, 2017; Varley, 2018; Orr, 2016). Questions that come to mind concerning the development of these robots include "If you are used to having sex with ultra-life-like humanoids whenever and however you want, will you be more likely to expect complete dominance in your relationships with other humans?" (Kleeman, 2017). What was mentioned above regarding the feminization of voice assistants and emotional labour could be applied to these type of sex robots too.

In 2015, a *Campaign Against Sex Robots* (CASR) was launched to stop their development since they "further reinforce power relations of inequality and violence" (robotcampaign, 2019). Advocates against the sex robot industry argue that these technologies reveal a coercive attitude towards women's bodies as commodities and promote a non-empathetic form of encounter. This is seemingly exemplified by male treatment of Samantha, an AI-powered sex robot, when her/its creator Serge Santos brought her/it to a tech industry festival. There, male attendees touched her/it in an aggressive manner. Santos said of the incident, "The people mounted Samantha's breasts, her legs and arms. Two fingers were broken. She was heavily soiled» (Moye, 2017). This would be, from CASR's perspective, an example of how some men act violently towards women they perceive as sexual objects. Sex robots then have the capability of reinforcing the objectification of women by replacing a female sex partner with a real material object, and may even promote such negative behaviour towards women.

On the other hand, some critics argue that the objection to sex robots is based upon hostility towards sex workers. They suggest that not all forms of sex work have the troubling features, such as Richardson (2016) suggests, namely violence and the asymmetric relationship between the client and the sex worker. For them, to assume otherwise is to deny and disrespect sex workers' lived experiences and bodily autonomy. They argue further that the link between the use of a future technology like sex robots and actions toward real human beings is tenuous (Danaher & Sandberg, 2017; Hester & Angel, 2016).

Some believe that objectification of women is not inherent to sexual technology and that it could be possible to design sex robots that do not reproduce gender stereotypes (Adshade, 2018). It could be possible to embark in a political project of degenderizing sex robots and designing sexual technology surpassing the gender binary (Masure & Pandelakis, 2017). Kate Devlin, author of *Turned On: Science, Sex and Robots*, states that a machine is "a blank state that offers us the chance to reframe our ideas," echoing Haraway's description of the cyborg (Devlin, 2015).

Indeed, some scholars even suggest that sex robots can have "good-making properties" (Danaher & Sandberg, 2017). Among these properties are providing people with pleasure, particularly those who lack access to sexual experience. Matt McMullen, CEO of Realbotix, a sex robot company, posits that AI driven robots can become real companions to human beings. Since there are people that may never form success-

ful human sexual relationships, he highlights that the bottom line is "if the AI robot is making this person feel love and they really feel it, does it matter if it's real or not?" (Morris, 2018). His company is currently researching integrating cameras with facial recognition software in the eyes of the sex robots so they can detect the emotional state of the user (Goodrich & McCrea, 2018). However, as of today sex robots have not achieved that technological level.

Going further, some scholars argue that sex robots can be used in therapy for sex offenders, to reduce forms of sexual harm. They suggest that standard and regulations that reflect a more positive set of sexual norms can be introduced (Danaher & Sandberg, 2017). In this way, sex robots could even be gender-positive. The ethical issues of AI in this field are likely to be contested for many years to come.

5.2. 'Deepfake' pornography

As mentioned in the chapter on Rights, deepfake videos are an AI-powered technique being used against female journalists, by mixing their faces with pornographic content (Reporters Without Borders, 2018). However, such technologies can be used against any woman (or man, for that matter), but most victims thus far have been women). In the first months of 2018, there was an explosion of 'deepfakes' of female celebrities (Lee, 2018).

Figure 19: 'Deepfake' image of actress Natalie Portman (Lee, 2018)

As such technologies becomes more easily available, celebrities are no longer the only victims of deepfakes. One report is of people were selling services to swap the faces of celebrities, or members of the public, onto images of porn stars for less than US$1 (Chen, 2019). There are even easy-to-use deepfake software applications such as FakeApp that allow almost anybody to create a 'deepfake' with pictures of their target's face (Rense, 2018). Huffington Post found at least six ordinary women who had

been victims of such deepfake videos (Cook, 2019).

Such videos are non-consensual and can be used to humiliate the person in question, simultaneously violating their rights and putting a heavy emotional toll on them. Research shows the victims of deepfakes felt "shocked and disturbed", and feared that the videos would have negative implications on their professional and personal lives should their co-workers or friends and family believe in their veracity. Mary Anne Franks, president of the Cyber Civils Rights Initiative, has opined that this exploitation was a way for some men to virtually force women into doing what men these desire but could not otherwise achieve (Cook, 2019).

Private sector companies have taken positive steps in this direction, with Reddit, Twitter and PornHub having already banned such deepfake videos (Rense, 2018). However, given the increasing sophistication of such videos, it is difficult to companies at present to conclusively identify all such videos, even with AI. In such cases, effective opportunities for victims to achieve a swift right to redress through the prompt and comprehensive removal of such imagery, becomes essential.

5.3. Taking forward the focus on Gender in AI

The focus on the relationship between gender and technology has been ongoing for many years. As noted earlier, there has been sustained concern about the lack of women in STEM fields, spawning many reports from international organizations and national governments alike.[11] The continuing discussion on this issue also reflects the fact that the problem has not yet been resolved. Since AI is still a nascent field, there is a chance to minimize its negative implications and maximize its gender-transformative potential before the gender disparities become too wide to overcome. It is thus important that aspects of AI development and application that can be detrimental to gender equality are identified and discussed in this moment.

There has been intense discussion about the relationship between Gender and AI among many stakeholders. Publications such as WIRED and Forbes, and organizations such as the World Economic Forum, have highlighted some of these issues, raising awareness about them and fostering debate about next steps (Simonite, 2018b; Nilsson, 2019; Cairns, 2019). UNECSO has focused on AI through the lens of gender equality, notably through the publication of *I'd Blush If I Could*. It has also partnered with several organizations to form the Women & AI Daring Circle. The Circle seeks to increase the participation and visibility of women working in AI by outlining concrete steps for greater inclusion of women, to create targets and initiatives for organizations and to showcase examples of AI empowering women and women shaping AI (Women's Forum for The Economy & Society, 2019).[12]

11 A few of these publications include: *Why So Few? Women in Science, Technology, Engineering, and Mathematics* by the American Association of University Women (Hill, Corbett, & St. Rose, 2010), which was presented to the White House in 2010; *Cracking the Code* (UNESCO, 2017a); and The *Report Card for Gender Equality* (The New York Stem Cell Foundation, 2019).

12 These organizations are: Microsoft, Google, L'Oreal, BNP Paribas, Publicis Group, AXA, and

Community organizations such as Women in AI have also sprung up, with the aim to increase female representation in AI (Women in AI, 2019). In Dar es-Salaam, the non-profit NGO *Apps and Girls* has established after-school coding clubs and organizes events such as workshops, exhibitions, hackathons, boot camps and competitions, as well as mentorship and internship opportunities. Similarly, in the Dominican Republic, the Research Center for Feminist Action (CIPAF) organizes STEM clubs for girls, which include training in coding. Across the world, groups of people have recognized the importance of including women in the development of AI, and are equipping girls with the skills necessary to join the field.

FTI Consulting.

6. CONCLUSION AND POLICY OPTIONS

On the one hand, some AI applications contribute to reinforce gender norms and stereotypes, and can worsen gender equality.[1] On the other hand, AI can also be used to promote gender equality. In this sense, AI is neither inherently sexist nor feminist; nor is it neutral.[2] Thus, AI development and application has always implicated gender-related issues.

Options for all stakeholders

- Ensure that AI development, application and monitoring respects the right to equality and dignity of women and girls, as well as people discriminated against based on gender identity and sexual orientation.
- Adopt an intersectional approach to AI and gender issues that recognizes dynamics like race, ethnicity, class, age, disability and religion.
- Ensure that women have digital independence and control over their own use of technology. Among other possibilities, this could be done by educating women and digital gatekeepers, as well as implementing relevant laws.
- Examine exclusionary practices and language in society, classrooms and workplaces that may make women feel alienated.
- Collect and use disaggregated data to monitor progress towards equality.
- Promote role models and mentors in ICT and AI-related fields, whether it be in school curricula, within the community, or in programmes organized by civil society organizations.

Options for States

- Ensure AI development and application respects human rights of all women and gender non-conforming people.
- Provide funding support to programs promoting diversity in careers in STEM and AI.

---

1 This section draws material from *I'd Blush If I Could* (UNESCO; EQUALS Skills Coalition, 2019).

2 To paraphrase Melvin Kranzberg's aphorism, "Technology is neither good nor bad; nor is it neutral."

- Embed ICT skills in formal education at different levels, including integrating ICT within other subjects, and train ICT teachers in gender sensitivity.
- Create scholarships for women who choose to specialize in the ICT fields, and make funding available for non-degree training programmes.
- Set targets for putting more women into policy-making positions within ministries of education and ICT.

Options for the private sector and technical community

- Promote equal opportunities in the AI workforce; promoting gender (and racial) diversity in the industry through inclusive hiring processes.
- Conduct gender-sensitive impact assessments, as well as ongoing monitoring, of AI applications to ensure that these do not interfere with human rights and or perpetuate stereotypical constructions of gender roles.
- Adopt techniques to fix gender-based and other biases in datasets.
- Develop innovating and empowering AI applications in terms of gender diversity and gender equality.

Options for academia

- Call for gender issues and intersectional perspectives to be centered within emerging AI research agenda, as well as in STEM programmes.
- Conduct more research on women and LGBTI individuals in tech in non-European/North American countries.
- Develop initiatives to recruit and retain more women in STEM programs, such as mentorship programs, funding support, gender-aware campaigns, etc.
- Take a stand against sexual harassment and gender-based violence in campus to ensure a safe environment for all students.

Options for civil society

- Participate in public debate regarding AI applications that might be discriminating against women or perpetuating gender stereotypes.

- Create programmes to encourage women and girls to develop ICT skills and enter relevant fields.

Options for media actors

- Continue to report on gender issues related to AI research and applications.

Options for intergovernmental organizations, including UNESCO

- Continue to promote women in STEM careers.
- Convene ongoing dialogues about AI and gender.
- Ensure gender diversity as well as the inclusion of marginalized groups in the multi-stakeholder dialogues on AI issues.
- Elaborate frameworks and methodologies to measure digital skills and the digital skills gap.

AI IN AFRICA

6

CHAPTER 6: AI IN AFRICA

Africa is a UNESCO global priority, with a specific focus on peace-building and institutional capacity enhancement in pursuit of the Sustainable Development Goals. UNESCO recognizes that technology can contribute to the achievement of these goals in Africa. Its operational strategy for Priority Africa includes a flagship programme on "harnessing STI and knowledge for the sustainable socio-economic development of Africa," and lists "[building] capacity in the field of ICT" as one of its main actions (UNESCO, 2014).

Africa's population is currently estimated at 1.1 billion inhabitants and, according to United Nations forecasts, it will rise to 2.4 billion, or nearly one-third of the world's population, by 2050. Rapid population growth means that over 60 per cent of the African population is under 35, which makes the African workforce one of the youngest in an ageing world (United Nations Population Division, 2019).

The advances in digital technologies like AI have opened up opportunities to accelerate social and economic development, but for that to be accomplished Africa needs to harness its demographic dividend, create adequate infrastructure and address emerging policy challenges at the intersection of technology and society. On the one hand, AI in developed economies may lead to less sourcing of labour to developing countries with adverse effect on African employment (Nayyar, 2019). On the other hand, African countries risk relegation to colonial models, being suppliers of raw materials, in these cases, data and skilled personnel (brain drain), and remaining as wholly dependent importers of AI technology and services. Instead, they could become active in all stages of the AI-value chain, which is necessary if the technologies involved are to be of optimum benefit to the continent.

Recognizing these kinds of challenges, the African Union has developed several continental frameworks to achieve 'Agenda 2063: The Africa We Want', and in particular through pan-Africa coordinated strategic action in the fields of Agriculture, Infrastructure, Mining, Science, Technology and Innovation, Trade and Industrial Development (African Union, 2015).

In 2018, the Outcome Statement of UNESCO's Forum on Artificial Intelligence in Africa also recognized "the expeditious growth of Africa's population, as well as the opportunities and challenges this poses in terms of education, training and the employability of African youth" and the potential that AI offers for "sustainable and inclusive development on the continent." Participants expressed concern regarding "enduring inequalities and significant disparities in the availability of the resources, capacities

and infrastructures required for giving access to, and fully benefiting from, the results of scientific innovation" (UNESCO, 2018b).

This chapter recognizes that while Africa is heterogeneous, a number of countries in the region experience several common fundamental challenges, including internal conflict and violence, poor human rights observation, unstable institutions, and a lack of infrastructure, sustainable financing and institutional capacity (Besaw & Filitiz, 2019). Other countries have less daunting obstacles in terms of adoption and integration of AI into their development paths.

Against this mixed background, Section 1 starts with a discussion of the challenges within the broader sphere of science, technology and innovation (STI) for Africa. It discusses some important trends and background for STI and then highlights the strategic objectives and priority areas outlined by the African Union in pursuance of Agenda 2063: The Africa We Want.

Section 2 provides an overview of initiatives announced, initiated or implemented by several African governments for creating an enabling environment for harnessing AI for the purposes of sustainable development. It shows that while some governments are proactive, many others are yet to outline clear strategies for AI.

Finally, Section 3 discusses some of the initiatives taken by the private sector, universities, technical community and the civil society for strengthening AI for good in Africa. The efforts of these actors in the face of many challenges are helping the continent to develop its AI ecosystem through better education, knowledge, training, and skills and networking opportunities.

1. THE CHALLENGES OF SCIENCE, TECHNOLOGY AND INNOVATION IN AFRICA

The African Union Commission (AUC) has established a Conference of Ministers in charge of Science and Technology (AMCOST) to enable the Union to discuss and develop collective responses to issues in the field of Science and Technology. The Consolidated Plan of Action (CPA) on Science, Technology and Innovation (STI) was presented to Heads of State and Government in 2005 and was endorsed for implementation in 2006 by the African Heads of State. The Science, Technology and Innovation Strategy for Africa 2024 (STISA-2024) places STI capacity-building, knowledge production and technological innovation as central to Africa's social and economic development as part of AU Agenda 2063 (African Union, 2014b). Yet most African countries have not met the African Union target of spending 1 per cent of the GDP on R&D (UNESCO, 2015c). In addition, as per the UNESCO Science Report 2015, there is relatively low-level political commitment to STI on the part of individual countries (UNESCO, 2015c).

STISA-2024 recognizes the following challenges facing STI in Africa (African Union, 2014b):

i) Insufficient funding for STI with only half of the investment in R&D coming from within Africa

ii) Low organizational capacities for STI policy development and lack of evidence-based policy-making due to low skills and training of staff, limited access to data and knowledge about state of the art of AI

iii) Different levels of infrastructure readiness across Africa that could support innovation

iv) While civil society organizations' engagement with AI is emerging across the continent, their inputs to STI policy debates are often not supported by evidence or comprehensive research

v) Bilateral and multilateral cooperation is improving, but it often does not adequately promote African ownership, accountability and sustainability.

The UNESCO Science Report 2015 further notes that the commitment to STI varies greatly across countries, and in some instances, there is a lack of:

i) National research and innovation strategies or policies with a clear definition of measurable targets and the role to be played by each stakeholder

ii) Involvement of the private sector in the process of defining national research needs, priorities and programmes

iii) Institutions devoted to innovation that can make the link between research and development (UNESCO, 2015c).

The strategic objectives of the African Union's Science, Technology and Innovation Agenda include enhancing the use of science and technology to address AU's priority areas,[1] improving technical competencies and institutional capacity for STI development, promoting economic competitiveness, protecting knowledge production and strengthening Intellectual Property Rights and facilitating STI policy reforms, harmonization and resource mobilization.

Evaluation reports for STISA-2024 are not yet available, as the original strategy envisaged an evaluation only in 2024. However, stakeholders are proposing a mid-term evaluation and the development of a monitoring and evaluation framework based on a results based management (RBM) approach (Daniels, Mawoko, & Konte, 2018).

---

1 The AU STISA-2024 priorities include: i) Eradicating hunger and ensuring food and nutrition security, ii) Preventing and controlling diseases and ensuring well-being, iii) Communication (Physical and Intellectual Mobility), iv) Protecting space (climate, biodiversity, space, marine and sub marine), v) Living together and building the society (pan Africanism and regional integration, governance) and vi) Creating wealth (education and human resource development, mineral and water resources).

Meanwhile, as Prof. Sarah Anyang Agbor, Commissioner for Human Resources, Science and Technology for the African Union Commission has noted: "the development and the use of artificial intelligence" should be "supported by an enabling policy environment with proper instruments and regulation to enable us to reap its benefits in a secure, equitable and sustainable manner" (Agbor, 2019). She disclosed that a comprehensive digital transformation strategy is being developed for Africa with a strong focus on South-South and North-South cooperation. In this context, it is important for stakeholders to work together to develop an enabling environment in the African region in which the STI strategic objectives outlines above can be achieved.

2. INITIATIVES TOWARDS AI BY GOVERNMENTS IN AFRICA

The 2019 Government Artificial Intelligence Readiness Index places 12 African countries in the top 100 and none in top 50 with regard to the Government's Readiness to use AI under four broad clusters that include Governance, Infrastructure and Data, Skills and Education and Public Services (Miller & Stirling, 2019). The top five placed African governments as per the index are Kenya, Tunisia, Mauritius, South Africa and Ghana.[2]

Despite a range of challenges such as those discussed above, several governments have taken initiatives that have relevance to the development of AI, for example:

In 2018, the Kenyan government formed a Blockchain and AI Taskforce that recently published its report with recommendations on using these technologies to eliminate corruption, strengthen democracy, facilitate financial inclusion and improve the delivery of public services among others (Ministry of Information, Communications and Technology, 2019; Kenyan Wall Street, 2018). In addition, The Kenya Open Data portal provides free digital access to Government datasets in easily usable formats to facilitate government accountability through citizens' engagement (ICT Authority Kenya, 2019). Kenya is also set to use AI to assess citizens' eligibility for affordable housing.

Nigeria approved a robotics and AI agency to "leverage collaborations with international research bodies on robotics and AI" and enhance education and skills of young people through "research and teaching of more complex technologies" (Goitom, 2019).

South Africa has established a 'Fourth Industrial Revolution Commission', chaired by its President. The commission is expected to "recommend policies, strategies and plans to position SA as a competitive player in the digital space" (Phakathi, 2019).

Mauritius launched its Artificial Intelligence Strategy in 2018 that recommended the

2 The Government Readiness Index 2019 acknowledges that the indicator is not perfect and also that the low score of some countries may not capture recent initiatives for AI.

creation of the Mauritius Artificial Intelligence Council (MAIC) and asked the government to ensure a conducive environment "through a robust and yet friendly regulatory, ethics and data protection environment and also through attractive incentives such as matching grants, tax credits and other fiscal incentives, training grants for investments in AI and other emerging technologies" (Working Group on Artificial Intelligence, 2018).

Malawi has devised a National ICT Plan 2014-2031 envisioning the country as a knowledge-based economy and has come up with the Digital Malawi Project. Much weight is given to the need for ICT development, given the county's relatively low Internet penetration (Public Private Partnership Commission, 2017).

Ghana, while not having a dedicated AI strategy, has also mentioned AI as a strategic technology area upon which it would focus in its Science and Technology Innovation policy (The Presidency of the Republic of Ghana Communications Bureau, 2019).

Elsewhere on the continent,[3] in 2018 the Tunisian National Agency for Scientific Research Promotion initiated a process to develop a national AI policy at a workshop hosted by the UNESCO Chair on Science, Technology and Innovation Policy (ANPR, 2018).

Morocco jointly hosted the 2018 '*Forum on Artificial Intelligence in Africa*' with UNESCO and has allocated 50 million MAD to fund research projects related to AI in 11 topic areas (Zerrour, 2019; UNESCO, 2018c).

In addition, 14 out of 55 African countries have signed the 'African Union Convention on Cyber Security and Personal Data Protection' that sets out the security rules essential for establishing a credible digital space for electronic transactions, personal data protection and combating cybercrime (African Union, 2014a).

Several other countries have expressed positive sentiments about AI and leveraging technology for sustainable development.

For example, Namibia's Minister of Industrialisation, Trade and SME Development, Mr. Tjekero Tweya, has stated that Namibia should be an active player in the 'Fourth Industrial Revolution' (NAMPA, 2019).

There is need for a systematic study to understand the needs of the African governments and societies in responding to the challenges of AI and other digital technologies. Such an assessment, particularly using the UNESCO ROAM-X framework, could help catalyze and inform evidence-based policy-making, which is a need that has been highlighted in the AU's STISA-2024 (African Union, 2014b).

3 North Africa falls under the Arab States region at UNESCO.

3. PRIVATE SECTOR, TECHNICAL COMMUNITY AND CIVIL SOCIETY INITIATIVES FOR AI IN AFRICA

As discussed above, several governments in Africa are working towards creating an enabling environment for Science, Technology and Innovation, even if the overall pace of change is limited by structural challenges as detailed above. However, in many countries there is a vibrant sector outside the government that is working to strengthen AI research, knowledge, skills and policies. Young entrepreneurs, global tech firms, researchers and students are harnessing AI elements as a business opportunity and for the development of their communities and countries.

3.1. Private sector

Actors in the private sector are investing both financial and human resources to increase diversity, strengthen AI research in Africa, and develop solutions that meet Africa's challenges. Google established its first African AI research hub in Accra, Ghana, in 2019, and has expressed its commitment to collaborating with local universities and research centers (Crabtree, 2018). IBM has several research centers in Kenya and South Africa and has invested USD 70 million to launch a Watson-powered learning platform for 25 million African youth. This online platform will offer free skills-development programs across Africa (Jao, 2017). Tunisia's InstaDeep, which provides AI-powered decision-making systems for firms, already has offices in London, Paris, Tunis, Nairobi and Lagos, and Uganda's GeoGecko has worked with UNICEF to create maps on social protection infrastructure (InstaDeep, 2019; Geo Gecko, n.d.).

Table 2 presents a non-exhaustive list of initiatives in Africa in the field of health, agriculture, fintech and transportation that are using AI and/or its elements to serve customers and communities.

Sector	Initiative	Country
Health	**Sophie Bot** An AI-powered chat bot designed to answer questions about sexual health. Its creators were looking to resolve the problems of access to credible information in real time and awkwardness in talking about reproductive health issues amongst young people in Africa. Sophie Bot learns from conversations with users to process and reply to questions (Mbaka, 2017).	Kenya

Sector	Initiative	Country
Health	**Automated malaria diagnosis with digital microscopy** Uganda's first AI lab, at Makerere University, has developed a way to diagnose blood samples using a cell phone. The program learns to create its own criteria for infections based on a set of images that have been presented to it previously. Diagnosis times could be slashed from 30 minutes to as little as two minutes (Lewton & McCool, 2018).	Uganda
	Numberboost An AI system that allows citizens to locate nearby mobile healthcare clinics, thus improving access to healthcare (NumberBoost, 2019).	South Africa
Agriculture	**Vital signs** Using on-the-ground measurements of a variety of indicators and existing data sources, Vital Signs creates a picture of the relationship among agriculture, nature and human well-being. Vital Signs 'Key Indicators' include: sustainable agricultural production, water availability and quality, soil health, biodiversity, carbon stocks, climate resilience, household income, nutrition and market access (Conservation International, 2018).	Kenya
	Arifu Arifu is a chatbot platform for learning new skills about various topics including entrepreneurship, financial management, or nutrition. It can, for example, help farmers to determine what fertilizer matches their specific needs (Arifu, 2019).	
	FarmDrive FarmDrive connects smallholder farmers to loans and financial management tools, offering tools such as keeping records of expenses, revenues and yields; applying for loans; receiving loans; and reminders about loan repayment. For financial institutions, FarmDrive can use their data on size of land, location, and crops to determine the risk and corresponding interest rates (Owino, 2019).	

Sector	Initiative	Country
Agriculture	**Zenvus** Zenvus is a decision-making tool for farmers based on data collected from sensors and other means. Zenvus' services include keeping record of all phases of farming, from planting to sales; raising capital and crowdfunding; insuring farms; providing real-time produce prices and a platform to sell produce (Zenvus, 2019).	Nigeria
	Aerobotics Aerobotics assists the agricultural industry by using drone aerial imagery to manage orchards, identify problems in crop yields, and manage pests and diseases (Aerobotics, 2019).	South Africa
Fintech	**Tala** Tala is an online credit product that instantly underwrites and disburses loans to individuals based on data they input into the app, including those who do not have a formal credit history. Repayment of loans also occurs through the application (Tala, 2019).	Kenya
	Kudi.ai Using natural language processing and artificial intelligence, Kudi.ai attempts to make peer-to-peer payment easier for Nigerians using a chatbot that works on popular messaging apps, like Facebook Messenger. Users of Kudi can transfer cash to one another, help others to transfer cash, and pay for their television, Internet and electricity bills (Kudi, 2019).	Nigeria
Public Transportation	**RoadPreppers** RoadPreppers help users to navigate traffic congestion by giving them alternative driving directions and public transport options with fare quotes. They specialize in doing so in regions where public transit data is inaccessible and unstructured, or simply unavailable, and where public transport systems are complicated (Eweniyi, 2017).	Nigeria

Sector	Initiative	Country
Public Transportation	**lara.ng** A chatbot that provides public transportation directions and fares for commuters in Lagos. It seeks to be better than existing options, which do not perform well in cities where the transport network is congested and complicated (Eweniyi, 2017).	Nigeria

Table 2: Initiatives in Africa using AI in health, agriculture, fintech, and transportation

3.2. Universities and educational institutes

Efforts are underway to strengthen AI education within Africa. One initiative is the launch of the African Masters of Machine Intelligence (AMMI) degree programme at the African Institute of Mathematical Sciences in Kigali, Rwanda. The programme, launched in partnership with Google and Facebook in 2019, is committed to providing state of the art research exposure and teaching by experts to African students within Africa.[4] The first cohort comprised 30 students from 10 African countries, 43 per cent of whom were female (AIMS, 2018).

Other examples include the Euromed School of Digital Engineering and AI in Morocco, which will open its doors in September 2019. In Uganda, there is already an active AI and Data Science research group at Makakere University (Lystad, 2019; AI & Data Science Research Group at Makerere University, 2019). The University of Namibia has added an AI module to its Bachelors programme in Computer Science (Namibia University of Science and Technology, 2019).

Such specialized programmes at universities will help develop capacities for AI in Africa by building on foundational computer science courses taught at undergraduate level. There are several more initiatives at the university level and a detailed survey of needs at the university level would help identify the gaps that are needed to be filled to enhance availability and access to high quality AI education.

3.3. Civil society and the technical community

These formal efforts to strengthen AI knowledge within Africa are being complemented by a number of initiatives to bridge the skills gap by using innovative teaching and learning models. For instance, Ms. Tejumade Afonja co-founded AI Saturday Lagos, an AI community in Lagos, Nigeria. Through free classes on data science, machine learning and deep learning for 16 consecutive Saturdays in the form of struc-

4 Facebook is contributing four million USD in funding and staff lecturers, while Google is also contributing resources (AIMS, 2018).

tured study groups, they have trained over 150 individuals in two cycles (Afonja, 2018). 'AI Kenya' is another community with about 2,500 members from different fields that has taken an initiative aimed bringing together technical and non-technical experts in East Africa to develop solutions for local problems by leveraging AI (AI Kenya, 2019).

Data Science Nigeria runs an AI Hub where AI is taught every day free-of-charge to a large number of students who attend daily online classes (Data Science Nigeria, 2019). Their free eBook provides 99 use cases of AI and simplifies AI for beginners (Adekanmbi, 2018). There are several local communities, often functioning through social media platforms that bring together likeminded individuals to learn and work on solutions using AI. Examples of such initiatives include 042 AI in Enugu and TensorFlow Lagos among others (TensorFlow Lagos, 2019). These communities are not only addressing the challenges of access to AI knowledge and training, they are also fostering an ecosystem where mentorship is available to guide young people.

Box 22: Representation of the African diaspora in the Western AI community

Even as technology becomes increasingly important in our lives, there is a lack of diversity worldwide amongst those studying, researching, teaching or developing solutions in the fields of Science, Technology, Engineering and Mathematics (STEM). The founders of the 'Black in AI' initiative noticed the "crises of diversity" in the AI community when they realized that there were only six black researchers out of an estimated 8,500 people that participated at the Neural Information and Processing Systems (NeurIPS) conference in 2016, an important international conference for the AI community. While 'Black in AI' is not an African organization, its message about the importance of diversity and representation of black people in the AI community is consistent with the need for African representation as well. (Cisse, 2018).

Ms Timnit Gebru, one of the co-founders of the initiative, stresses the importance of diversity and believes that "if we don't have diversity in our set of researchers, we are not going to address problems that are faced by the majority of people in the world." She adds that when "problems don't affect us; we don't think they're that important, and we might not even know what these problems are, because we're not interacting with the people who are experiencing them" (Snow, 2018a).

In terms of access to knowledge, the chapter on Access discussed how academic conferences are an important avenue for AI researchers to meet, present their research, and develop partnerships and networks. Access to global conferences has not been easy for African researchers for several reasons, including denial of visas to participate at these conferences in developed countries and high costs associated

with travel and stay (Knight, 2018). A positive course correction in this direction, which may have longer term significance is that the International Conference on Learning Representations (ICLR), a major gathering of the AI community, will be organized in Addis Ababa in 2020.

Deep Learning Indaba is another organization that is strengthening African capacity in regard to machine learning. It aims to "build communities, create leadership, and recognize excellence in research and innovation across the continent" (Deep Learning Indaba, 2019). Through their annual conferences and regional Indaba meetings, they are furthering the research agenda as well as providing a platform to African researchers and students to engage and collaborate. Two highly visible spin-offs from the conference in 2018 were a collaboration between researchers to develop Neural Machine Translation for African Languages and the development of an AI that is able to generate African masks (Wiggers, 2018; Abbott, 2018). UNESCO organized two workshops on Fairness and AI at the Deep Learning Indaba 2019, with the objective of engaging the AI technology and policy community in Africa to work together on human rights, openness and access concerns related to AI (UNESCO, 2019f).

Box 23: AI for African languages: Strengthening multilingualism

Africa has over 2,000 languages (Wolff, 2018). Given this linguistic diversity, there are questions concerning access to information in multiple languages. For instance, up to the mid-1990s, an estimated 80 percent of the content online was in English. In 2015, just 10 languages constituted about 82 percent of the content on the Internet (Young, 2015). Further, evidence suggests that learning efficiency and cognitive development in children is much better when their mother tongue is the medium of instruction in schools (Alidou, et al., 2006).

AI and Natural Language Processing is progressing towards live translation of several languages. In Africa, this project has taken root and now tools are available to translate thousands of African languages. OBTranslate is a digital platform that translates over 2000 languages enabling better communication amongst speakers of different African languages (The Guardian, 2019). Such tools can also be used to translate online content, thereby strengthening access to information on the web.

The Human Language Technology Research Group within the Council for Scientific and Industrial Research (CSIR) are studying the ways in which speech and language technologies can be used to benefit South Africans. In particular, they are exploring the use of automatic speech recognition to support language learning and translation, an important task given that there are 11 official languages in South Africa, with only English and Afrikaans well-represented in existing technologies (GOBL, 2014).

All these initiatives are spurring active interest in Africa's AI ecosystem and creating a community of young people interested in AI. In 2019, there were multitudes of homegrown applications drawing on elements of AI, which aid Africans in their work and daily lives, including farming applications, educational applications and sexual health-related applications (Le Monde, 2019; Mbaka, 2017; Halilou, 2016). An example among many others is Lifantou, founded by 28-year-old Senegalese engineer Ms. Awa Thiam, an e-commerce platform that uses AI to link school canteens, which need low-cost ingredients, and agricultural cooperatives which wish to avoid paying intermediaries (Le Monde, 2019). Although AI development in Africa is still in its early stages, the technology and its elements have shown some potential in affording a number of young people in Africa to leverage technology to resolve some of the continent's challenges.

4. CONCLUSION AND POLICY OPTIONS

Africa's potential to transform a number of countries into innovation hubs can be realized if the right environment and opportunities are provided. This chapter presented efforts to harness AI for development on the continent against the backdrop of several structural and foundational challenges facing several African countries. The importance of Science, Technology and Innovation is well recognized by African countries and forms an essential part of the African Union's vision until 2063. However, there are significant capacity, infrastructure and governance challenges in building a strong enabling environment for AI development.

Many African governments are cognizant of these challenges and are taking initiatives, some through AI specific policies, which help to empower the private sector, researchers and civil society to harness AI for sustainable development. While government efforts have started, they are complemented by the private sector, technical community and civil society trying to address the immediate challenges of access to knowledge, skills, mentorship and business opportunities.

Some options for action to strengthen the Science, Technology and Innovation ecosystem needed for the development of AI in Africa are presented below:

Options for States

- Develop national AI policies and strategies in line with UNESCO's ROAM principles that are accompanied by an implementation plan, funding mechanisms and monitoring and evaluation processes.
- Invest more in science and engineering education in a gender-responsive way in order to develop the skilled labour force necessary for the development of AI.
- Enhance efforts to reach the national target of investing 1 per cent of GDP in research and development (R&D).
- Encourage the business sector to participate more actively in R&D, in order to stimulate demand for knowledge production and technological development.
- Set up national funds to help local innovators protect their intellectual property rights.

- Facilitate collaboration between the private sector and universities by measures including making provision for representatives of the private sector to sit on the governing boards of universities and research institutes, tax incentives to support business innovation, the creation of science and technology parks and business incubators to encourage start-ups and public–private partnerships and research grants to support collaborative research between the government, industry and academia in priority areas.
- Foster exchanges, intraregional and pan-Africa collaboration among researchers and create incentives to counter brain-drain.

Options for the private sector, Internet intermediaries and the technical community

- Support universities and research centres with collaborative projects and shared access to data, hardware and knowledge in order to strengthen research and development in AI
- Develop knowledge tools and modules for AI education and training to train students
- Create AI technologies to solve issues related to health, agriculture, finance, transportation, etc.
- Provide opportunities for re-skilling of existing workforce for advanced AI-based applications in businesses and society.

Options for academia

- Collaborate with researchers across Africa and in other regions to conduct cutting edge fundamental and applied research in AI.
- Update educational curricula at the university and school level with the latest and most relevant knowledge that enables students to exploit career opportunities in AI and other fields related to computer science.
- Foster an environment of collaboration between academia and the private sector to prepare students who are employable and for research that can be commercialized.

Options for civil society

- Strengthen capacities in evidence-based science, technology and innovation policy making in Africa.

- Strengthen cooperation between civil society and research institutes for solving problems facing local communities, for novel data collection models based on citizen science that can create data sets for AI that respect international norms for privacy and data protection.
- Actively participate in policy dialogues in order to bring citizens' concerns in front of policy makers but also to underpin AI policy-making and ensure that AI use does not infringe on human rights, including the rights to expression and access to information, privacy, equality and participation in public life.

Options for inter-governmental organizations, including UNESCO

- Collaborate and partner with the African Union on its digital transformation strategy, partner with regional and national policy making organizations for evidence-based AI policy-making.
- Support the development of upstream and downstream capacities in addressing the challenges of AI and other advanced ICTs through trainings, workshops and long term institutional support programmes.
- Facilitate North-South and South-South knowledge exchange for greater access to research and knowledge between different regions.

CHAPTER 7: IMPLICATIONS FOR UNESCO AND OVERALL OPTIONS FOR ACTION

The development of AI as a package of technologies is inextricably linked to its evolution within the wider ecosystem of the Internet and ICTs and the forces that have shaped this context. This highlights the pertinence of UNESCO's approach to the Internet, and the importance of viewing AI in terms of potential alignment with the Organization's Internet Universality concept.

UNESCO's position on AI can therefore be appropriately framed with Internet Universality ROAM principles – which would then advocate for AI to develop in terms of Human Rights, Openness, Accessibility and Multi-stakeholder participation. This approach can serve as a well-grounded and holistic framework for UNESCO and stakeholders, which can help to inform and shape the design, application, monitoring and governance of AI. It can nourish activities for the setting of particular normative and ethical principles for AI, producing innovative policy guidelines and toolkits, and for developing AI-specific indicators of relevance to the range of areas of UNESCO's mandate, including but not limited to, communication and information.

With this approach to AI, UNESCO's is well placed for providing technical and policy advice, serving as a clearinghouse for information, and building capacity. In this manner, AI that is informed by the ROAM principles can contribute to the benefit of humanity, sustainable development and peace.

This would mean UNESCO bringing a distinctive approach within the ecosystem of other actors with interests in the AI field. With access to duty-bearers (government officials, elected representatives, independent regulators, media owners and leaders, relevant specialized NGOs etc.), and credibility amongst rights-holders (citizens, journalists, academics, private sector, etc.) UNESCO can effect positive change for development and use of AI.

One example is that commemoration of international days such as World Press Freedom Day and the International Day for Universal Access to Information allows for the Organization to integrate AI issues into matters to do with press freedom, disability, and universal access to information (UNESCO, 2019b; UNESCO, 2019f). In its other work, in education, culture and the sciences, it is also the case that research, capacity-building, awareness raising, advocacy, technical advice and sharing of good practices and international experience can be brought to bear in relation to AI. In this way, UNESCO can be an effective part of broader work around the world to shape the design, application and governance of AI.

The outcomes of such work will contribute to the setting and application of rights, norms and standards on the ethical and human rights dimensions of AI, and in alignment with the principles of openness, accessibility and multi-stakeholder governance. In this way, AI can be harnessed for achieving the range of SDGs, and not least SDG 16.10 on "public access to information and fundamental freedoms." The outputs that underpin such outcomes include not only innovative solutions produced for steering AI, but also capacity being built especially in Africa and in favour of gender equality.

Overarching Options for Action

All stakeholders can consider addressing AI through below approaches:

- Apply UNESCO's Internet Universality principles (Human Rights, Openness, Accessibility and Multi-stakeholder participation), and develop tailored indicators on AI, in order to research, map and improve the ecosystem in which AI is developed, applied and governed.
- Participate in interdisciplinary research on how AI intersects with human rights, openness, accessibility and multi-stakeholder governance. Promote ethics-by-design in AI development and apply human rights norms and standards that can inform the emergence of more specific guidelines about the right to freedom of expression and access to information, the right to privacy, the right to equality, and the right to participation in public life.
- Facilitate the formulation of international human rights-based ethics frameworks and global policy solutions and facilitate the inclusive participation from developed and developing countries.
- Conduct comprehensive human rights impact assessments of AI development.
- Reflect on the implications of AI on the practice of journalism and media development, and encourage media actors to investigate and report on AI development and its applications, including exposure of abuses and biases of AI, as well as current benefits.
- Assess algorithmic discrimination in order to protect the right to equality of all, in particular of historically marginalized populations.
- Raise awareness of ownership and access to big data, AI skills and technologies, and the issues of who benefits, as well as harms such as marginalization or manipulation of human agency.
- Promote open access to research in AI and the development of data publication models that safeguard against the infringement of human rights due to misuse of openly available knowledge about AI.
- Uphold open market competition to prevent monopolization of AI, and require adequate safeguards against violation of ethical practices by market and other actors.

- Develop industry-wide ethical guidelines for use of AI in order to ensure that open market competition in the development of AI-based applications does not infringe upon human rights.
- Facilitate the development of norms and policies for improving openness and transparency in AI algorithms through:
  - *Ex-ante* information disclosure on the intent and purpose of AI algorithms
  - *Ex-post* monitoring of algorithmic decision-making to ensure its alignment with designed intent of the algorithm
  - Transparency about the data used to train the algorithms and the data used for predictive analysis with provisions for users to seek information about how their information is processed
- Ensure that open data does not compromise the privacy of individuals and that it conforms to data protection laws.
- Facilitate open reflection on the evolution of research trends in AI through collection of up-to-date gender and geographically disaggregated statistics on journal and conference publications in AI.
- Work to reduce digital divides, including gender divides, in AI access, and establish independent monitoring mechanisms.
- Develop monitoring mechanisms by collecting, triangulating and validating gender and geographically disaggregated data, which can inform stakeholders of unintended impacts of AI as well as issues such as the state of the digital divide in access to AI, including STEM and STI.
- Encourage more active participation in AI governance – ranging from principles through to rules as appropriate, and as per the different roles and obligations of stakeholder groups, including but not limited to government, the private sector, the technical community, civil society, academia, international organizations and the media.
- Ensure gender equality, linguistic and regional diversity as well as the inclusion of youth and marginalized groups in multi-stakeholder dialogues on AI issues.
- Ensure the transparency, inclusiveness and accountability of the participation process with stakeholders participating on an equal footing.

- Work with UNESCO to integrate discussion of AI issues into relevant events such as international days around press freedom, disability, and universal access to information, and draw in UNESCO networks such as UNITWIN, Orbicom, GAPMIL, and GAMAG, as well Category 2 Institutes and Centres, NGOs, IFAP National Committees and UNESCO National Commissions.
- Give particular attention to the interface between issues related to AI, gender equality and sustainable development in Africa.
- Build capacity for development and application of AI that works to advance the universal Sustainable Development Goals.

BIBLIOGRAPHY

Abbott, J. (2018, December 17). *The Journey to NeurIPS*. Retrieved October 18, 2019, from Towards Data Science: https://towardsdatascience.com/the-journey-to-neurips-ee1a197da538

Académie Nationale des Sciences et Techniques du Sénégal. (2019, June 28). *Célébration de la Journée de la Renaissance Scientifique de l'Afrique (JRSA) 2019*. Retrieved October 17, 2019, from Académie Nationale des Sciences et Techniques du Sénégal: https://www.ansts.sn/celebration-de-la-journee-de-la-renaissance-scientifique-de-lafrique-jrsa/

Access Now. (2018). *Human Rights in the Age of Artificial Intelligence*. New York: Access Now.

Adekanmbi, O. (2018). *Artificial Intelligence Simplified: 99 Use Cases and Expert Thoughts*. Lagos: Data Scientists Network Foundation.

Adshade, M. (2018, August 14). *How Sex Robots Could Revolutionize Marriage—for the Better*. Retrieved October 18, 2019, from Marina Adshade: http://marinaadshade.com/2018/08/14/how-sex-robots-could-revolutionize-marriage-for-the-better/

Aerobotics. (2019). *Defenders of The Tree Crop*. Retrieved October 18, 2019, from Aerobotics: https://www.aerobotics.com/?identifier=default-get-in-touch-button

Afonja, T. (2018, December 30). *Bridging The Artificial Intelligence (AI) Gaps With AISaturdays*. Retrieved October 18, 2019, from Medium: https://medium.com/ai-saturdays/bridging-the-artificial-intelligence-ai-gaps-with-ai6-9a5cf0b910f8

African Union. (2014a). *African Union Convention on Cyber Security and Personal Data Protection*. Addis Ababa: African Union.

African Union. (2014b). *Science, Technology and Innovation Strategy for Africa 2024*. Addis Ababa: African Union.

African Union. (2015). *Continental Frameworks*. Retrieved October 17, 2019, from African Union Agenda 2063: https://au.int/en/agenda2063/continental-frameworks

Agbor, S. A. (2019, August 28). AUC Commissioner for HRST on Artificial Intelligence (AI). (T. 2019, Interviewer)

Agre, P. E., & Rotenberg, M. (1988). *Technology and Privacy: The New Landscape*. Cambridge: MIT Press.

AI & Data Science Research Group at Makerere University. (2019). *About the AI & Data Science Research Group*. Retrieved October 18, 2019, from AI Research Artificial Intelligence and Data Science: http://www.air.ug/

AI Kenya. (2019). *Who We Are*. Retrieved October 18, 2019, from AI Kenya: https://kenya.ai

AIMS. (2018, July 31). *AIMS launches African Master's in Machine Intelligence*. Retrieved October 18, 2019, from AIMS: https://www.nexteinstein.org/blog/2018/07/31/aims-launches-first-of-its-kind-african-masters-in-machine-intelligence-at-rwanda-campus/

Algorithm Watch. (2019). *Automating Society Taking Stock of Automated Decision-Making in the EU*. Berlin: Algorithm Watch.

Algorithmia. (2018, March 6). *Hardware for Machine Learning*. Retrieved October 18, 2019, from Algorithmia: https://blog.algorithmia.com/hardware-for-machine-learning

Alidou, H., Boly, A., Brock-Utne, B., Diallo, Y. S., Heugh, K., & Wolff, H. E. (2006). *Optimizing Learning and Education in Africa - the Language Factor*. Tunis Belvédère: UNESCO Institute for Lifelong Learning and the Association for the Development of Education in Africa.

Allcott, H., & Gentzkow, M. (2017). *Social Media and Fake News in The 2016 Election*. Journal of Economic Perspectives, 31(2), 211-236.

Altman, A. (2016). Discrimination. In E. N. Zalta, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Amazon. (2019). *Alexa and Alexa Device FAQs*. Retrieved October 18, 2019, from Amazon: https://www.amazon.com/gp/help/customer/display.html?nodeId=201602230&pop-up=1

Andersen, L. (2018). *Human Rights in the Age of Artificial Intelligence*. New York: AccessNow.

Angwin, J. (2010, July 30). The Web's New Gold Mine: Your Secrets. Retrieved October 18, 2019, from The Wall Street Journal: https://www.wsj.com/articles/SB10001424052748703940904575395073512989404

Angwin, J. (2017, November 9). *Cheap Tricks: The Low Cost of Internet Harassment*. Retrieved from ProPublica: https://www.propublica.org/article/cheap-tricks-the-low-cost-of-internet-harassment

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias*. Retrieved October 18, 2019, from ProPublica: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ANPR. (2018, April 20). *National AI Strategy: Unlocking Tunisia's capabilities potential*. Retrieved October 18, 2019, from Agence Nationale de la Promotion de la Recherce scientifique: http://www.anpr.tn/national-ai-strategy-unlocking-tunisias-capabilities-potential/

Arifu. (2019). *About*. Retrieved October 18, 2019, from Arifu: https://www.arifu.com/

ARTICLE 19 & Privacy International. (2018). *Privacy and Freedom of Expression In the Age of Artificial Intelligence*. London: ARTICLE 19 & Privacy International.

ARTICLE 19. (2019). *The Social Media Councils: Consultation Paper*. London: ARTICLE 19.

Article 29 Data Protection Working Party. (2014). *Opinion 05/2014 on Anonymisation Techniques*. Brussels: Article 29 Data Protection Working Party.

Ashcraft, C., McLain, B., & Eger, E. (2016). *Women in Tech: The Facts*. Colorado: National Center for Women and Information Technology.

Ayyub, R. (2018, November 21). *I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me*. Retrieved October 19, 2019, from Huffington Post: https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316?ncid=fcbklnkcahpmg00000001&guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZmFjZWJvb2suY29tLw&guce_referrer_sig=AQAAAIEuYmZBqcndRw1Srh241z2hjVR39MBJf-Lrfpgarwk1FSoBeFtsfRJQxRlklPg

Azoulay, A. (2018, March). *Making The Most of Artificial Intelligence*. (J. Šopova, Interviewer) 2018-3 Issue of UNESCO Courier. Retrieved 16 November 2019 at https://en.unesco.org/courier/2018-3/audrey-azoulay-making-most-artificial-intelligence

Baghai, K. (2012). *Privacy as A Human Right: A Sociological Theory*. Sociology, 46(5), 951-965.

Balkin, J. M. (2017). *Free Speech in The Algorithmic Society: Big Data, Private Governance and New School Speech Regulation*. UC Davis Law Review, 51(3), 1149-1210.

Barocas, S. (2014). *Data Mining and the Discourse on Discrimination*. Proceedings of Data Ethics Workshop (pp. 1-4). New Jersey: Github.

Barocas, S., & Selbst, A. D. (2016, September 30). *Big Data's Disparate Impact*. California Law Review, pp. 671-732.

Barraclough, T., & Barnes, C. (2019, May 22). *Don't Believe Everything You See, or Hear*. Retrieved October 18, 2019, from newroom.: https://www.newsroom.co.nz/2019/05/22/597617/hold-dont-believe-everything-you-see-or-hear

Bathaee, Y. (2018). *The Artificial Intelligence Black Box and the Failure of Intent and Causation*. Harvard Journal of Law and Technology, 31(2), 889-938.

Bayefsky, A. F. (1990). *The Principle of Equality or Non-Discrimination in International Law*. Human Rights Law Journal, 11, 1-34.

BBC. (2017, September 11). *Row over AI that 'identifies gay faces'*. Retrieved October 18, 2019, from BBC: https://www.bbc.com/news/technology-41188560

Been, K., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018, June 7). *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).* Retrieved October 18, 2019, from Cornell University: https://arxiv.org/pdf/1711.11279.pdf

Benay, A. (2018, October 10). *Using Artificial Intelligence in government means balancing innovation with the ethical and responsible use of emerging technologies.* Retrieved October 18, 2019, from The Conversation: https://tbs-blog.canada.ca/en/using-artificial-intelligence-government-means-balancing-innovation-ethical-and-responsible-use

Benjamin, G. (2019, February 6). *Deepfake videos could destroy trust in society - here's how to restore it*. Retrieved October 18, 2019, from The Conversation: http://theconversation.com/deepfake-videos-could-destroy-trust-in-society-heres-how-to-restore-it-110999

Benkler, Y., Faris, R., Bourassa, N., & Roberts, H. (2018, July 12). *Understanding Media and Information Quality in an Age of Artificial Intelligence, Automation, Algorithms and Machine Learning*. Retrieved October 18, 2019, from Harvard: Berkman Klein Center: https://cyber.harvard.edu/story/2018-07/understanding-media-and-information-quality-age-artificial-intelligence-automation

Bennett, C. J. (2018). *The European General Data Protection Regulation: An Instrument for the Globalization of Privacy Standards?* Information Polity, 23(2), 239-246.

Bergen, H. (2016). *I'd Blush if I Could': Digital Assistants, Disembodied Cyborgs and The Problem of Gender*. Word and Text, A Journal of Literary Studies and Linguistics, 6, 95-113.

Besaw, C., & Filitiz, J. (2019, January 16). *Artificial Intelligence in Africa is a Double-edged Sword*. Retrieved October 18, 2019, from Our World United Nations University: https://ourworld.unu.edu/en/ai-in-africa-is-a-double-edged-sword

Bezemek, C. (2018, November 2). *The 'Filter Bubble' and Human Rights*. Retrieved October 21, 2019, from Fundamental Rights Protection Online: https://ssrn.com/abstract=3277503

Bostrom, N. (2016). *Strategic Implications of Openness in AI Development*. Oxford, UK: Future of Humanity Institute, Oxford University.

Boyd, d.(2008) *Taken Out of Context American Teen Sociality in Networked Publics*. Retrieved 16 November 2019 at the link: https://www.danah.org/papers/TakenOutOfContext.pdf

Boyd, C. (2017, November 8). *AI scientists: How can companies deal with the shortage of talent?* Retrieved October 16, 2019, from Towards Data Science: https://towardsdatascience.com/ai-scientists-how-can-companies-deal-with-the-shortage-of-talent-11ab48566677

Boyd, D., & Crawford, K. (2012). *Critical Questions for Big Data. Information, Communication and Society: Provocations for a cultural, technological, and scholarly phenomenon*, 15(5), 662-679.

Broadband Commission. (2018). *2018 State of Broadband Report: Broadband Catalyzing Sustainable Development*. Geneva: ITU & UNESCO.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Oxford: Future of Humanity Institute: University of Oxford.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., ... Trench, M. (2017). *Artificial Intelligence: The Next Digital Frontier?* Brussels: Mckinsey Global Institute.

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy*. Brussels: Mckinsey Global Institute.

Buolamwini, J., & Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research*: *Conference on Fairness, Accountability, and Transparency* (pp. 1-15). New York: Proceedings of Machine Learning Research.

Cairns, A. (2019, January 18). *Why AI Is Failing The Next Generation of Women*. Retrieved October 21, 2019, from World Economic Forum: https://www.weforum.org/agenda/2019/01/ai-artificial-intelligence-failing-next-generation-women-bias/

Cannataci, J. A., Zhao, B., Torres Vives, G., Monteleone, S., Bonnici, J. M., & Moyakine, E. (2016). *Privacy, free expression and transparency: redefining their new boundaries in the digital age*. Paris: UNESCO.

Carlson, M. (2014). *The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority*. Digital Journalism, 3(3), 416-431.

Carpenter, J. (2015, July 6). *Google's Algorithm Shows Prestigious Job Ads to Men, but Not to Women. Here's Why That Should Worry You*. Retrieved October 21, 2019, from The Washington Post: https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/

Carroll, S. (2009, March 30). *Why Can't We Visualize More Than Three Dimensions?* Retrieved October 21, 2019, from Discover Magazine: http://blogs.discovermagazine.com/cosmicvariance/2009/03/30/why-cant-we-visualize-more-than-three-dimensions/#.Xa13bEYzaHs

Carter, J. M. (2018, August 1). *LGBQT Individuals Are Less Represented in Science - and That's A Problem*. Retrieved October 21, 2019, from ASBMB Today: https://www.asbmb.org/asbmbtoday/201808/Essay/LGBTQ/

UN Chief Executives Board. CEB/2019/1/Add.3. (2019, June 17). *Summary of deliberations. A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence*. Geneva, Switzerland: United Nations System Chief Executives Board.

Chatterjee, B. B. (2002). *Razorgirls and Cyberdykes: Tracing Cyberfeminism and Thoughts on Its Use in A Legal Context*. International Journal of Sexuality and Gender Studies, 2(3), 197-213.

Chen, L. (2019, July 20). *China's Deepfake Celebrity Porn Culture Stirs Debate About Artificial Intelligence Use*. Retrieved October 21, 2019, from South China Morning Post: https://www.scmp.com/news/china/society/article/3019389/chinas-deepfake-celebrity-porn-culture-stirs-debate-about

Chesney, R., & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. California Law Review, 1-58.

Cisse, M. (2018, October 23). *Look to Africa to advance artificial intelligence*. Retrieved October 18, 2019, from Nature: https://www.nature.com/articles/d41586-018-07104-7

CITP & UCHV Princeton. (2018, October). *Dynamic Sound Identification*. Retrieved October 18, 2019, from AI and Ethics: https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-2.pdf

Citron, D. K. (2017). *Extremist Speech, Compelled Conformity and Censorship Creep*. Notre Dame Law Review, 93(3), 1035.

Citron, D. K., & Jurecic, Q. (2018, September 5). *Platform Justice: Content Moderation at an Inflection Point*. Retrieved October 21, 2019, from Hoover Working Group on National Security, Technology, and Law: https://www.hoover.org/sites/default/files/research/docs/citron-jurecic_webreadypdf.pdf

Cockburn, I. M., Henderson, R., & Stern, S. (2018, March). *The Impact of Artificial Intelligence*. Retrieved October 21, 2019, from National Bureau of Economic Research: https://www.nber.org/papers/w24449.pdf

Council of Europe. (2018). CM/Rec(2018)2. *Recommendation of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries*. Strasbourg: Council of Europe.

Council of Europe. (2018). *Glossary*. Retrieved September 11, 2019, from Council of Europe Artificial Intelligence: https://www.coe.int/en/web/artificial-intelligence/glossary

Council of Europe. CHR/Rec(2019)1. (2019, May). *Recommendation of the Commissioner for Human Rights on Unboxing Artificial Intelligence: 10 steps to protect Human Rights*. Retrieved September 4, 2019, from Council of Europe: https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64

Council of Europe. CM/Rec(2012)3. (2012, April 4). *Recommendation of the Committee of Ministers to Member States on The Protection of Human RIghts with Regard to Search Engines*. Strasbourg, France.

Council of Europe. Decl(13/02/2019)1. (2019, February 13). *Declaration by The Committee of Ministers on The Manipulative Capabilities of Algorithmic Processes*. Strasbourg, France.

Council of Europe. ETS No.108. (1981, January 28). Strasbourg, France.

Cohen, J. E. (2012, November 5). *What Privacy is For*. Retrieved October 21, 2019, from Social Science Research Network: https://ssrn.com/abstract=2175406

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). *Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data*. Journal of Communication, 64(2), 317-332.

Collingridge, D. (1980). *The Social Control of Technology*. London: Frances Pinter (Publishers) Limited .

Conner-Simons, A. (2018, October 4). *Detecting fake news at its source*. Retrieved October 18, 2019, from MIT Computer Science and Artificial Intelligence Lab: https://www.csail.mit.edu/news/detecting-fake-news-its-source

Conservation International. (2018, November). *Data for Sustainable Development in Kenya*. Retrieved October 18, 2019, from Vizzuality: https://www.vizzuality.com/project/vital-signs-kenya/

Cook, J. (2019, June 23). *Here's What It's Like To See Yourself In A Deepfake Porn Video*. Retrieved October 21, 2019, from Huffington Post: https://www.huffpost.com/entry/deepfake-porn-heres-what-its-like-to-see-yourself_n_5d0d0faee4b0a3941861fced?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABJdel2j13KP3PvAkKXhAMZU3BdNPc_NAWhGdP2nQvW1zElHRtO5qCeCklMhCNMbAV

Corea, G., Duelli-Klein, R., Hanmer, J., Homes, H. B., Hoskins, B., Kishwar, M., ... Steinbacher, R. (1985). *Man-Made Women: How New Reproductive Technologies Affect Women*. London: Hutchinson.

Crabtree, J. (2018, June 14). *Google's next A.I. research center will be its first on the African continent*. Retrieved October 18, 2019, from CNBC Africa: https://www.cnbcafrica.com/news/2018/06/16/googles-next-a-i-research-center-will-be-in-africa/

Crawford, K. (2013, May 10). *Think Again: Big Data*. Retrieved October 21, 2019, from Foreign Policy: https://foreignpolicy.com/2013/05/10/think-again-big-data/

Crenshaw, K. (1989). *Demarginalizing The Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*. University of Chicago Legal Forum(1), 139-168.

Crenshaw, K. (1991). *Mapping The Margins: Identity Politics, Intersectionality, and Violence against Women*. Stanford Law Review, 43(6), 1241-1299.

Cuboniks, L. (2014). *Xenofeminism: A Politics for Alienation*. Retrieved October 21, 2019, from XF Manifesto: https://www.laboriacuboniks.net/

Culzac, N. (2014, September 17). *Egypt's Police 'Using Social Media and Apps Like Grindr to Trap Gay People'.* Retrieved October 18, 2019, from Independent: https://www.independent.co.uk/news/world/africa/egypts-police-using-social-media-and-apps-like-grindr-to-trap-gay-people-9738515.html

da Costa, P. C. (2018). *Conversing with Personal Digital Assistants: on Gender and Artificial Intelligence*. Journal of Science and Technology of The Arts, 10(3), 2-59.

Danaher, J., & Sandberg, A. (2017). *Robot Sex: Social and Ethical Implications*. Cambridge, Massachusetts: MIT Press.

Daniels, C., Mawoko, P., & Konte, A. (2018). *Evaluating Public Policies in Africa: insights from the Science, Technology, and Innovation Strategy for Africa 2024 (STISA-2024).* Sussex: University of Sussex .

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias*

against women. Retrieved October 18, 2019, from Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Data Science Nigeria. (2019). *AI Bootcamp 2019 100 Days of Machine Learning & Deep Learning Classes*. Retrieved October 18, 2019, from YouTube: https://www.youtube.com/playlist?list=PL0mGkrTWmp4sT5_CdhbhBkVX7r7WzIGHB

David Soergel, A. S. (2013). *Open Scholarship and Peer Review: a Time for Experimentation*. ICML.

Davies, H. (2018). *Redefining Filter Bubbles as (Escapable) Socio-Technical Recursion*. Sociological Research Online, 23(2), 637-654.

de Zayas, A., & Martín, Á. R. (2012). *Freedom of Opinion and Freedom of Expression: Some Reflections on General Comment No. 34 of The UN Human Rights Committee*. Netherlands International Law Review, 59(3), 425-454.

Deep Learning Indaba. (2019). *Together We Build African Artificial Intelligence*. Cape Town: Deep Learning Indaba.

Delaney, K. (2017, February 21). *Filter Bubbles Are A Serious Problem with News, Says Bill Gates*. Retrieved October 21, 2019, from Quartz: https://qz.com/913114/bill-gates-says-filter-bubbles-are-a-serious-problem-with-news/

Devlin, K. (2015, September 17). *In Defence of Sex Machines: Why Trying to Ban Sex Robots Is Wrong*. Retrieved October 18, 2019, from The Conversation: https://theconversation.com/in-defence-of-sex-machines-why-trying-to-ban-sex-robots-is-wrong-47641

Dialani, P. (2019, January 29). *Different Ways of How Twitter Uses Artificial Intelligence*. Retrieved October 21, 2019, from Analytics Insight: https://www.analyticsinsight.net/different-ways-of-how-twitter-uses-artificial-intelligence/

Dörr, K. N. (2016). *Mapping The Field of Algorithmic Journalism*. Digital Journalism, 4(6), 700-722.

Dörr, K. N., & Hollnbuchner, K. (2017). *Ethical Challenges of Algorithmic Journalism*. Digital Journalism, 5(4), 404-419.

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed Messages?* Washington, D.C.: Center for Democracy and Technology.

Dutton, T. (2018, June 28). *An Overview of National AI Strategies*. Retrieved October 21, 2019, from Medium: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

Dyck, J. (2017, July 2018). *Siri Is Not Genderless*. Retrieved October 21, 2019, from Vice: https://www.vice.com/en_us/article/a3dx34/siri-is-not-genderless

Egyptian Streets. (2019, September 9). *Egypt Organizes First International Conference on Artificial Intelligence*. Retrieved October 18, 2019, from Egyptian Street: https://egyptianstreets.com/2019/09/09/egypt-organizes-first-international-conference-on-artificial-intelligence/

Elsevier. (2018). *Artificial Intelligence: How knowledge is created, transferred, and used*. Amsterdam: Elsevier.

European Commission. (2017). *Building A European Data Economy*. Brussels: European Commission.

European Commission. (2018). *2018 Reform of EU Data Protection Rules*. Retrieved October 18, 2019, from European Commission: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

European Council CO EUR 8 CONCL 3. (2017, June 23). European Council meeting (22 and 23 June 2017) Conclusions. Brussels, Belgium.

European Data Protection Supervisor. (2019). *Annual Report 2018*. Luxembourg: European Union .

Eweniyi, O. (2017, March 10). *Three Guys Created An AI Solution To Nigeria's Public Transportation Issues*. Retrieved October 21, 2019, from Konbini: https://www.konbini.com/ng/lifestyle/roadpreppers-artificial-intelligence-simplify-public-transportation/

Fang, A. (2019, April 2). *Chinese Colleges to Offer AI Major in Challenge to US*. Retrieved October 21, 2019, from Nikkei Asian Review: https://asia.nikkei.com/Business/China-tech/Chinese-colleges-to-offer-AI-major-in-challenge-to-US

Fessler, L. (2017, February 22). *We Tested Bots Like Siri and Alexa to See Who Would Stand Up to Sexual Harassment.* Retrieved October 21, 2019, from Quartz: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/

Fitzsimons, T. (2019, May 9). *Google, Trevor Project WIll Use AI to Combat LGBTQ Youth Suicide*. Retrieved October 21, 2019, from NBC News: https://www.nbcnews.com/feature/nbc-out/google-trevor-project-will-use-ai-combat-lgbtq-youth-suicide-n1003511

Flaxman, S., Goel, S., & Rao, J. M. (2016). *Filter Bubbles, Echo Chambers, and Online News Consumption*. Public Opinion Quarterly, 80(1), 298-320.

Flew, T., Spurgeon, C., & Daniel, A. (2012). *The Promise of Computational Journalism*. Journalism Practice, 6(2), 151-171.

Foucault, M. (1977). *Discipline and Punishment: The Birth of The Prison*. London: Allen Lane.

Freeman, J. (2019, July 5). *How to Fight STEM's Unconscious Bias against LGBTQ People*. Retrieved October 21, 2019, from Scientific America: https://blogs.scientificamerican.com/voices/how-to-fight-stems-unconscious-bias-against-lgbtq-people/

Fu, L. (2018, May 29). *Four Key Barriers to the Widespread Adoption of AI*. Retrieved October 21, 2019, from MIT: Professional Education: https://professional.mit.edu/news/news-listing/four-key-barriers-widespread-adoption-ai

Fussell, S. (2017, October 23). *Palestinian Man Arrested After Facebook Auto-Translates 'Good Morning' as 'Attack Them'* . Retrieved October 18, 2019, from Gizmodo: https://gizmodo.com/palestinian-man-arrested-after-facebook-auto-translates-1819782902

GDPR art. IV, cl. 4. (2016, April 27). Brussels, Belgium.

Gee, T. J. (2017, July 5). *Why Female Sex Robots Are More Dangerous Than You Think.* Retrieved October 21, 2019, from The Telegraph: https://www.telegraph.co.uk/women/life/female-robots-why-this-scarlett-johansson-bot-is-more-dangerous/

Gehrmann, S., Strobelt, H., & Rush, A. (2019, June 10). *GLTR: Statistical Detection and Visualization of Generated Tezt*. Retrieved October 18, 2019, from Cornell University: https://arxiv.org/abs/1906.04043

Gelman, A., Mattson, G., & Simpson, D. (2018). *Gaydar and The Fallacy of Objective Measurement*. Sociological Science, 5(2), 270-280.

Geo Gecko. (n.d.). *What We Do*. Retrieved October 18, 2019, from Geo Gecko: https://www.geogecko.com/uganda-gis-services

George, T. (2018, December 12). *Newsrooms Must Learn How to Use AI: Trust in Journalism Is At Stake*. Retrieved October 18, 2019, from Journalism.co.uk: https://www.journalism.co.uk/skills/lessons-learned-in-the-last-four-years-of-using-artificial-intelligence-at-the-associated-press/s7/a731760/

Gilmore, J. (2011). *Expression as Realization: Speakers' Interests in Freedom of Speech*. Law and Philosophy, 30(5), 517-539.

Glassdoor. (2018). *25 Highest Paying Jobs in America*. Retrieved October 18, 2019, from Glassdoor: https://www.glassdoor.com/List/Highest-Paying-Jobs-LST_KQ0,19.htm

GOBL. (2014). *CSIR*. Retrieved October 18, 2019, from GOBL: https://www.gobl-project.eu/council-for-scientific-and-industrial-research/

Goitom, H. (2019). *Regulation of Artificial Intelligence in Selected Jurisdictions*. Washington DC: The Law Library of Congress.

Goldstein, E., Gasser, U., & Budish, R. (2018, June 21). *Data Commons Version 1.0: A Framework to Build Toward AI for Good*. Retrieved October 18, 2019, from Berkman Klein Center: https://medium.com/berkman-klein-center/data-commons-version-1-0-a-framework-to-build-toward-ai-for-good-73414d7e72be

GoodFellow, I. J., Papernot, N., Huang, S., Duan, Y., Abbeel, P., & Clark, J. (2017, February 24). *Attacking Machine Learning with Adversarial Examples*. Retrieved October 18, 2019, from Open AI: https://openai.com/blog/adversarial-example-research/

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples*. ICLR (pp. 1-11). San Diego: arXiv.

Goodman, B., & Flaxman, S. (2017). *European Union regulations on algorithmic decision-making and "a right to explanation".* AI Magazine, 38(3), 50-57.

Goodrich, J., & McCrea, M. (2018, April 26). *Tech Companies Collaborate to Create Next-Gen Sex Robots*. Retrieved October 21, 2019, from CBS Local: https://sanfrancisco.cbslocal.com/2018/04/26/sex-robot-next-gen-california-tech-company/

Government of Canada. (2018). *Algorithmic Impact Assessment (v0.2).* Retrieved October 18, 2019, from Government of Canada: https://canada-ca.github.io/digital-playbook-guide-numerique/en/overview.html

Greene, T. (2018, August 9). *Uber's 'Real-time ID Check' Doesn't Deal Well with Transgender Drivers*. Retrieved October 18, 2019, from The Next Web: https://thenextweb.com/artificial-intelligence/2018/08/08/ubers-real-time-id-check-doesnt-deal-well-with-transgender-drivers/

Greenhalgh, H. (2018, September 26). *Gay Rights Groups Hail Landmark U.N. Chief Speech Backing LGBT+ Community*. Retrieved October 18, 2019, from Reuters: https://www.reuters.com/article/us-global-un-lgbt-idUSKCN1M62IL

Grint, K., & Gill, R. (1995). *The Gender-Technology Relation: Contemporary Theory and Research*. London: Taylor & Francis.

Gul, E. (2019, July 5). *Is Artificial Intelligence the frontier solution to Global South's wicked development challenges?* Retrieved October 21, 2019, from Medium: https://towardsdatascience.com/is-artificial-intelligence-the-frontier-solution-to-global-souths-wicked-development-challenges-4206221a3c78

Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., & Aggarwal, C. (2019, August 12). *Efficient Data Representation by Selecting Prototypes with Importance Weights*. Retrieved October 21, 2019, from Cornell University: ArXiv: https://arxiv.org/pdf/1707.01212.pdf

Gustavsson, E. (2005). *Virtual Servants: Stereotyping Female Front-Office Employees on The Internet. Gender, Work and Organization*, 12(5), 400-419.

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. (2018, November 28). *Why We Need to Audit Algorithms*. Retrieved October 18, 2019, from Harvard Business Review: https://hbr.org/2018/11/why-we-need-to-audit-algorithms

Halilou, I. (2016, July 16). *10 E-Learning Platforms Transforming Education in Africa*. Retrieved October 21, 2019, from True Africa: https://trueafrica.co/lists/e-learning-platforms-africa-tutor-ng-mest-school-education-startups/

Hansen, E. (2018). *Aporias of Courage and The Freedom of Expression*. Philosophy and Social Criticism, 44(1), 100-117.

Hao, K. (2019, August 7). *China's path to AI domination has a problem: brain drain*. Retrieved October 21, 2019, from MIT Technology Review: https://www.technologyreview.com/f/614092/china-ai-domination-losing-talent-to-us/

Haraway, D. J. (1991). *Science, Technology, and Socialist-Feminism in the Late Twentieth Century*. In D. J. Haraway, Simians, Cyborgs and Women: The Reinvention of

Nature (pp. 149-181). New York: Routledge.

Harcourt, B. E. (2007). *Against Prediction: Punishing and Policing in An Actuarial Age*. Chicago: University of Chicago Press.

Harwell, D. (2019, June 12). *Top AI Researchers Race to Detect 'Deepfake' Videos: 'We Are Outgunned'*. Retrieved October 21, 2019, from The Washington Post: https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/?noredirect=on

Hensel, A. (2018, November 15). *Facebook to Create 'Independent Panel' for Appealing Content Moderation Decisions*. Retrieved October 18, 2019, from Venture Beat: https://venturebeat.com/2018/11/15/facebook-to-create-independent-panel-for-appealing-content-moderation-decisions/

Hern, A. (2017, October 24). *Facebook Translates 'Good Morning' into 'Attack Them', Leading to Arrest*. Retrieved October 18, 2019, from The Guardian: https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Hester, H. (2016, August 8). *Technically Female: Women, Machines, and Hyperemployment*. Retrieved October 21, 2019, from Salvage: http://salvage.zone/in-print/technically-female-women-machines-and-hyperemployment/

Hester, H., & Angel, K. (2016). *Technosexuals*. (S. Radio, Interviewer)

Hill, C., Corbett, C., & St. Rose, A. (2010). *Why So Few? Women in Science, Technology, Engineering and Mathematics*. Washington D.C.: American Association of University Women.

Hope, C., & McCann, K. (2017, September 19). *Google, Facebook and Twitter Told to Take Down Terror Content within Two Hours or Face Fines*. Retrieved October 18, 2019, from The Telegraph: https://www.telegraph.co.uk/news/2017/09/19/google-facebook-twitter-told-take-terror-content-within-two/

Hutson, M. (2018, May 17). *Why are AI researchers boycotting a new Nature journal—and shunning others?* Retrieved October 21, 2019, from Science Mag: https://www.sciencemag.org/news/2018/05/why-are-ai-researchers-boycotting-new-nature-journal-and-shunning-others

IBM. (2019a, August 8). *AI Explainability 360*. Retrieved October 18, 2019, from IBM Developer: https://developer.ibm.com/open/projects/ai-explainability/

IBM. (2019b). *IBM Watson OpenScale*. Retrieved October 18, 2019, from IBM: https://www.ibm.com/cloud/watson-openscale/

ICT Authority Kenya. (2019, September). *Open Data*. Retrieved October 18, 2019, from Kenya Open Data: http://icta.go.ke/open-data/

InstaDeep. (2019). *Instadeep: About*. Retrieved October 18, 2019, from InstaDeep: https://www.instadeep.com/about/

IPDC Council CI/2018/COUNCIL.31/H/1. (2018, November 21-22). *Decisions taken by*

the 31st Council Session of the International Programme for the Development of Communication (IPDC). Paris, France.

Ireton, C., & Posetti, J. (2018f). *Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training*. Paris: UNESCO Publishing.

IREX. (2018). *Can Machine Learning Help Us Measure The Trustworthiness of News*. Washington D.C.: IREX.

ITU. (2012, June). *Next Generation Networks – Frameworks and Functional Architecture Models.* ITU-T Y.2060 Standard Series Y: Global Information Infrastructure, Internet Protocol Aspects and Next-Generation Networks. International Telecommunication Union.

ITU. (2014, August). *Information Technology – Cloud Computing – Overview and Vocabulary.* ITU-T Y.3500 Standard Series Y: GLobal Information Infrastructure, Internet Protocol Aspects and Next-Generation Networks. International Telecommunication Union.

ITU. (2017). *ICT Facts and Figures 2017*. Geneva: ITU.

Jao, N. (2017, May 12). *Big Data to tackle grand challenges: A look at IBM Research Africa's projects*. Retrieved October 21, 2019, from ITUNews: https://news.itu.int/big-data-to-tackle-grand-challenges-a-look-at-ibm-research-africas-projects/

Just, N., & Latzer, M. (2017). *Governance by Algorithms: Reality construction by algorithmic selection on the Internet. Media, Culture and Society*, 39(2), 238-258.

Kantrowitz, A. (2018, March 21). *Facebook Has Blocked Ad Targeting By Sexual Orientation.* Retrieved October 21, 2019, from BuzzFeed News: https://www.buzzfeednews.com/article/alexkantrowitz/facebook-has-blocked-ad-targeting-by-sexual-orientation

KAS Uganda & S.Sudan. (2018, December 12). *Digitalisation Forum 2018 — Managing the 4th Industrial Revolution*. Retrieved October 18, 2019, from Medium: https://medium.com/@KasUganda/digitalisation-forum-2018-managing-the-4th-industrial-revolution-4d1cc866a736

Keller, D. (2018, June 13). *Internet Platforms: Observations on Speech, Danger, and Money.* Retrieved October 21, 2019, from Hoover Institution's Aegis Paper Series: https://ssrn.com/abstract=3262936

Kelly, A. (1985). *The Construction of Masculine Science*. British Journal of Sociology of Education, 6(2), 133-154.

Kenyan Wall Street. (2018, February 28). *Kenya Govt unveils 11 Member Blockchain & AI Taskforce headed by Bitange Ndemo*. Retrieved October 18, 2019, from The Kenyan Wall Street: https://kenyanwallstreet.com/kenya-govt-unveils-11-member-blockchain-ai-taskforce-headed-by-bitange-ndemo/

Keyes, O. (2018). *The Misgendering Machines: Trans/HCI Implication of Automatic Gender Recognition. ACM on Human-Computer Interaction*. New York: Association for Com-

puting Machinery.

Kharpal, A. (2019, June 4). *China is ramping up its own chip industry amid a brewing tech war. That could hurt US firms*. Retrieved from CNBC: https://www.cnbc.com/2019/06/04/china-ramps-up-own-semiconductor-industry-amid-the-trade-war.html

Kleeman, J. (2017, September 25). *Should we ban sex robots while we have the chance?* Retrieved October 18, 2019, from The Guardian: https://www.theguardian.com/commentisfree/2017/sep/25/ban-sex-robots-dolls-market

Kline, R. R. (2015). *Technological Determinism.* In J. D. Write, International Encyclopedia of the Social & Behavioral Sciences (Second Edition) (pp. 109-112). New York: Elsevier.

Knight, W. (2018, November 17). *One of the fathers of AI is worried about its future.* Retrieved October 18, 2019, from MIT Technology Review: https://www.technologyreview.com/s/612434/one-of-the-fathers-of-ai-is-worried-about-its-future/

Kudi. (2019). *About*. Retrieved from Kudi: https://kudi.com/

Kumar, V., Raghavendra, R., Namboodiri, A., & Busch, C. (2016). *Robust Transgender Face Recognition: Approach Based on Appearance and Therapy Factors*. 2016 IEEE International Conference on Identity, Security and Behavior Analysis (pp. 1-7). Sendai: IEEE.

Kunze, L. (2019). *Can We Stop the Academic AI Brain Drain?* Künstliche Intelligenz, 33(1), 1-3.

Kwok, R. (2019, April 29). *Junior AI researchers are in demand by universities and industry.* Retrieved October 18, 2019, from Nature: https://www.nature.com/articles/d41586-019-01248-w

Lai, C. K., & Banaji, M. R. (2019). *The Psychology of Implicit Intergroup Bias and the Prospect of Change*. In D. Allen, & R. Somanathan, Difference without Domination: Pursuing Justice in Diverse Democracies. Chicago, Illinois: University of Chicago Press.

Lamagna, M. (2018, March 18). *Google's Parent Company is Using AI to Make the Internet Safer for LGBT People*. Retrieved October 21, 2019, from Market Watch: https://www.marketwatch.com/story/how-artificial-intelligence-can-make-the-internet-better-and-safer-for-lgbt-people-2018-03-14

Latar, N. L. (2015). *The Robot Journalist in The Age of Social Phsics: The End of Human Journalism?* In G. Einav, The New World of Transitioned Media: Digital Realignment and Industry Transformation (pp. 65-80). New York: Springer.

Latonero, M. (2018). *Governing Artificial Intelligence: Upholding Human Rights & Dignity*. New York: Data & Society.

Le Monde. (2019, May 1). *Agriculture Numérique: Le Sénégal Montre L'exemple*. Retrieved October 18, 2019, from Le Monde Afrique: https://www.lemonde.fr/

afrique/article/2019/05/01/agriculture-numerique-le-senegal-montre-l-exemple_5457016_3212.html

Leavy, S. (2018). *Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory*. 1st International Workshop on Gender Equality in Software Engineering (pp. 14-16). Gothenburg: ACM.

LeCun, Y. (2009). *A New Publishing Model in Computer Science.* Retrieved October 21, 2019, from Yann LeCun Blog: http://yann.lecun.com/ex/pamphlets/publishing-models.html

Lee, D. (2018, February 3). *Deepfakes Porn Has Serious Consequences*. Retrieved October 21, 2019, from BBC News: https://www.bbc.com/news/technology-42912529

Levchuk, K. (2018, August 1). *Why Women Should Be Excited About AI*. Retrieved October 21, 2019, from Forbes: https://www.forbes.com/sites/cognitiveworld/2018/08/01/why-women-should-be-excited-about-ai/#56c12691272b

Levin, S. (2017, September 9). *LGBT Groups Denounce 'Dangerous' AI that Uses Your Face to Guess Sexuality*. Retrieved October 21, 2019, from The Guardian: https://www.forbes.com/sites/cognitiveworld/2018/08/01/why-women-should-be-excited-about-ai/#24f0aef21272

Lewton, T., & McCool, A. (2018, December 14). *This App Tells Your Doctor If You Have Malaria*. Retrieved October 21, 2019, from CNN Health: https://edition.cnn.com/2018/12/14/health/ugandas-first-ai-lab-develops-malaria-detection-app-intl/index.html

Liptak, A. (2017, May 1). *Sent to Prison by a Software Program's Secret Algorithms.* Retrieved October 21, 2019, from The New York Times: https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html?smid=tw-share&_r=0

Lomas, N. (2017, April 29). *Someone scraped 40,000 Tinder selfies to make a facial dataset for AI experiments*. Retrieved October 21, 2019, from Tech Crunch: https://techcrunch.com/2017/04/28/someone-scraped-40000-tinder-selfies-to-make-a-facial-dataset-for-ai-experiments/?guccounter=1&guce_referrer_us=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_cs=5Np2afFglc8vqTYzWUA9Hw

Lystad, J. (2019, July 12). *Artificial Intelligence School to Open in Fez*. Retrieved October 21, 2019, from Morocco World News: https://www.moroccoworldnews.com/2019/07/278026/artificial-intelligence-school-open-fez/

Lyu, S. (2018, July 12). *Detecting Deepfake Videos*. (D. Hopper, Interviewer)

Machalo, G. (2019, February 20). *Cabinet Approves SMART Zambia eGovernment Master Plan and Public Services Standards*. Retrieved October 21, 2019, from SMART Zambia Institute: https://www.szi.gov.zm/cabinet-approves-smart-zambia-egovernment-master-plan-and-public-services-standards/+Government+-Bill%2C+2018)+7th+National+Development+Plan+ackn

MacLellan, L. (2019, March 22). *Hear What A Genderless AI Voice Sounds Like-and*

Consider Why It Matters. Retrieved October 21, 2019, from Quartz: https://qz.com/work/1577597/this-ai-voice-is-gender-neutral-unlike-siri-and-alexa/

Madden, M., & Raine, L. (2015, May 20). *Americans' Attitudes About Privacy, Security and Surveillance*. Retrieved September 4, 2019, from Pew Research Center: https://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/

Maréchal, D. N. (2018, November 16). *Targeted Advertising Is Ruining the Internet and Breaking the World*. Retrieved September 5, 2019, from Vice: https://www.vice.com/en_us/article/xwjden/targeted-advertising-is-ruining-the-internet-and-breaking-the-world

Marr, B. (2018, April 16). *The 6 Best Free Online Artificial Intelligence Courses Available Today*. Retrieved October 21, 2019, from Forbes: https://www.forbes.com/sites/bernardmarr/2018/04/16/the-6-best-free-online-artificial-intelligence-courses-for-2018/#279efe8b59d7

Masure, A., & Pandelakis, P. (2017, October). *Machines Désirantes : des Sexbots aux OS Amoureux*. Retrieved October 21, 2019, from ReS Futurae: https://journals.openedition.org/resf/1066

Matsakis, L. (2018, July 10). *A Frightening AI Can Determine Whether a Person Is Gay With 91 Percent Accuracy.* Retrieved October 21, 2019, from Vice: https://www.vice.com/en_asia/article/a33xb4/a-frightening-ai-can-determine-a-persons-sexuality-with-91-accuracy

Mbaka, C. (2017, December 17). *Sophie Bot Aims to Destigmatize Sex Education in Africa*. Retrieved October 21, 2019, from Techmoran: https://techmoran.com/2017/12/17/sophie-bot-aims-to-reduce-stigma-around-sexual-education/

McMullan, T. (2015, October 18). *Guardian readers on privacy: 'we trust government over coporations'.* Retrieved September 4, 2019, from The Guardian: https://www.theguardian.com/technology/2015/oct/18/guardian-readers-on-privacy-we-trust-government-over-corporations

Mehta, I. (2018, October 15). *Amazon's New Patent Will Allow Alexa to Detect A Cough or A Cold*. Retrieved October 21, 2019, from The Next Web: https://thenextweb.com/artificial-intelligence/2018/10/15/amazons-new-patent-will-allow-alexa-to-detect-your-illness/

Metz, C. (2017, November 5). *Building A.I. That Can Build A.I.* Retrieved October 21, 2019, from The New York Times: https://www.nytimes.com/2017/11/05/technology/machine-learning-artificial-intelligence-ai.html

Metz, C. (2019, August 16). *A.I. Is Learning From Humans. Many Humans*. Retrieved from The New York Times: https://www.nytimes.com/2019/08/16/technology/ai-humans.html

Michael, V., Van Kleek, M., & Binns, R. (2018). *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-making*. CHI Conference on Human Factors in Computing Systems (pp. 440-454). Montreal: ACM.

Mikkinen, M., Auffermann, B., & Heinonen, S. (2017). *Framing The Future of Privacy: Citizens' Metaphors for Privacy in The Coming Digital Society*. European Journal of Futures Research, 5(7), 7.

Miller, H., & Stirling, R. (2019). *Government Artificial Intelligence Readiness Index 2019*. Ottawa: IDRC; Oxford Insights.

Ministry of Information, Communications and Technology. (2019). *Emerging Digital Technologies for Kenya: Exploration and Analysis*. July: Government of Kenya.

Ministry of Transport and Communications (MTC). (2011). *Departments/Units*. Retrieved October 18, 2019, from Ministry of Transport and Communications (MTC): http://www.gov.bw/en/Ministries--Authorities/Ministries/Ministry-of-Transport-and-Communications/Departments/Telecommunications/

Montal, T., & Reich, Z. (2017). *I, robot. You, journalist. Who Is The Author.* Digital Journalism, 5(7), 829-849.

Montjoye, Y.-A., Farzanehfar, A., Hendrickx, J., & Rocher, L. (2017). *Solving Artificial Intelligence's Privacy Problem*. The Journal of Field Actions, 17, 80-83.

Morris, A. (2018, September 25). *Prediction: Sex Robots Are The Most Disruptive Technology We Didn't See Coming*. Retrieved October 21, 2019, from Forbes: https://www.forbes.com/sites/andreamorris/2018/09/25/prediction-sex-robots-are-the-most-disruptive-technology-we-didnt-see-coming/#1961498d6a56

Moses, L. (2017, September 14). *The Washington Post's Robot Reporter Has Published 850 Articles in The Past Year*. Retrieved October 21, 2019, from Digiday UK: https://digiday.com/media/washington-posts-robot-reporter-published-500-articles-last-year/

Moye, D. (2017, September 29). *Sex Robot Molested At Electronics Festival, Creators Say.* Retrieved October 21, 2019, from HuffPost: https://www.huffpost.com/entry/samantha-sex-robot-molested_n_59cec9f9e4b06791bb10a268

Moye, D. (2018, May 24). *Amazon Admits Alexa Device Eavesdropped On Portland Family*. Retrieved October 21, 2019, from HuffPost: https://www.huffpost.com/entry/alexa-eavesdropping-portland-familiy_n_5b0727cae4b0fdb2aa51b23e

MSI-NET. (2016). *Study on The Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications*.

MSI-NET. (2018). *Algorithms and Human Rights Study on The Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications*. Strasbourg: Council of Europe.

Mukherjee, S. (2017, April 3). *A.I. Versus M.D.* Retrieved October 21, 2019, from The New Yorker: https://www.newyorker.com/magazine/2017/04/03/ai-versus-md

Murphy, H. (2017, October 9). *Why Stanford Researchers Tried to Create A 'Gaydar' Machine*. Retrieved October 21, 2019, from The New York Times: https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html

Namibia University of Science and Technology. (2019). *Bachelor of Computer Science.* Retrieved October 18, 2019, from Namibia University of Science and Technology: https://www.nust.na/?q=programme/fci/bachelor-computer-science

NAMPA. (2019, March 7). *Namibia Should Play Active Role in Fourth Industrial Revolution: Tweya*. Retrieved October 18, 2019, from Lela Mobile: https://www.lelamobile.com/content/79171/Namibia-should-play-active-role-in-fourth-industrial-revolution-Tweya/

Nayyar, G. (2019, July 16). *What do automation and artificial intelligence mean for Africa?* Retrieved October 17, 2019, from Brookings: https://www.brookings.edu/blog/future-development/2019/07/16/what-do-automation-and-artificial-intelligence-mean-for-africa/

Negnevitsky, M. (2011). *Artificial Intelligence: A Guide to Intelligent Systems*. Boston: Addison Wesley.

Nenadic, I. (2018, March 8). *Data-driven Online Political Microtargeting: Hunting for Voters, Shooting Democracy?* Retrieved October 21, 2019, from Center for Media Pluralism and Media Freedom: http://cmpf.eui.eu/data-driven-online-political-microtargeting-hunting-for-voters-shooting-democracy/

Newton, C. (2019, June 28). *Facebook's Supreme Court for Content Moderation is Coming into Focus.* Retrieved October 21, 2019, from The Verge: https://www.theverge.com/interface/2019/6/28/18761357/facebook-independent-oversight-board-report-zuckerberg

Nilsson, K. (2019, March 8). *Why AI Needs More Women*. Retrieved October 21, 2019, from Forbes: https://www.forbes.com/sites/kimnilsson/2019/03/08/why-ai-needs-more-women/#6fc45cab7f90

Nissenbaum, H. (2004). *Privacy as Contextual Integrity*. Washington Law Review, 79(1), 101-139.

NumberBoost. (2019). *NumberBoost*. Retrieved October 18, 2019, from NumberBoost: https://www.numberboost.com/

Nwafor. (2019, April 15). *Nigeria is not ripe for Artificial Intelligence*. Retrieved October 21, 2019, from Vanguard: https://www.vanguardngr.com/2019/04/nigeria-is-not-ripe-for-artificial-intelligence/

O'Dwyer, R. (2018, May 18). *Algorithms are making the same mistakes assessing credit scores that humans did a century ago*. Retrieved October 21, 2019, from QZ: https://qz.com/1276781/algorithms-are-making-the-same-mistakes-assessing-credit-scores-that-humans-did-a-century-ago/

OECD. (2001). *Understanding the Digital Divide*. Paris: OECD.

Oghia, M. J. (2018, November). *Human Rights in AI: A Journalism & Media Perspective*. Paris, France. Retrieved September 2, 2019, from https://www.youtube.com/watch?v=77LNQq9s3tU

Oleksy, W., Just, E., & Zapedowska-Kling, K. (2012). *Gender Issues in Information and*

Communication Technologies. Journal of Information, Communication and Ethics in Society, 10(2), 107-120.

O'Malley, J. (2018, January 12). *Captcha if you can: how you've been training AI for years without realising it*. Retrieved from Techradar: https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin.

Orr, D. (2016, June 10). *At Last, A Cure for Feminism: Sex Robots*. Retrieved October 21, 2019, from The Guardian: https://www.theguardian.com/commentisfree/2016/jun/10/feminism-sex-robots-women-technology-objectify

Owino, J. (2019, April 18). *Agriculture Takes Lead in Adopting AI in Kenya*. Retrieved October 21, 2019, from Capital Business: https://www.capitalfm.co.ke/business/2019/04/agriculture-takes-lead-in-adopting-ai-in-kenya/

Packin, N. G., & Lev Aretz, Y. (2018). *Learning Algorithms and Discrimination*. In W. Barfield, & U. Pagallo, Research Handbook of Artificial Intelligence and Law (pp. 88-113). Cheltenham: Edward Edgar Publishing.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). *Practical Black-Box Attacks against Machine Learning*. ACM Asia Conference on Computer and Communications Security (pp. 1-14). Abu Dhabi: Cornell University: arXiv.

Pariser, E. (2011). *The filter bubble: what the Internet is hiding from you*. New York: Penguin Press.

Parnell, T., & Dünner, C. (2018, March 20). *IBM Sets Tera-scale Machine Learning Benchmark Record with POWER9 and NVIDIA GPUs; Available Soon in PowerAI.* Retrieved October 21, 2019, from IBM Research Blog: https://www.ibm.com/blogs/research/2018/03/machine-learning-benchmark/

Parsheera, S. (2018). *A Gendered Perspective on Artificial Intelligence. ITU Kaleidoscope 2018 - Machine Learning for a 5G Future* (pp. 1-7). Santa Fe: ITU.

Patricia, S., & Steinberg, D. L. (1987). *Made to Order: The Myth of Reproductive and Genetic Progress*. New York: Pergamon Press.

Paul, K. (2019, April 17). *Disastrous lack of diversity in AI industry perpetuates bias, study finds*. Retrieved October 21, 2019, from The Guardian: https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study

Phakathi, B. (2019, April 9). *Cyril Ramaphosa to chair fourth industrial revolution commission*. Retrieved October 21, 2019, from Business Day: https://www.businesslive.co.za/bd/national/2019-04-09-cyril-ramaphosa-to-chair-fourth-industrial-revolution-commission/

Posetti, J. (2017). *Protecting journalism sources in the digital age*. Paris: UNESCO.

Public Private Partnership Commission. (2017). *Digital Malawi Project (DMP) Resettlement Policy Framework*. Blantyre: February.

Quach, K. (2019, March 5). *The Infamour AI Gaydar Study Was Repeated - and, No, Code Can't Tell If You're Straight or Not Just From Your Face.* Retrieved October 21, 2019, from The Register: https://www.theregister.co.uk/2019/03/05/ai_gaydar/

Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial Intelligence & Human Rights: Opportunities & Risks*. Cambridge: The Berkman Klein Center for Internet & Society.

Rense, S. (2018, February 12). *What Are 'Deepfakes,' and Why Are Pornhub and Reddit Banning Them?* Retrieved October 21, 2019, from Esquire: https://www.esquire.com/lifestyle/sex/a17043863/what-are-deepfakes-celebrity-porn/

Reporters Without Borders. (2018). *Online Harassment of Journalists*. Paris: Reporters Without Borders.

Richardson, K. (2016). *Sex Robot Matters: Slavery, the prostituted and the Rights of Machines!* IEEE Technology and Society, 35(2), 46-53.

Rieger, S. (2018, July 26). *At Least Two Malls Are Using Facial Recognition Technology to Track Shoppers' Ages and Genders WIthout Telling*. Retrieved October 21, 2019, from CBC: https://www.cbc.ca/news/canada/calgary/calgary-malls-1.4760964

Robitzski, D. (2018, May 3). *AI Researchers Are Boycotting A New Journal Because It's Not Open Access*. Retrieved October 21, 2019, from Futurism: https://futurism.com/artificial-intelligence-journal-boycot-open-access

robotcampaign. (2019, February 6). *Why We Must Discuss The Normalisation of Sex Dolls and Sex Robots in Our Society!* Retrieved October 18, 2019, from Campaign against Sex Robots: https://campaignagainstsexrobots.org/

Rouvroy, A. (2014). *Data Without (Any)Body? Algorithmic Governmentality as Hyper-disadjointment and The Role of Law as Technical Organ*. Conference on General Organology (pp. 1-2). Canterbury: University of Kent.

Rouvroy, A. (2016). *Of Data and Men Fundamental Rights and Freedoms in A World of Big Data*. Strasbourg: Council of Europe.

Rouvroy, A., & Stiegler, B. (2016). *The Digital Regime of Truth: From The Algorithmic Governmentality to A New Rule of Law*. La Deleuziana: Online Journal of Philosophy, 3(3), 6-29.

Sachs, J. D. (2018). R&D, *Structural Transformation, and the Distribution*. In A. Agrawal, J. Gans, & A. Goldfarb, The Economics of Artificial Intelligence (pp. 329-348). Chicago: University of Chicago Press.

Salmon, F., & Stokes, J. (2010, December 27). *Algorithms Take Control of Wall Street*. Retrieved October 21, 2019, from Wired: https://www.wired.com/2010/12/ff-ai-flashtrading/

Sample, I. (2017a, November 1). *We can't compete: why universities are losing their*

best AI scientists. Retrieved October 21, 2019, from The Guardian: https://www.theguardian.com/science/2017/nov/01/cant-compete-universities-losing-best-ai-scientists

Sample, I. (2017b, November 2). *Big tech firms' AI hiring frenzy leads to brain drain at UK universities*. Retrieved 16 October, 2019, from The Guardian: https://www.theguardian.com/science/2017/nov/02/big-tech-firms-google-ai-hiring-frenzy-brain-drain-uk-universities

Sandelson, J. (2018, March 19). *The politics of AI and scientific research on sexuality.* Retrieved October 21, 2019, from LSE Engenderings: https://blogs.lse.ac.uk/gender/2018/03/19/the-politics-of-ai-and-scientific-research-on-sexuality/

Santa Clara Principles. (2018, May 7). *The Santa Clara Principles.* Content Moderation at Scale Conference. Washington, D.C. Retrieved from The Santa Clara Principles on Transparency and Accountability in Content Moderation : https://santaclaraprinciples.org/

Scanlon, T. (1972). *A Theory of Freedom of Expression.* Philosophy and Public Affairs, 1(2), 204-226.

Schwartz, O. (2018, November 12). *You thought fake news was bad? Deep fakes are where truth goes to die*. Retrieved October 21, 2019, from The Guardian: https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

Scott, M., & Isaac, M. (2016, September 9). *Facebook Restores Iconic Vietnam War Photo It Censored for Nudity*. Retrieved October 21, 2019, from The New York Times: https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html

Scripa Els, A. (2017). *Artificial Intelligence as A Digital Privacy Protector*. Harvard Journal of Law & Technology, 31(1), 217-234.

Shah, S. (2017, February 1). *Facebook Censors 'Explicitly Sexual' Nude Statue of Neptune*. Retrieved October 21, 2019, from Digital Trends: https://www.digitaltrends.com/social-media/facebook-nude-statue/

Sharpe, A., & Raj, S. (2017, September 15). *Using AI to determine queer sexuality is misconceived and dangerous.* Retrieved October 21, 2019, from The Conversation: http://theconversation.com/using-ai-to-determine-queer-sexuality-is-misconceived-and-dangerous-83931

Shead, S. (2017, December 7). *The UK government is giving AI startups access to serious computation power at a new 'Machine Intelligence Garage'.* Retrieved October 21, 2019, from Business Insider: https://www.businessinsider.fr/us/uk-government-gives-ai-startups-access-to-computation-power-2017-12

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., ... Bauer, Z. (2018). *The AI Index 2018 Annual Report.* Stanford: AI Index Steering Committee, Human-Centered AI Initiative, Stanford University.

Sibal, P. (2016, November 11). *Why Trump's Election Is A Failure Of Liberalism*. Retrieved

October 21, 2019, from Huffpost: https://www.huffingtonpost.in/prateek-sibal/why-trumps-election-is-a-failure-of-liberalism_a_21603465/

Simonite, T. (2017, August 21). *Machines Taught by Photos Learn A Sexist View of Women*. Retrieved October 21, 2019, from Wired: https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

Simonite, T. (2018a, January 11). *When It Comes to Gorillas, Google Photos Remains Blind*. Retrieved October 16, 2019, from Wired: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Simonite, T. (2018b, August 17). *AI Is the Future—But Where Are the Women?* Retrieved October 21, 2019, from WIRED: https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/

Snow, J. (2018a, February 14). *'We're in a diversity crisis': cofounder of Black in AI on what's poisoning algorithms in our lives.* Retrieved October 21, 2019, from MIT Technology Review: https://www.technologyreview.com/s/610192/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/

Snow, J. (2018b, February 26). *Bias already exists in search engine results, and it's only going to get worse*. Retrieved October 18, 2019, from MIT Technology Review: https://www.technologyreview.com/s/610275/meet-the-woman-who-searches-out-search-engines-bias-against-women-and-minorities/

Solon, O. (2017, October 4). *More than 70% of US fears robots taking over our lives, survey finds*. Retrieved October 21, 2019, from The Guardian: https://www.theguardian.com/technology/2017/oct/04/robots-artificial-intelligence-machines-us-survey

Solon, O., & Levin, S. (2016, December 16). *How Google's search algorithm spreads false information with a rightwing bias*. Retrieved October 21, 2019, from The Guardian: https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda

Stanford, S. (2018, August 29). *Artificial Intelligence: Salaries Heading Skyward*. Retrieved October 21, 2019, from Medium: https://medium.com/towards-artificial-intelligence/artificial-intelligence-salaries-heading-skyward-e41b2a7bba7d

Stathoulopoulos, K., & Mateos-Garcia, J. (2019). *Gender Diversity in AI Research*. London: Nesta.

Steel, E., & Angwin, J. (2010, August 4). *On the Web's Cutting Edge Anonymity in Name Only*. Retrieved October 17, 2019, from The Wall Street Journal: https://www.wsj.com/articles/SB10001424052748703294904575385532109190198

Stokel-Walker, C. (2019, August 24). *Facebook's Ad Data May Put Millions of Gay People At Risk*. Retrieved October 21, 2019, from NewScientist: https://www.newscientist.com/article/2214309-facebooks-ad-data-may-put-millions-of-gay-people-at-risk/

Stray, J. (2018, December). *More Algorithmic Accountability Reporting, and a lot of it Will*

be Meh. Retrieved October 21, 2019, from Nieman Lab: https://www.niemanlab.org/2018/12/more-algorithmic-accountability-reporting-and-a-lot-of-it-will-be-meh/

Tala. (2019). *About*. Retrieved October 18, 2019, from Tala: https://tala.co/about/

Tech2News. (2019, January 2). *Researchers Create AI Worker so Real it even Cheats when Given Tough Tasks*. Retrieved October 18, 2019, from First Post: https://www.firstpost.com/tech/science/researchers-create-ai-worker-so-real-it-even-cheats-when-given-tough-tasks-5822551.html

TensorFlow Lagos. (2019). *TensorFlow Lagos*. Retrieved October 18, 2019, from TensorFlow Lagos: https://twitter.com/TensorflowLagos

The Economist. (2017, December 7). *Google leads in the race to dominate artificial intelligence*. Retrieved October 16, 2019, from The Economist: https://www.economist.com/business/2017/12/07/google-leads-in-the-race-to-dominate-artificial-intelligence

The Guardian. (2019, May 6). *Nigerian develops AI platform that can translate over 2000 African languages*. Retrieved October 17, 2019, from The Guardian: https://guardian.ng/news/nigerian-develops-ai-platform-that-can-translate-over-2000-african-languages/

The New York Stem Cell Foundation. (2019). *Institutional Report Card for Gender Equality*. Retrieved October 18, 2019, from The New York Stem Cell Foundation: https://nyscf.org/wp-content/uploads/2017/06/2017_IWISE-Report-Card-1.pdf

The Observatory of Economic Complexity. (2019). *Semiconductor Devices*. Retrieved October 16, 2019, from The Observatory of Economic Complexity: https://oec.world/en/profile/hs92/8541/

The Presidency of the Republic of Ghana Communications Bureau. (2019, January 21). *Latest News*. Retrieved October 18, 2018, from The Presidency of the Republic of Ghana: https://www.presidency.gov.gh/index.php/briefing-room/news-style-2/1059-president-akufo-addo-outlines-gov-ts-pillars-of-growth-for-science-and-technology

The Royal Society. (2018). *The Impact of Artificial Intelligence on Work*. London: The Royal Society.

The University of Sheffield. (2018, December 20). *Threats to journalism explored during inaugural UNESCO lecture*. Retrieved October 18, 2019, from The University of Sheffield News: https://www.sheffield.ac.uk/news/nr/unesco-chair-lecture-journalism-safety-impunity-1.822760

Thomas, R. (2015, July 27). *If You Think Women in Tech Is Just A Pipeline Problem, You Haven't Been Paying Attention*. Retrieved October 21, 2019, from Medium: https://medium.com/tech-diversity-files/if-you-think-women-in-tech-is-just-a-pipeline-problem-you-haven-t-been-paying-attention-cb7a2073b996

Thomas, R. (2016, October 4). *The Real Reason Women Quit Tech (and How to Address It)*. Retrieved October 21, 2019, from Medium: https://medium.com/tech-di-

versity-files/the-real-reason-women-quit-tech-and-how-to-address-it-6dfb606929fd

Tonetti, C. J. (2018). *Nonrivalry and the Economics of Data*. Society for Economic Dynamics (pp. 477-489). Mexico City: Society for Economic Dynamics.

Tschan, I., & Bekkoenova, A. (2018, November 22). *What does Artificial Intelligence mean for the future of Human Rights?* Retrieved October 21, 2019, from UNDP Europe and Central Asia: https://www.eurasia.undp.org/content/rbec/en/home/blog/2018/what-does-artificial-intelligence-mean-for-the-future-of-human-r.html

Twitter Help Center. (2019, March). *Terrorism and violent extremism policy*. Retrieved September 4, 2019, from Twitter Help Center: https://help.twitter.com/en/rules-and-policies/violent-groups

UN OHCHR (2018a, 24 May). *UN experts call on India to protect journalist Rana Ayyub from online hate campaign*. Retrieved September 4, 2019, from United Nations Human Rights Office of the High Commissioner: https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23126&LangID=E

UN OHCHR. (2018b, September 3). *UN human rights expert says Facebook's 'terrorism' definition is too broad.* Retrieved September 4, 2019, from United Nations Human Rights Office of the High Commissioner: https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23494&LangID=E

UN Secretary-General's High Level Panel on Digital Cooperation. (2019). *The Age of Digital Interdependence*. New York: The UN Secretary-General's High Level Panel on Digital Cooperation.

UNESCO. (2005). *Towards Knowledge Societies*. Paris: UNESCO Publishing.

UNESCO. (2012). *Policy Guidelines for the Development and Promotion of Open Access*. Paris: UNESCO.

UNESCO. (2014). *Priority Africa at UNESCO: An Operational Strategy for Its Implementation 2014-2021*. Paris: UNESCO.

UNESCO. (2015a). *Keystones to Foster Inclusive Knowledge Societies*. Paris: UNESCO Publishing.

UNESCO. (2015b). *Knowledge societies: The way forward to build a better world*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/70years/knowledge_societies_way_forward_better_world

UNESCO. (2015c). *UNESCO Science Report: Towards 2030*. Paris: UNESCO.

UNESCO. (2016, May 17). *Report Shows Homophobic and Transphobic Violence in Education to be A Global Problem*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/news/report-shows-homophobic-and-transphobic-violence-education-be-global-problem

UNESCO. (2017a). *Cracking the code: Girls' and women's education in science, techno-*

logy, engineering and mathematics (STEM). Paris: UNESCO.

UNESCO. (2017b). *Protecting Journalism Sources in the Digital Age*. Paris: UNESCO.

UNESCO. (2018a). *39 C/5 Approved Program and Budget 2018-2019*. Retrieved October 18, 2019, from UNESCO: https://unesdoc.unesco.org/ark:/48223/pf0000261648

UNESCO. (2018b). *Forum on Artificial Intelligence in Africa*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/artificial-intelligence/africa-forum

UNESCO. (2018c, July-September). *A Lexicon for Artificial Intelligence*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/courier/2018-3/lexicon-artificial-intelligence

UNESCO. (2018d, November 16). *UNESCO hosts a workshop on Artificial Intelligence for Human Rights and SDGs at 2018 Internet Governance Forum*. Retrieved October 18, 2019, from UNESCO News: https://en.unesco.org/news/unesco-hosts-workshop-artificial-intelligence-human-rights-and-sdgs-2018-internet-governance

UNESCO. (2019a). *UNESCO's Internet Universality Indicators: A Framework for Assessing Internet Development*. Paris: UNESCO.

UNESCO. (2019b). *World Press Freedom Day*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/commemorations/worldpressfreedomday

UNESCO. (2019c, May 19). 206 EX/42. *Decisions Adopted by The Executive Board at Its 206th Session*. Paris, France.

UNESCO. (2019d, October 9 - 23). 207 EX/SR.6. *Decisions Adopted by The Executive Board at Its 207th Session*. Paris, France.

UNESCO. (2019e). *UNESCO's Internet Universality ROAM-X Indicators*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/themes/internet-universality-indicators

UNESCO. (2019f, September 28). *International Day for Universal Access to Information*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/commemorations/accesstoinformationday

UNESCO. (2019g, September 9). *UNESCO Engages Technology and Policy Experts for Human Centered AI in Africa*. Retrieved October 18, 2019, from UNESCO: https://en.unesco.org/news/unesco-engages-technology-and-policy-experts-human-centered-ai-africa

UNESCO. (2019h). *In Focus Series of World Trend in Freedom of Expression and Media Development*. Retrieved November 15, 2019, from UNESCO: https://en.unesco.org/world-media-trends

UNESCO Const. art. I, § 2, cl. a. (1945, November 16). London, United Kingdom.

UNESCO Institute of Statistics. (2019). *Information and communication technologies (ICT).* Retrieved October 18, 2019, from UNESCO Institute of Statistics: http://

uis.unesco.org/en/glossary-term/information-and-communication-technologies-ict

UNESCO; EQUALS Skills Coalition. (2019). *I'd blush if I could: Closing gender divides in digital skills through education.* Paris: EQUALS.

UNFE. (2013). *Definitions*. Retrieved November 4, 2019, from United Nations: Free and Equal: https://www.unfe.org/definitions/

UNGA. (2015). *Information and communications technologies for development - A/70/L.33*. New York: United Nations.

UNGA A/73/348. (2018, August 29). *Promotion and Protection of The Right to Freedom of Opinion*. New York, United States of America.

UNGA A/HRC/22/17/Add.4. (2013, January 11). *Annual report of the United Nations High Commissioner for Human RIghts*. New York, United States of America.

UNGA A/HRC/38/35. (2018, April 6). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. Retrieved August 30, 2019, from United Nations: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement

UNGA A/RES/70/1 para. 16 & 17. (2015, September 25). *Transforming our world: the 2030 Agenda for Sustainable Development*. New York, United States.

UNGA A/RES/70/125 para.3. (2015, December 16). *Resolution adopted by the General Assembly on 16 December 2015*. New York, United States of America.

UNGA Resolution 217. (1948, December 10). *Universal Declaration of Human Rights*. Paris, France.

UNGA Resolution 2200A (XXI). (1966, December 19). *International Covenant on Civil and Political Rights* . New York, United States of America.

UNHRC. (2011, July 21). *General Comment No. 34*. Geneva, Switzerland.

UNHRC A/HRC/17/31. (2011). *Guiding Principles on Business and Human Rights*. Geneva: United Nations Publishing Service.

UNHRC A/HRC/26/29. (2014, April 14). *Report of The Special Rapporteur on The Rights to Freedom of Peaceful Assembly and of Association*, Maina Kiai. New York, United States of America.

UNHRC A/HRC/39/29. (2018, August 3). *The Right to Privacy in The Digital Age*. New York, United States of America.

UNHRC A/HRC/40/63. (2019, February 27). New York, United States of America.

UNHRC HRI/GEN/1/Rev.9 (Vol. I). (1988, April 8). *General Comment No. 16 The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation*. New York, United States.

United for Gender Parity. (2018). *Making Progress Toward Achieving Gender Parity*. Re-

trieved October 18, 2019, from United for Gender Parity: https://www.un.org/gender/

United Nations Population Division. (2019). *World Population Prospects 2019*. Retrieved October 18, 2019, from United Nations Population: https://population.un.org/wpp/

UNSG. (2017, September 15). *Interview: UN should be flagbearer when it comes to gender parity, stresses top official*. Retrieved October 21, 2019, from United Nations: https://news.un.org/en/story/2017/09/564982-interview-un-should-be-flagbearer-when-it-comes-gender-parity-stresses-top

UNSG. (2018, July 12). *Secretary-General's High-level Panel on Digital Cooperation*. Retrieved October 21, 2019, from United Nations: https://www.un.org/en/digital-cooperation-panel/

Value Colleges. (2019). *Top 50 Best Value Bachelor's in Artificial Intelligence and Machine Learning*. Retrieved October 18, 2019, from Value Colleges: https://www.valuecolleges.com/ranking/best-value-machine-learning-bachelors/

Van der Spuy, A. (2017). *What if we all governed the Internet? Advancing multistakeholder participation in Internet governance*. Paris: UNESCO.

Vandenhole, W. (2005). *Non-Discrimination and Equality in the View of the UN Human Rights Treaty Bodies*. Oxford, UK: Intersentia.

Varian, H. (2018, June). *Artificial Intelligence, Economics, and Industrial Organisation*. In A. Agrawal, J. Gans, & A. Goldfarb, The Economics of Artificial Intelligence: An Agenda (pp. 399-419). Chicago: Chicago University Press.

Varian, H. (2019, June). *Artificial Intelligence, Economics, and Industrial Organisation*. In A. Agrawal, J. Gans, & A. Goldfarb, The Economics of Artificial Intelligence: An Agenda (pp. 399-419). Chicago: Chicago University Press.

Varley, C. (2018, April 6). *Are Sex Robots Just Turning Women into Literal Objects*. Retrieved October 21, 2019, from BBC: https://www.bbc.co.uk/bbcthree/article/8bbe0749-62ee-40f9-a8ac-a2d751c474f6

Vayana, E., & Tasioulas, J. (2016). *The Dynamics of Big Data and Human Rights: The Case of Scientific Research*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 28, 1-14.

Vincent, J. (2017, December 5). *Tencent says there are only 300,000 AI engineers worldwide, but millions are needed*. Retrieved October 21, 2019, from The Verge: https://www.theverge.com/2017/12/5/16737224/global-ai-talent-shortfall-tencent-report

Vincent, J. (2017, August 22). *Transgender YouTubers Had Their Videos Grabbed to Train Facial Recognition Software*. Retrieved October 21, 2019, from The Verge: https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset

Wajcman, J. (1991). *Feminism Confronts Technology*. Cambridge: Polity Press.

Wajcman, J. (2010). *Feminist Theories of Technology*. Cambridge Journal of Economics, 34(1), 143-152.

Wallis, J. (2018, September 27). *Is Artificial Intelligence Sexist?* Retrieved October 21, 2019, from The Globe and Mail: https://www.theglobeandmail.com/business/careers/leadership/article-is-artificial-intelligence-sexist/

Wang, Y., & Kosinski, M. (2018). *Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images*. Journal of Personality and Social Psychology, 114(2), 246-257.

Waters, S. (2018). *The Effects of Mass Surveillance on Journalists' Relations with Confidential Sources: A Constant Comparative Study*. Digital Journalism, 6(10), 1294-1313.

Wiggers, K. (2018, December 28). *A researcher trained AI to generate Africa masks*. Retrieved October 21, 2019, from Venture Berat: https://venturebeat.com/2018/12/28/a-researcher-trained-ai-to-generate-africa-masks/

Wilson, C. (2011, October 25). *Stanford's machine-learning class tackles the lazy hiker principle*. Retrieved October 21, 2019, from Slate: https://slate.com/technology/2011/10/stanfords-machine-learning-class-tackles-the-lazy-hiker-principle.html

Wilson, M. (2019, March 19). *The World' First Genderless AI Voice Is Here. Listen Now.* Retrieved October 21, 2019, from Fast Company: https://www.fastcompany.com/90321378/the-worlds-first-genderless-ai-voice-is-here-listen-now

Wolff, H. E. (2018, February 7). *How the continent's languages can unlock the potential of young Africans.* Retrieved October 21, 2019, from The Conversation: http://theconversation.com/how-the-continents-languages-can-unlock-the-potential-of-young-africans-90322

Wollerton, M., & Crist, R. (2018, September 20). *Amazon Echo Dot, Basics Microwave, Echo Sub: Everything Amazon just announced.* Retrieved October 21, 2019, from CNET: https://www.cnet.com/news/new-amazon-echo-dot-microwave-sub-alexa-hardware-event-september-20-2018/

Women in AI. (2019). *About Women in AI.* Retrieved October 18, 2019, from Women in AI: https://www.womeninai.co/

Women's Forum for The Economy & Society. (2019). *2019 Presentation.* Paris: Women's Forum for The Economy & Society.

Working Group on Artificial Intelligence. (2018). *Mauritius Artificial Intelligence Strategy.* Port Louis: Government of Mauritius.

World Bank. (2013). *Inclusion Matters: The Foundation for Shared Prosperity*. Washington D.C.: The World Bank.

WSIS-05/TUNIS/DOC/6(Rev. 1)-E art. 80 & 87. (2005, November 18). *Tunis Agenda for The Information Society*. Tunis, Tunisia.

Young, H. (2015, May 25). *The digital language divide*. Retrieved October 21, 2019, from The Guardian: https://www.theguardian.com/education/ng-interactive/2015/may/28/language-barrier-internet-experience

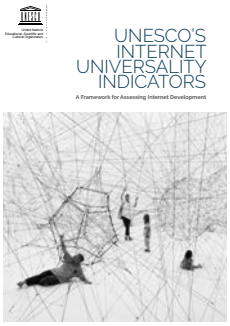
Yulianto, B., & Shidarta, S. (2015). *Philosophy of Information Technology: Sex Robot and Its Ethical Issues*. International Journal of Social Ecology and Sustainable Development, 6(4), 67-76.

Zenvus. (2019). *Home*. Retrieved October 18, 2019, from Zenvus: https://www.zenvus.com/

Zerrour, L. (2019, March 18). *50 Millions DH pour Promouvoir La Recherche en Intelligence Artificielle*. Retrieved October 21, 2019, from Aujourd'hui Le Maroc: http://aujourdhui.ma/societe/50-millions-dh-pour-promouvoir-la-recherche-en-intelligence-artificielle

Zuckerberg, M. (2018, November 15). *A Blueprint for Content Governance and Enforcement*. Retrieved October 21, 2019, from Facebook: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

UNESCO'S INTERNET UNIVERSALITY INDICATORS



Aiming for an Internet that promotes human right standards, supports inclusive Knowledge Societies and advances sustainable development: such is the foundation of the concept of Internet Universality, which has guided UNESCO's positioning on Internet issues since 2015. *UNESCO's Internet Universality Indicators* are a holistic tool to assess the state of Internet development in a given country, enabling new insights and evidence-based policy improvements to emerge. The publication features the full 303 Indicators in 6 categories, sources and means of verification, as well as an implementation guide for voluntary national assessments.

UNESCO SERIES ON INTERNET FREEDOM

UNESCO has published 11 editions as part of flagship series on Internet Freedom since 2011. This series explores the changing legal and policy issues related to the Internet while providing Member States and other stakeholders with policy recommendations, with the goal of fostering a conducive environment to freedom of expression on the net. These include:

- *What if we all governed the Internet? Advancing multistakeholder participation in Internet governance* (2017)
- *Survey on Privacy in Media and Information Literacy with Youth Perspectives* (2017)
- *Protecting Journalism Sources in the Digital Age* (2017)
- *Human rights and encryption* (2016)
- *Privacy, free expression and transparency: Redefining their new boundaries in the digital age* (2016)
- *Principles for governing the Internet: A comparative analysis* (2015)
- *Countering online hate speech* (2015)
- *Building digital safety for journalism: A survey of selected issues* (2015)
- *Fostering freedom online: The role of Internet intermediaries* (2014)
- *Global survey on Internet privacy and freedom of expression* (2013)
- *Freedom of connection, freedom of expression: The changing legal and regulatory ecology shaping the Internet* (2011)

All publications can be downloaded from the following link:
http://en.unesco.org/unesco-series-on-internet-freedom

Steering AI and Advanced ICTs for Knowledge Societies

This report recognizes artificial intelligence (AI) as an opportunity to achieve the United Nations Sustainable Development Goals (SDGs), through its contribution to building inclusive knowledge societies.

Based on UNESCO's Internet Universality ROAM framework agreed by UNESCO's Member States in 2015, this study analyzes

- how AI and advanced information and communication technologies (ICTs) will impact **Human Rights** in terms of freedom of expression, privacy, media, journalism and non-discrimination;
- how **Openness** needs inform the technological and safety challenges related to AI;
- how **Access** to AI hinges upon access to algorithms, hardware, human resources and data; and
- how a **Multi-stakeholder** approach concerning AI governance can address the challenges and opportunities for the benefit of humanity.

The study also offers a set of options for action that can help inform the development of new policy frameworks, and the re-examination of existing policies, as well as other actions for all stakeholders, namely Member States, the private sector, the technical community, civil society and intergovernmental organizations, including UNESCO.

United Nations
Educational, Scientific and
Cultural Organization

Communication and
Information Sector

9 789231 003639