TOWARDS REGULATION OF AI SYSTEMS

Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law

Compilation of contributions
DGI (2020)16

Prepared by the
CAHAI Secretariat

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

TOWARDS REGULATION OF **AI SYSTEMS**

Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe's standards on human rights, democracy and the rule of law

Compilation of contributions
prepared by the CAHAI Secretariat
December 2020

This publication has been funded through a voluntary contribution of Japan

Authors:

Isaac BEN-ISRAEL
Jorge CERDIO
Arisa EMA
Leehe FRIEDMAN
Marcelo IENCA
Alessandro MANTELERO
Eviatar MATANIA
Catelijne MULLER
Hideaki SHIROYAMA
Effy VAYENA

Council of Europe Study
DGI (2020)16

Layouts and cover page: Council of Europe, Information Society Department

Images: Shutterstock

This publication has not been copy-edited by the SPDP Editorial Unit to correct typographical and grammatical errors.

Contents

OPENING WORDS

Claudia Luciani

Director – Directorate of Human Dignity, Equality and Governance
Director, Council of Europe

Jan Kleijssen

Director – Directorate of Information Society and Action against crime, Council of Europe

Artificial intelligence (AI) systems are increasingly being used in almost every kind of human activity. The benefits of this technology are recognised and are already part of our everyday life. In the context of the fight against Covid-19, numerous applications have been deployed to accelerate research, improve case detection and measure the pandemic. However, the development of this technology raises public concern and it is the responsibility of States to ensure that, in this new technological era, human rights, democracy and the rule of law continue to be fully protected.

The Council of Europe is the continent's leading human rights organisation with 47 member States and protects and promotes human rights, democracy and the rule of law at pan-European level. The Council of Europe thus has a clear role to address the issue of the development and uses of artificial intelligence. The Organisation has already produced pioneering global legally-binding standards involving complex technological issues, such as the protection of personal data, bioethics and cybercrime, reconciling innovation and human rights protection. Moreover, the Council of Europe has developed many instruments addressing the impact of AI systems on human rights, democracy and the rule of law.

A step further has been taken with the setting up of the Ad hoc Committee on Artificial Intelligence (CAHAI), which is responsible for examining, on the basis of broad multi-stakeholder consultations, the feasibility and potential elements of a legal framework for the development, design and application of artificial intelligence, based on Council of Europe standards in the field of human rights, democracy and the rule of law. The Committee has a unique composition bringing together member and observer States, as well as observers from civil society, academia and the private sector. Convinced of the importance of a global reflection and of combining efforts in this field, the CAHAI works in close co-operation with other international institutions, such as UNESCO, the OECD and the European Union.

This publication aims to support the ongoing reflections within the CAHAI on the analysis of the challenges arising from AI systems and possible regulatory responses. Firstly, it sets out to inform the reader of the progress of the work of the CAHAI and presents several studies produced under auspices of the CAHAI on the impact of AI systems on human rights, rule of law and democracy, as well as on existing international legally-binding instruments and ethical guidelines on AI. Secondly, it brings in national perspectives from different States in order to support the development of an international legal framework on the use of certain AI systems based on the standards established by the Council of Europe.

We are confident that this publication will provide a valuable perspective on the current debate on the regulation of AI systems, and highlight the unique contribution that the Council of Europe and the CAHAI provide.

SUMMARY

TITLE 1. INTERNATIONAL PERSPECTIVE

Preliminary Chapter introduces the present report, submitted by the CAHAI to the Committee of Ministers and details the progress achieved to date, taking into account the impact of COVID-19 pandemic measures. It also includes reflections on working methods, synergy and complementarity with other relevant stakeholders and proposals for further action by the CAHAI by means of a robust and clear roadmap.

Chapter 1 outlines the impact of AI on human rights, democracy and rule of law. It identifies those human rights, as set out by the European Convention on Human Rights ("ECHR"), its Protocols and the European Social Charter ("ESC"), that are currently most impacted or likely to be impacted by AI.

Chapter 2 maps the relevant corpus of soft law documents and other ethical-legal frameworks developed by governmental and non- governmental organisations globally with a twofold aim. First, we want to monitor this ever-evolving spectrum of non-mandatory governance instruments. Second, we want to prospectively assess the impact of AI on ethical principles, human rights, the rule of law and democracy.

Chapter 3 aims to contribute to the drafting of future AI regulation by building on the existing binding instruments, contextualising their principles and providing key regulatory guidelines for a future legal framework, with a view to preserving the harmonisation of the existing legal framework in the field of human rights, democracy and the rule of law.

TITLE II. NATIONAL PERSPECTIVES

Chapter 1 sets forth the current state of play in Israel's policy development, with respect to the opportunities and challenges presented by artificial intelligence (AI) in relation to human rights and ethics. It is based, to a large extent, on the report of Israel's National Initiative for Secured Intelligent Systems, which has been recently submitted to the Israeli government.

Chapter 2 summarizes the nature of AI governance and the characteristics of the discussions in Japan. Since this report mainly deals with governance attempts and discussions up to 2018, it is easy to imagine that the details contained within it will change with future technological developments and changes in social conditions. However, organizing discussions at a fixed point is useful for future discussions on AI governance and for comparative research with other cutting-edge technology governance.

Chapter 3 aims to present some of the most pervasive and extended uses of Artificial Intelligence Applications (AI-Applications) in Mexico as well as the regulatory framework applicable to AI-Applications. We aim at representing with a high degree of accuracy the context under which each AI system operates. The context, plus a brief description of the system, will hopefully provide the reader with enough information to produce comparisons to other jurisdictions and to shed some light on the complexities surrounding the potential regulation of AI-Applications in Mexico and elsewhere.

TITLE I. INTERNATIONAL PERSPECTIVE OF AI SYSTEMS REGULATION BASED ON THE COUNCIL OF EUROPE'S STANDARDS

PRELIMINARY CHAPTER. Progress report of the Ad hoc Committee on Artificial Intelligence (CAHAI)[1]

I. Executive summary

1. On 11 September 2019, the Committee of Ministers adopted the terms of reference of the Ad hoc Committee on Artificial Intelligence (CAHAI), mandating the Committee to examine, on the basis of broad multi-stakeholder consultations, the feasibility and potential elements of a legal framework for the development, design and application of artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law. The present report is submitted by the CAHAI to the Committee of Ministers and details the progress achieved to date, taking into account the impact of COVID-19 pandemic measures. It also includes reflections on working methods, synergy and complementarity with other relevant stakeholders and proposals for further action by the CAHAI by means of a robust and clear roadmap.

2. The report underlines that the Council of Europe has a crucial role to play to ensure that Artificial Intelligence (AI) complies with the Organisation's standards on human rights, democracy and the rule of law. The CAHAI has focused its work on the mapping of relevant international and national legal frameworks and ethical guidelines, as well as on the analysis of the risks and opportunities arising from artificial intelligence, notably their impact on human rights, the rule of law and democracy.

3. The preliminary analysis carried out so far confirms the importance of deepening the reflections on the feasibility and the development of the elements of a "horizontal", cross-cutting legal framework to regulate the use and effects of AI applications, which would draw on the Organisation's unique expertise and work carried out at "vertical", sectorial level. Its clear relevance has also been confirmed and reinforced by the recent COVID-19 pandemic. Such a legal framework could consolidate existing standards in this field or develop additional standards necessary in the digital age. It could be based on a human rights impact assessment approach and contain practical operational mechanisms. Finally, it could also provide the basis for other initiatives and instruments, which remain indispensable for comprehensively addressing the challenges posed by AI applications in the relevant fields of activity of the Council of Europe.

4. The work undertaken by the CAHAI also provides an opportunity to contribute and complement other international initiatives in this area (i.e. by the OECD, the European Union - in particular the European Commission - UNESCO and the United Nations in general) by enacting a concrete instrument based on human rights, as part of a global legal mechanism for the regulation of digital technologies, an area in which the Council of Europe can bring real added value. Close co-ordination and synergies with the work of these organisations in the field of artificial intelligence will continue on a regular basis.

5. Finally, the report includes the working version to date of the table of contents of the future feasibility study and a clear roadmap setting out key timelines of the process. New working methods have been considered, including the creation of three thematic working groups in charge of specific tasks pertaining to the CAHAI's feasibility study, namely the preparation of proposals on specific policy aspects, the preparation of multi-stakeholder consultations and the elaboration of legal frameworks, which should ensure that substantive progress in the fulfilment of CAHAI's mandate is achieved over the coming months and until December 2021.

[1] 23 September 2020, CM(2020)90-final.

II. Background and CAHAI's mandate

6. For several years, the Council of Europe has been assessing and anticipating the impact of digital technologies on human rights, democracy and the rule of law. It has been developing relevant legal instruments and two Internet Governance strategies2 to ensure that our society and individuals fully reap the benefits of innovative practices. Among these technologies, those based on artificial intelligence (AI) stand out from traditional computer applications because of their autonomy. Increasingly used in a wide range of public and private services, these particular technologies offer great opportunities for development but also raise important and complex ethical, legal, political and economic issues.

7. The Council of Europe has committed to framing their scope and implications in most of its specialised areas of activity, such as justice, data protection, equality and non-discrimination3. The impact of AI on individuals and society has also been examined from different angles by the Committee of Ministers, the Parliamentary Assembly and the Commissioner for Human Rights4. The Organisation is now taking a complementary approach to the initiatives implemented to-date, having adopted a cross-cutting strategic approach based on the examination of the feasibility of binding and non-binding frameworks with the aim of ensuring that the design, development and application of AI are in line with European standards of human rights, democracy and the rule of law. A high-level conference, organised by the Finnish Chairmanship of the Committee of Ministers and the Council of Europe in Helsinki (Finland) on 26-27 February 2019, noted the importance and urgency of policy responses to the impact of AI on human rights, democracy and the rule of law, and gave the necessary impetus to the creation of the Ad hoc Committee on Artificial Intelligence (CAHAI), whose terms of reference were adopted by the Committee of Ministers on 11 September 2019.

8. The CAHAI is responsible for examining the feasibility and potential elements of a legal framework for the development, design and application of AI, based on Council of Europe standards in the field of human rights, democracy and the rule of law. A broad multi-stakeholder consultation and close co-ordination with other international organisations will be ensured within the framework of its mandate.

9. Economic and other benefits arising from the application of AI and the importance of AI for Europe to keep pace with the rest of the world feature amongst the issues which the CAHAI might examine as part of its work.

III. Progress of work

10. To date, the CAHAI held its first plenary meeting on 18, 19 and 20 November 2019[5] and a second plenary meeting took place on 6-8 July 2020[6]. During its first plenary meeting, the Committee elected its Chairperson (Gregor Strojin, Slovenia), Vice-Chairperson (Peggy Valcke, Belgium) and its Bureau (composed, in addition to the Chairperson and Vice-Chairperson, of five members from Estonia, France, Germany, Italy and Switzerland). To-date, the Bureau has held

[2] Internet Governance – Council of Europe Strategy 2016 - 2019 and 2012 - 2015.
[3] European Commission for the Efficiency of Justice (CEPEJ), European Ethical Charter for the Use of Judicial Intelligence in Judicial Systems and their Environment; Consultative Committee of the Convention for the Protection of Individuals with regard to the Processing of Personal Data (Convention 108), Guidelines on Artificial Intelligence and Data Protection; European Commission against Racism and Intolerance (ECRI), Study, Discrimination, Artificial Intelligence and Algorithmic Decisions.
[4] Declaration of the Committee of Ministers on Manipulative Capabilities of Algorithmic Processes, Recommendation of the Parliamentary Assembly of the Council of Europe on Technological Convergence, Artificial Intelligence and Human Rights and Recommendation of the Commissioner for Human Rights "Unboxing AI: 10 steps to protect human rights".
[5] See the abridged report of the first plenary meeting of CAHAI.
[6] See the report of the second plenary meeting of CAHAI. The meeting originally scheduled for 11-13 March 2020 had to be postponed due to the international situation concerning the COVID-19 virus and in particular the various restrictions put in place within the Council of Europe and its member States.

three meetings, respectively on 19 November 2019[7], on 23 and 24 January 2020[8] and an online meeting on 27 March 2020[9]. The Committee also appointed the Gender Equality Rapporteur (Jana Novohradska, Slovak Republic) at its first plenary meeting.

11. During its first plenary meeting, the CAHAI specified that the feasibility study should include a mapping of national and international legal instruments (both of the Council of Europe and other international organisations) and ethical frameworks related to AI applications, as well as a mapping of risks and opportunities arising from the development, design and application of AI, including its impact on human rights, democracy and the rule of law. The findings resulting from the mappings should be considered for deciding the appropriateness of a definition of AI and for defining a suitable legal framework for the design, development and application of AI based on Council of Europe standards. Three independent experts were commissioned to frame the feasibility study, with a view to supporting CAHAI decision-making on this matter, and submitted their reports to the CAHAI at the second plenary meeting.

12. With regard to the mapping of international legal instruments applicable to AI, the approach chosen is to identify, in the many sectors of activity of the Council of Europe and with the support of the relevant intergovernmental committees, the relevant international instruments, the guiding legal principles and the main values emerging from these instruments, as well as their potential gaps. In particular, the possible emergence of common AI-specific "horizontal" and “transversal” principles, which are the basis for specialised "vertical" principles defined (or being defined) in the different fields of activity of the Council of Europe, should be examined.

13. The results of online consultations conducted among the member States of the Council of Europe in November 2019 – February 2020 are being used to map national initiatives on AI.

14. The mapping of ethical and non-binding frameworks analyses the vast production of charters, declarations and principles on AI developed in recent years by private, scientific and civil society actors, with the aim of identifying those that could contribute to the establishment of a possible legal framework on AI in line with Council of Europe standards.

15. The mapping of risks and opportunities arising from the development, design and application of AI, including the impact of AI on human rights, democracy and the rule of law, is carried out specifically in relation to the rights protected by the Convention for the Protection of Human Rights and Fundamental Freedoms (hereafter: the European Convention on Human Rights) and aims to determine to what extent their exercise could be impacted by certain applications of AI and what strategies should be adopted to address this. The answers received in the framework of the online consultation may also offer some help in respect of the strategies in place in member States to mitigate the impact of AI on the rights protected by the Convention.

16. These mappings shall aim to identify applicable human rights legal frameworks, which are important in a digital age. Nevertheless, the preliminary analysis carried out so far points out, on the one hand, the limitations of existing legal regimes, drafted prior to the development of AI applications, which reduce their effectiveness in responding adequately to current challenges; and, on the other hand, the significant impact of AI on a number of standards defined in Council of Europe instruments and in particular on the rights protected by the European Convention on Human Rights. In this regard, the question of the role and responsibilities of all relevant stakeholders, including States, business enterprises, research institutions and civil society, in developing and deploying AI technologies which respect human rights should be further explored.

17. Faced with these challenges, the Council of Europe has a key role to play in ensuring that the development, design and application of AI technologies is in line with the Organisation's

[7] See the abridged report of the first meeting of the CAHAI Bureau.
[8] See the abridged report of the second meeting of the CAHAI Bureau.
[9] See the abridged report of the third meeting of the CAHAI Bureau.

standards. Therefore, it is important to deepen the reflections on the development of the elements of a legal framework to regulate the use and effects of AI applications, drawing on the Organisation's unique expertise and based on a human rights impact assessment approach. Furthermore, this legal framework could consider the advisability of consolidating existing standards in this field (such as those of the European Convention on Human Rights or the 108+ Convention) or developing additional standards necessary in a digital age. Finally, it could also provide the basis for a number of sectorial initiatives and instruments, which remain indispensable for comprehensively addressing the challenges posed by AI applications in the relevant fields of activity of the Council of Europe.

18. It is also an opportunity to support other international initiatives (OECD, the European Union - in particular the European Commission - UNESCO and the United Nations in general) by enacting a concrete instrument based on human rights, as part of a global legal mechanism for the regulation of digital technologies in which the Council of Europe has a real added value.

19. Furthermore, it would be important to also examine, in the framework of the feasibility study, the mechanisms and actions that could be taken to ensure the effectiveness of the proposed legal framework, both at the level of the Council of Europe and at the level of the member States. This would include, in particular, questions relating to the creation of mechanisms of collaborative law-making (in particular the use, in some specific cases, of "sandboxing[10]") and of *ex-ante* verification and/or certification of AI applications or control mechanisms by independent authorities, as well as the advisability of regulating the professions of mega-data experts (for instance the opportunity to establish a professional order for *data scientists*, to develop deontological charters based on ethical principles, a "Hippocratic Oath" for AI professionals or the creation of a "Driver's License" for AI, etc.).

20. Through the process of preparing the feasibility study, including in activities of working groups and in the framework of multi-stakeholder consultation, the CAHAI will aim to integrate a gender equality perspective, contribute to building cohesive societies and to promoting and protecting rights of persons with disabilities.

21. During the second plenary meeting of the CAHAI, member States expressed their views on the contents of the different items of the draft table of contents of the feasibility study, which is included in Annex I of this report.

IV. The impact of the COVID-19 pandemic on CAHAI's activities

22. AI is one of the many tools used by States to contain and combat the COVID-19 pandemic. Several applications of AI have been developed in areas such as supporting research for vaccine development, formulating diagnostics to support healthcare personnel and developing predictive models for the possible evolution of the pandemic. AI has been used to facilitate the analysis of the thousands of research papers published on the pandemic and to ensure better sharing of scientific knowledge, as well as for disease screening purposes. Furthermore, AI might be useful for combating coronavirus misinformation, as long as data protection rights and freedom of speech are adequately protected.

23. Other uses aimed at controlling epidemic risks were also identified: for example, facial recognition and biometric devices, geolocation devices and drones were used to ensure compliance with containment measures by at-risk or infected individuals or populations. Applications have also been developed to warn users that they have been in contact with people potentially carrying the virus. This type of use has a significant impact on certain rights and freedoms protected by international human rights instruments, including the European Convention on Human Rights, such

[10] The process where regulators allow a new technology to be used and tested within a closed or limited environment and in close dialogue with policymakers. In addition to the technology being assessed, this mechanism also allows regulators to try out new rules and observe their impact on the technology in an environment where wider damage or danger to the public is limited.

as the right to respect for private and family life (right to privacy in particular) and freedom of movement, and should be considered by the CAHAI. The existing or envisaged use of AI applications in certain countries for epidemic monitoring and control as well as for the prevention of new disease outbreaks could lead to the trivialisation of the use of mass surveillance of populations in the absence of adequate safeguards.

24. COVID-19 lockdown measures resulted in an unintentional discrimination of children and young people who do not have access to laptops and do not attend schools which offer online courses. This caused an interruption to their education and an impediment to the advancement of their digital skills including AI as an intersectional field of study. In times of economic crisis, which may come as a result of the COVID-19 crisis, there is a historical pattern of discrimination of girls and young women when families make decisions as to whether to invest into higher or technical education of children. Girls and young women are often disadvantaged, and side lined in favour of boys and young men. This will even further exacerbate the status quo of under representation of women in AI development roles and AI related decision-making roles. The CAHAI, through its Policy Development Group (CAHAI-PDG), could pay attention to this issue and focus on measures aimed to mitigate this historical pattern of discrimination of girls and young women in times of economic crisis. The application of temporary positive gender mainstreaming measures within the AI policy could be considered to protect and support women and young girls, so they are not left behind and given access to decision making roles within the AI ecosystem.

25. In this context, the relevance of a legal framework on the design, development and application of AI based on the principles and values of the Council of Europe is enhanced. The CAHAI might wish to consider the implications and, where appropriate, the need for a legal framework to cater for situations such as the use of AI applications in specific situations, as is the case for other Council of Europe' treaties, and in accordance with the principles and values of the Council of Europe.

V. Working methods, documents and roadmap

26. During its first plenary meeting, the CAHAI entrusted its Bureau, in accordance with Resolution CM/Res(2011)24[11], with the task of carrying out the necessary mapping for the feasibility study. The CAHAI, taking note of the legal opinion of the Council of Europe's Jurisconsult, also authorised the ad hoc involvement of other member States in the work of the Bureau, in accordance with the modalities defined by the CAHAI and its Bureau.[12]

27. In order to ensure more active participation of the member States in CAHAI and its transparent and balanced approach to decision-making, the CAHAI held an exchange of views on a proposal by the Russian Federation to increase the number of Bureau members by six (6), to a total number of thirteen (13). It was proposed that new members could participate in the Bureau's meetings without defrayal of their expenses by the Council of Europe.[13]

[11] See in particular Article 13 of Appendix 1 to Resolution CM/Res(2011)24.
[12] See in this regard paragraphs 19 and 20 of the report of the 1st online plenary meeting. See also the Legal Opinion delivered by the Council of Europe's Jurisconsult, DD(2020)16 24/01/2020, paragraph 2, distributed at the request of the Secretariat during the GR-J meeting of 16 January 2020: "Resolution CM/Res(2011)24 on intergovernmental committees and subordinate bodies, their terms of reference and working methods does not prohibit a committee from exceptionally associating other members to the work of the Bureau provided the following conditions are fulfilled:
− the CAHAI adopts this decision by consensus (no objection has been expressed);
− the possibility to attend Bureau meetings will be offered to all members of the CAHAI on equal footing;
− the additional members would not become Bureau members and would not be entitled to the defrayal of their expenses;
− the Bureau retains the right to reserve certain parts of its meetings to the participation of the elected Bureau members only;
− Resolution CM/Res(2011)24 remains fully applicable, in particular no decision on substantive issues would be taken by the Bureau".
[13] See paragraphs 86 and 87 of the report of the second CAHAI plenary meeting.

28. Following the postponement of the meeting scheduled from 11 to 13 March 2020 due to the COVID-19 pandemic, concrete proposals for adapting working methods were discussed by the Bureau in order to ensure that CAHAI could continue its work effectively and in accordance with the objectives set, unless it would prove challenging to do so and until the next plenary meeting could be held, notably under conditions that would guarantee the safety and participation of members and observers. The CAHAI Bureau instructed the Secretariat to organise a written consultation of CAHAI members on several working documents and issues, including the adoption of this progress report and the consideration of new working methods for CAHAI, such as the establishment of small working groups. On the basis of this consultation, CAHAI members agreed to establish three working groups, dealing respectively with policy development, legal frameworks, consultations and outreach (in connection with the multi-stakeholder consultation), whose terms of reference are contained in Annex III to this document. Moreover, CAHAI adopted a document on working methods and functioning modalities of the working groups at its second plenary meeting[14].

29. As noted above, the feasibility study and the potential elements of the future legal framework should be based on a broad multi-stakeholder consultation. In this regard, CAHAI members indicated by submitting written comments in a first online consultation following the first plenary meeting that it should be addressed as a matter of priority to representatives of the private sector, civil society and the scientific community, as well as to other international organisations including technical standard-setting organisations. The CAHAI roadmap envisages that a preliminary draft of the feasibility study including the main elements of a future legal framework will be considered by the CAHAI at its third plenary meeting (30 November - 2 December 2020). This would provide a first deliverable that could be discussed in an open and transparent manner by the different stakeholders and subsequently be the basis for the multi-stakeholder consultations that would start in 2021. The year 2021 would also be possibly dedicated to the finalisation of the elements of the above-mentioned legal framework.

30. CAHAI's draft roadmap is attached as Annex II to this report.

VI. Synergy and complementarity of CAHAI's work with that of other international organisations

31. Co-ordination with existing or ongoing work on AI in other international organisations, in particular the European Union (such as the ongoing process initiated by the European Commission on the White Paper on AI) and the OECD, has been actively sought in order to promote synergies and avoid duplication. In addition to participation in CAHAI meetings and regular contacts between the institutions, the Chair of CAHAI and the Secretariat participated in the launch of the OECD AI Policy Observatory on 27 February 2020. In addition, exchanges with representatives of OECD, the European Union, ITU, ISO and UNESCO on the ongoing work within each organisation are scheduled to take place regularly during CAHAI plenary meetings. The objective is to promote synergies in the development of a legal framework on AI, in which the specific contribution and expertise of each organisation (in the case of CAHAI, expertise in human rights, democracy and the rule of law in particular) can be highlighted and complement each other.

VII. Budgetary implications

32. The CAHAI's budget, as adopted by the Committee of Ministers, only covers the holding of plenary and Bureau meetings and partially the holding of a limited number of working group meetings. The budget allocated does not allow for complementary actions,[15] such as strengthening

[14] See CAHAI (2020) 10ADD rev1.
[15] The cancellation of the 11-13 March 2020 plenary meeting and the impossibility of recovering some costs already incurred also have an impact on the budget envelope available for CAHAI in 2020.

the dimension of the multi-stakeholder consultations, involving other high-level experts, and/or organising actions likely to enhance the visibility of the Council of Europe's work in this field.

33. Therefore, CAHAI member and observer States are invited to consider contributing financially to the strengthening of CAHAI's activities, through voluntary contributions or any other means they deem appropriate.

VIII. Conclusions and concrete proposals for action

34. The Council of Europe has a crucial role to play today to ensure that AI applications are in line with human rights protection and comply with the Organisation's standards when they exist. Thus, CAHAI's initiative remains unique and complementary to those undertaken by other international organisations. It is important that work on the elaboration of potential elements of a legal framework can start as early as January 2021 and that CAHAI be provided with additional financial means in the implementation of its mandate.

35. The Committee of Ministers is invited to instruct the CAHAI to undertake the following:

(i) make substantive progress in the drafting of its feasibility study on a legal framework by November 2020, with a view to starting in January 2021 a reflection on the elements of a legal framework that would be the subject of a broad multi-stakeholder consultation; this legal framework could regulate the design, development and application of AI that have a significant impact on human rights, democracy and the rule of law. It could also consider the desirability of consolidating existing standards through an interpretation of the norms, principles and values already enacted in this area or developing new standards required for the digital age. Finally, it would lay the foundations on which a number of initiatives and instruments could be further developed in the different sectors of activity of the Council of Europe, which remains indispensable to comprehensively address the challenges posed by AI applications in the relevant fields of activity of the Council of Europe;

(ii) propose, simultaneously, complementary measures to operationalise the above-mentioned legal framework: in particular, reference could be made to the prior human rights impact assessment procedure, the means of validation or certification of algorithms and AI systems or the training and organisation of certain professions involved in the application of AI tools.

36. The Committee of Ministers is also invited to take note of the CAHAI's roadmap and to propose to member or observer States, which so wish, to contribute financially to the strengthening of CAHAI's activities, through voluntary contributions or any other means they consider useful.

Annexes

I. Draft table of contents of the feasibility study

II. CAHAI's roadmap

III. Terms of reference of the Working Groups

Annex I. Draft table of contents of the Feasibility Study

(working version: 16 June 2020)

1. **General Introduction**

2. **Scope of a Council of Europe legal framework on artificial intelligence**

3. **Opportunities and risks** arising from the design, development and application of artificial intelligence on human rights, the rule of law, and democracy. “Green” and “red” areas” - meaning respectively positive and problematic examples of artificial intelligence applications from a human rights, the rule of law and democracy perspective, while considering the context-sensitive environment for artificial intelligence design, development and application in Europe and developments at global level.

4. **The Council of Europe's work on artificial intelligence to date**

5. **Mapping of instruments applicable to artificial intelligence**
   - i. International legal instruments applicable to artificial intelligence
   - ii. Ethical Guidelines applicable to artificial intelligence
   - iii. Overview of national instruments, policies and strategies related to artificial intelligence
   - iv. Advantages, disadvantages and limitations of existing international and national instruments and ethical guidelines on artificial intelligence
   - v. International legal instruments, ethical guidelines and private actors

6. **Key findings of the multi-stakeholder consultations**
   - i. Outline of the feasibility study
   - ii. Main findings on the type and content of a legal framework for the development, design and application of artificial intelligence, based on Council of Europe’s standards on human rights, democracy and the rule of law

7. **Main elements of a legal framework for the design, development and application of artificial intelligence**
   - i. Key values, rights and principles deriving - in a bottom-up perspective - from sectorial approaches and ethical guidelines; in a top-down perspective - from human rights, democracy and the rule of law requirements.
   - ii. Role and responsibilities of member States and of private actors in developing applications which are in line with such requirements
   - iii. Liability for damage caused by artificial intelligence

8. **Possible options** for a Council of Europe legal framework on the design, development and application of artificial intelligence based on human rights, rule of law and democracy (for each option: content, addressees, added value, role of private actors, member States' expectations resulting from the submitted written comments)
   - i. Updating of existing legally binding instruments

ii. Convention

iii. Framework Convention

iv. Soft law instrument(s)

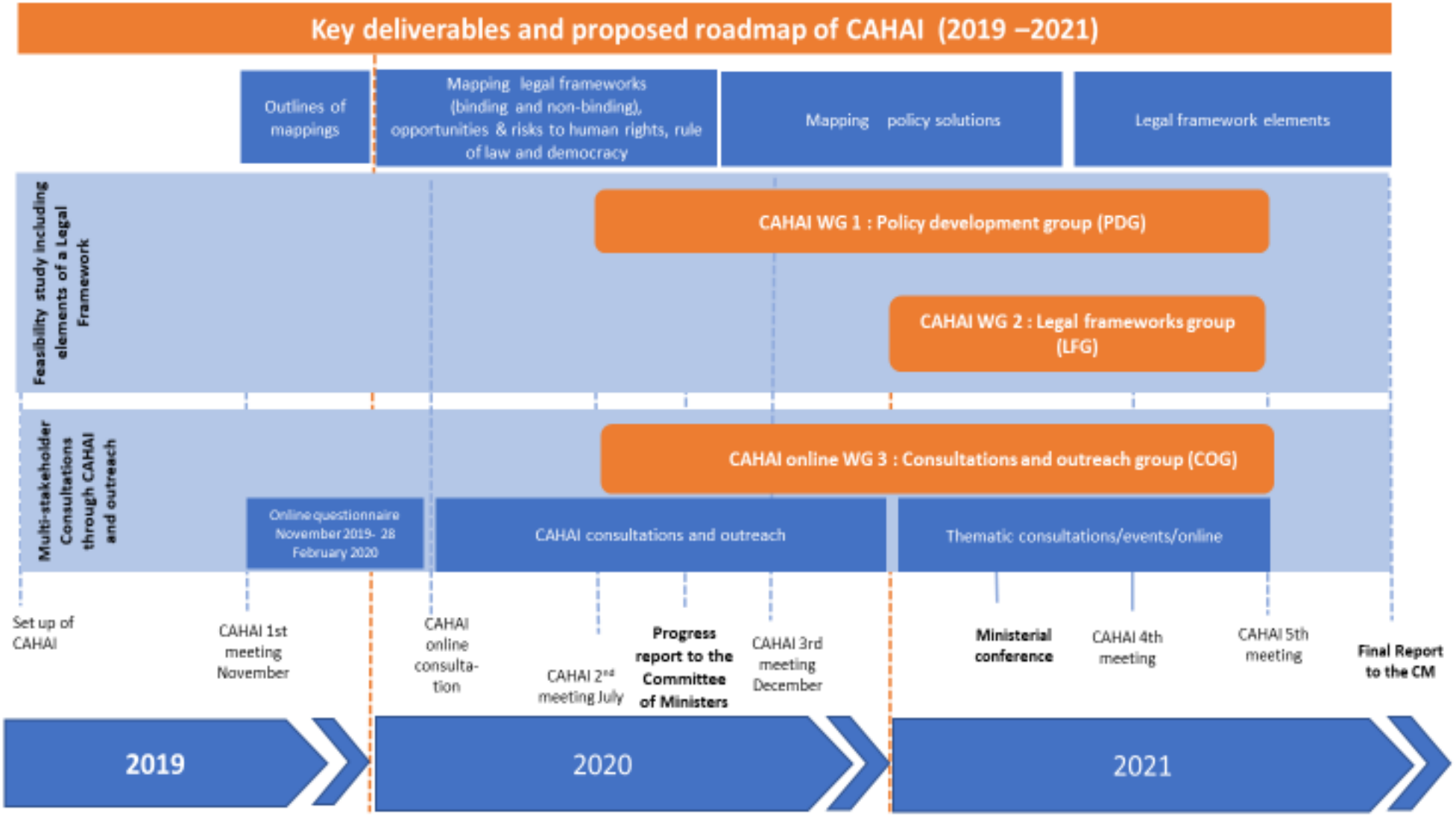
v. Other type of support to member States such as identification of best practices;

vi. Possible complementarity between the horizontal and cross-cutting elements that could form part of a conventional-type instrument and the vertical and sectoral work that could give rise to specific instruments of a different nature.

9. Possible practical mechanisms to ensure compliance with and effectiveness of the legal framework (such as for instance the creation of a mechanism of ex-ante verification and/or certification, oversight by independent authorities, sandboxing, etc.)

10. Final considerations

Annex II. CAHAI Roadmap



Annex III. Terms of reference of the Working Groups

Policy Development Group (CAHAI- PDG)

Composition

The CAHAI-PDG shall be composed of up to 20 experts designated by member States, and of participants and observers which express interest in taking part in this working group and contribute professionally and continuously to its work.

Tasks The CAHAI - PDG shall:

1) Contribute to the development of the feasibility study of a legal framework on artificial intelligence applications, building upon the mapping work already undertaken by the CAHAI and focusing in particular on the following issues:

a. To identify and analyse how AI applications by States and private actors impact on human rights, rule of law and democracy, including the risks, threats, and opportunities posed or brought by AI, including where appropriate through an evaluation of proposals on AI regulation made by member States and other stakeholders; and taking into account the importance of keeping pace with AI developments worldwide;

b. Based on the conclusions of the mapping, to prepare key findings and proposals on policy and other measures, to ensure that international standards and international legal instruments in this area are up-to-date and effective and prepare proposals for a specific legal instrument regulating artificial intelligence;

c. To identify high-risk and high-opportunity AI applications, consider and develop policy and other measures to address risks posed by them and ensure the protection of human rights, rule of law and democracy;

d. In connection with the work carried out under points b) and c) prepare proposals to facilitate implementation of relevant international and European human rights obligations, and aimed at supporting the implementation of effective human rights risk assessments and any other necessary actions to respond as necessary to significant new and emerging opportunities, threats and risks;

2) Develop proposals for engaging and consulting with the relevant stakeholders on the above-mentioned elements in close co-ordination and consultations with the Consultations and Outreach Group (CAHAI-COG).

Working methods

The CAHAI-PDG will hold 2 meetings in 2020, and subject to a plenary decision to that effect, any other additional meetings as required in 2021. It will conduct its work through technology, such as e-mail or other similar means of electronic communication, virtual meetings, and/or physical meetings. The group shall designate a Chair and a Co-chair among its members at its first meeting.

The Council of Europe's budget will cover the costs of participation of the chairpersons (for physical meetings) and those for the organisation of meetings through technological means, to enable the working group to perform its functions and responsibilities through online exchanges.

Participation of members for physical meetings will as a principle be borne by the sending institutions and organisations without defrayal of expenses.

Duration July 2020-December 2021

The Consultations and Outreach Group (CAHAI-COG)

Composition

The CAHAI - COG shall be composed of up to 20 experts designated by member States and of participants and observers which express interest in taking part in this working group and contribute professionally and continuously to its work.

Tasks

The CAHAI - COG shall take stock of the analysis undertaken by the secretariat of responses to the first online consultation and elaborate for CAHAI's consideration and approval:

1) A stakeholder analysis and mapping (due: September 2020)

2) On the basis of an outreach to countries having expressed interest in holding in-country consultations, a methodology and template(s) for use by member States in in-country consultations (due: November 2020)

3) A calendar of thematic consultations (due: November 2020) and an initial consultative document (due: December 2020)

4) An analysis of ongoing developments and reports which are directly relevant for CAHAI's working groups' tasks (due: ongoing – October 2020) as well as an analysis of contributions by respondents to online consultations for consideration by the CAHAI (due: 2021)

Working methods

The CAHAI-COG is expected to work primarily online and will be supported by the CAHAI secretariat. The group shall designate a Chair and a Co-chair among its members at its first meeting.

Duration July 2020-January 2021

Legal Frameworks Group (CAHAI- LFG)

Composition

The CAHAI - LFG shall be composed of up to 20 experts designated by member States, and of participants and observers which express interest in taking part in this working group and in contributing substantively to its work.

Tasks

The CAHAI - LFG shall:

1) Prepare key findings and proposals on possible elements and provisions of a legal framework with a view to elaborating, as appropriate, draft legal instrument(s), for consideration and approval by the CAHAI, taking into account the review of existing legal instruments applicable to artificial intelligence and policy options set out in the feasibility study approved by the CAHAI;

2) Develop specific proposals for regulation for the development, design and application of AI in the areas identified as risky by member States and other stakeholders, taking into account member States' regulatory approaches.

Working methods

The CAHAI-LFG will hold 2 meetings in 2021, and subject to a plenary decision to that effect, any other additional meetings as required. It will conduct its work through technology, such as e-mail or other similar means of electronic communication, virtual meetings, and/or physical meetings. The group shall designate a Chair and a Co-chair among its members at its first meeting.

The Council of Europe's budget will cover the costs of participation of the chairpersons (for physical meetings) and those for the organisation of meetings through technological means, to enable the working group to perform its functions and responsibilities through online exchanges.

Participation of members to physical meetings will as a principle be borne by the sending institutions and organisations without defrayal of expenses.

Duration January 2021 - December 2021

CHAPTER I. The Impact of AI on Human Rights, Democracy and the Rule of Law[16]

Catelijne Muller[17]

I. Introduction

1. AI, as a general purpose technology, has an impact on the entire fabric of society, In 2017, the European Economic and Social Committee, in what is widely considered the 'inception report' on the broader societal impact of AI, identified the most important societal impact domains including: safety; ethics; laws and regulation; democracy; transparency; privacy; work; education and (in)equality.[18] This means that AI has an impact on our human rights, democracy and the rule of law, the core elements upon which our European societies are built.

2. In 2019, the AI High Level Expert Group on AI presented Ethics Guidelines for Trustworthy AI.[19] These guidelines define trustworthy AI as being lawful, ethical and socio-technically robust. For the ethical element of trustworthy AI, the guidelines explicitly take fundamental rights as a basis for AI ethics.[20] While these guidelines do contain elements that are derived directly from existing (human) rights, they are not yet legally binding by themselves. Recently, the call for stronger (existing or new) legally binding instruments for AI has become louder. The European Commission announced potential elements of a legislative framework in its Whitepaper on AI[21] and stresses the importance of AI being in line with EU fundamental rights and the laws that ensure those rights.

3. This paper outlines the impact of AI on human rights, democracy and rule of law. It identifies those human rights, as set out by the European Convention on Human Rights ("ECHR"), its Protocols and the European Social Charter ("ESC"), that are currently most impacted or likely to be impacted by AI (Chapter II). In Chapters III and IV, it aims to provide a number of possible strategies that could be implemented simultaneously, if necessary. Chapter III looks at addressing the impact within the existing framework of human rights, democracy and the rule of law. Chapter IV looks at strategies, should the existing framework fail to adequately protect us. As technology and society are evolving quickly this paper cannot be exhaustive but prioritises the most relevant impacts to the extent that they can be identified today.

[16] Report prepared for the CAHAI. Strasbourg, 24 June 2020, named CAHAI(2020)06-fin
[17] President ALLAI, Member of the EU High Level Expert Group on AI, EESC Rapporteur on AI.
[18] EESC opinion on AI & Society (INT/806, 2017)
[19] Ethics Guidelines for Trustworthy AI, High Level Expert Group on AI to the European Commission, 2019
[20] Subsequently, the guidelines describe 7 requirements for trustworthy (i.e. lawful, ethical and robust) AI: 1) Human agency and oversight, 2)Technical robustness and safety, 3)Privacy and data governance, 4)Transparency, 5) Diversity, non-discrimination and fairness, 6) Social and environmental well-being, 7) Accountability
[21] Whitepaper on artificial intelligence, COM(2020) 65 final

II. Defining AI

4. AI has a myriad of applications that have already been introduced into society: biometric (including facial) recognition, object recognition, risk and success prediction, algorithmic decision making or support, automatic translation, recommender systems, and so on. These applications have found their way into sectors such as law enforcement, justice, human resource management, financial services, transport, healthcare, public services, etc.

5. AI remains an essentially contested concept, as there is no universally accepted definition. Nevertheless, definitions can broadly be clustered in two camps: rationalist and human-centric definitions. The most prominent rationalist definition, defines AI as "an agent created by humans that decides and performs actions based on its perception".[22] The best-known example of a human-centric definition is the Turing test, which is passed by a computer, as soon as it performs a task that would otherwise require human (conversational) intelligence. The High-Level Expert Group on AI has provided a definition of AI in 2019[23]:

Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

6. Often, AI is described as a collection of technologies that combine data, algorithms and computing power. While this is correct for the most widely used AI-systems at present, this is only a very limited description of what AI is. AI is a container term for many computer applications, some of which combine data and algorithms, but other, non-data- driven AI approaches, also exist, e.g. expert systems, knowledge reasoning and representation, reactive planning, argumentation and others.

7. Most AI systems that have been penetrating our societies lately, are indeed examples of data-driven AI, with particular impact on human rights, democracy and rule of law.

8. In the following the most relevant terms are defined:

- Narrow AI: AI systems that can perform only very specific 'narrow' tasks;
- Big (historical) Data: AI systems that need a lot of (historical) data to perform well. The quality, volume and content of the data influence the operation of these AI system, and often lead to replication and amplification of errors, gaps and biases in the data;
- Correlation: many AI systems only look for relations in data, which do not

[22] Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: A modern approach (3rd ed). Prentice Hall.
[23] EU High Level Expert Group on AI. A definition of AI, main capabilities and scientific disciplines, 2019.

establish (or “see”) a causal relationship between a 'case' and a decision, but merely makes a prediction based on shared characteristics with other 'cases';

- Black boxes: many AI systems often are so-called black boxes, within which (decision) processes take place that cannot be fully explained in human terms;
- Common sense: AI systems do not have common sense, meaning that while a system might be able to recognize a cat or a cancer cell, it has no conception of the idea of what a cat or a cancer cell is. It merely provides a label to a specific pattern. It also cannot use the information about a cat or a cancer cell to identify a dog or a headache.

All these characteristics can make current AI brittle, unstable and unpredictable, but also popular and widely applied.

9. Most importantly, AI systems are more than just the sum of their software components. AI systems also comprise the socio-technical system around it. When considering governance, the focus should not just be on the technology, but also on the social structures around it: the organisations, people and institutions that create, develop, deploy, use, and control it, and the people that are affected by it, such as citizens in their relation to governments, consumers, workers or even entire society.

i. Defining AI for regulatory purposes

10. A complicating factor is that legal definitions differ from pure scientific definitions whereas they should meet a number of requirements[24] (such as inclusiveness, preciseness, comprehensiveness, practicability, permanence), some of which are legally binding, and some are considered good regulatory practice[25]

III. Impact of AI on Human Rights, Democracy and the Rule of Law

11. Taking an 'AI lifecycle approach' is important, in order to consider not only the development stage of AI, but also the deployment and use stages. Another element to keep in mind is that most AI-applications currently being used could enshrine, exacerbate and amplify the impact on human rights, democracy and the rule of law at scale, affecting larger parts of society and more people at the same time.

12. Four "Families of Human Rights" under the ECHR, its Protocols ESC are impacted by AI:

i) Respect for Human Dignity
ii) Freedom of the Individual

[24] A Legal Definition of AI Jonas Schuett Goethe University Frankfurt September 4, 2019 *(Legal definitions must be: (i) inclusive: the goals of regulation must not over- or under-include. (Julia Black. Rules and Regulators. Oxford University Press, 1997. [32] Robert Baldwin, Martin Cave, and Martin Lodge. Understanding Regulation: Theory, Strategy, and Practice. Oxford University Press, 2nd edition, 2012.); (ii) Precise: it should be clear which case falls under the definition and which does not; (iii) Comprehensive: the definition should be understandable by those who are regulated; (iv) Practicable: legal professionals should be able to easily determine whether a case falls under the definition; (v) Permanent: the need for continued legal updating should be avoided.*

[25] Inclusiveness can be derived from the principle of proportionality in EU law (art. 5(4) of the Treaty on European Union. The criteria precision and comprehensiveness are based on the principle of legal certainty in EU law. The criteria practicability and permanent are considered good legislative practice.

iii) Equality, Non-Discrimination and Solidarity
iv) Social and Economic Rights

Moreover, AI has ample impact on:

v) Democracy
vi) The Rule of Law

It is important to note that many AI-systems or uses can impact various human rights, democracy and the rule of law at the same time, or adversely affect one person's human rights while positively affecting another's.

i. AI & Respect for Human Value

13. Respect for human value is reflected by the ECHR in various rights, such as the right to liberty and security (art. 5), the right to a fair trial (art. 6), the right to no punishment without law (art. 7) and the right to a private life and physical and mental integrity (art. 8). AI can impact these rights in the following ways.

Liberty and Security, Fair Trial, No Punishment without Law (art. 5, 6, 7 ECHR)

14. The fact that AI can perpetuate or amplify existing biases, is particularly pertinent when used in law enforcement and the judiciary. In situations where physical freedom or personal security is at stake, such as with predictive policing, recidivism risk determination and sentencing, the right to liberty and security combined with the right to a fair trial are vulnerable**.** When an AI-system is used for recidivism prediction or sentencing it can have biased outcomes. When it is a black box, it becomes impossible for legal professionals, such as judges, lawyers and district attorneys to understand the reasoning behind the outcomes of the system and thus complicate the motivation and appeal of the judgement.

15. Less obvious is the impact of AI on the right to reasonable suspicion and prohibition of arbitrary arrest. AI-applications used for predictive policing merely seek correlations based on shared characteristics with other 'cases'. Suspicion in these instances is not based on actual suspicion of a crime or misdemeanour by the particular suspect, but merely on shared characteristics of the suspect with others (such as address, income, nationality, debts, employment, behaviour, behaviour of friends and family members and so on). Moreover, the actual characteristics used in the AI-system and the 'weights' given to those characteristics are often unknown.

16. If applied responsibly however, certain types or uses of AI can however also improve security, for example AI applications that can 'age' missing people in pictures to improve chances of finding them or AI-driven object recognition that can scan luggage at an airport for suspected contents.

Private and Family Life, Physical, Psychological and Moral Integrity (art. 8 ECHR)

17. Many AI-systems and uses have a broad and deep impact on the right to privacy. Privacy discussions around AI currently tend to focus primarily on data privacy and the indiscriminate processing of personal (and non-personal) data. It should however be noted

that, while data privacy is indeed an important element, the impact of AI on our privacy goes well beyond our data. Art. 8 of the ECHR encompasses the protection of a wide range of elements of our private lives, that can be grouped into three broad categories namely: (i) a person's (general) privacy, (ii) a person's physical, psychological or moral integrity and (iii) a person's identity and autonomy.[26] Different applications and uses of AI can have an impact on these categories, and have received little attention to date.

18. AI-driven (mass) surveillance, for example with facial recognition, involves the capture, storage and processing of personal (biometric) data (our faces)[27], but it also affects our 'general' privacy, identity and autonomy in such a way that it creates a situation where we are (constantly) being watched, followed and identified. As a psychological 'chilling' effect, people might feel inclined to adapt their behaviour to a certain norm, which shifts the balance of power between the state or private organisation using facial recognition and the individual.[28] In legal doctrine and precedent the chilling effect of surveillance can constitute a violation of the private space, which is necessary for personal development and democratic deliberation.[29] Even if our faces are immediately deleted after capturing, the technology still intrudes our psychological integrity.

19. And while for facial recognition the impact on our 'general' right to privacy and our psychological integrity might be more obvious, one could argue that the indiscriminate on and offline tracking of all aspects of our lives (through our online behaviour, our location data, our IoT data from smart watches, health trackers, smart speakers, thermostats, cars, etc.), could have the same impact on our right to privacy, including our psychological integrity.

20. Other forms of AI-driven biometric recognition have an even greater impact on our psychological integrity. Recognition of micro-expressions, gate, (tone of) voice, heart rate, temperature, etc. are currently being used to assess or even predict our behaviour, mental state and emotions.

21. It should be noted upfront that no sound scientific evidence exists corroborating that a person's inner emotions or mental state can be accurately 'read' from a person's face, gate, heart rate, tone of voice or temperature, let alone that future behaviour could be predicted by it. In a recent meta-study, a group of scientists[30] concluded that AI-driven emotion recognition could, at the most, recognize how a person subjectively *interprets* a certain biometric feature of another person. An interpretation does not align with how that person actually feels, and AI is just labelling that interpretation which is highly dependent on context and culture. Far-fetched statements, that AI could for example determine whether someone will be successful in a job based on micro-expressions or tone of voice, are simply without scientific basis.

---

[26] Guidance to art. 8 ECHR, Council of Europe.

[27] The jurisprudence of the European Court of Human Rights (ECtHR) makes clear that the capture, storage and processing of such information, even only briefly, impacts art. 8 ECHR.

[28] Examined Lives: Informational Privacy and the Subject as Object, Julie E. Cohen, 2000.

[29] The chilling effect describes the inhibition or discouragement of the legitimate exercise of a right. It has been shown that once people know that they are being surveilled they start to behave and develop differently.
Staben, J. (2016). Der Abschreckungseffekt auf die Grundrechtsausübung: Strukturen eines verfassungsrechtlichen Arguments. Mohr Siebeck.

[30] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest, 20(1), 1–68.

22. More importantly, the widespread use of these kinds of AI techniques, for example in recruitment, law enforcement, schools, impacts a person's physical, psychological or moral integrity and thus elements of that person's private life.

23. It should be noted that GDPR restricts the processing of biometric data only to some extent. Biometric data according to the GDPR is "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person. The last part of the sentence is crucial, because if biometric recognition is not aimed at identification (but for example at categorization, profiling or affect recognition), it might not fall under the GDPR-definition. In fact, recital 51 of the GDPR says that 'the processing of photographs [is considered] biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person.'

24. Many biometric recognition technologies are not aimed at processing biometric data to uniquely identify a person, but merely to assess a person's behaviour (for example in class) or to categorize individuals (for example for the purpose of determining their insurance premium based on their statistical prevalence to health problems). These uses might not fall under the definition of biometric data (processing) under the GDPR.

25. Going back to data privacy, personal and non-personal data is not only being used to train AI systems, but also to profile and score people for various purposes such as predictive policing, insurance acceptance, social benefits allowance, performance prediction in hiring and firing processes. Moreover, massive amounts of 'data points' on how we go about our daily lives are used not only to send us targeted advertising, but also to push/influence/induce/nudge us towards certain information and thus influence our options, affecting our moral integrity.

ii. AI & Freedom of the Individual

26. Freedom of the individual is reflected by the ECHR in various rights, such as freedom of expression (art. 10) and freedom of assembly and association (art. 11). AI can have a 'chilling' effect on these freedoms as well.

Freedom of Expression (art. 10 ECHR)

27. Art. 10 of the ECHR provides the right to freedom of expression and information, including the freedom to hold opinions, and to receive information and ideas. AI being used to profile, survey, track and identify people and screen, define, sort and influence or nudge behaviour not only has a potential impact on the right to moral integrity as described above, it can also have a chilling effect on these particular freedoms.

28. Using facial recognition in public areas may interfere with a person's freedom of opinion and expression, simply because of the fact that the protection of 'group anonymity' no longer exists, if everyone in the group could potentially be recognized. This could lead to those individuals changing their behaviour for example by no longer partaking in peaceful demonstrations.[31]

[31] Privacy Impact Assessment Report for the Utilization of Facial Recognition Technologies to Identify Subjects in the Field, 30 June 2011, p. 18.

29. The same goes for the situation where all our data is used for AI-enabled scoring, assessment and performance (e.g. to receive credit, a mortgage, a loan, a job, a promotion, etc.). People might become more hesitant to openly express a certain opinion, read certain books or newspapers online or watch certain online media.

30. With regards to the right to receive and impart information and ideas, AI used in media and news curation, bringing ever more 'personalized' online content and news to individuals, raises concerns. Search engines, video recommendation systems and news aggregators often are opaque, both where it comes to the data they use to select or prioritize the content, but also where it comes to the purpose of the specific selection or prioritization.[32] Many business models are based on online advertising revenue. In order to have people spend as much time on a platform or website as possible, they might be selecting and prioritizing content that will do only that: keep people on their platform, irrespective of whether this content is objective, factually true, diverse or even relevant.

31. Beyond commercial motives, political or other motives might lead to AI-systems being optimized to select or prioritize particular content in an effort to coerce and influence individuals towards certain points of view, for example during election processes.[33]

32. Moreover, AI is becoming more capable of producing media footage (video, audio, images) resembling real people's appearance and/or voice (also known as 'deep fakes'), enabling the deceptive practices for various purposes.

33. All this can give rise to filter bubbles and proliferation of fake news, disinformation and propaganda, and affects the capacity of individuals to form and develop opinions, receive and impart information and ideas and thus impact our freedom of expression.[34]

Freedom of Assembly and Association (art. 11 ECHR)

34. The internet and social media have shown to be helpful tools for people to exercise their right to peaceful assembly and association. At the same time however, the use of AI could also jeopardize these rights, when people or groups of people are automatically tracked and identified and perhaps even 'excluded' from demonstrations or protests.[35]

35. As already mentioned, the use of facial recognition in public areas in particular might discourage people to attend demonstrations and join in peaceful assembly, which is one of the most important elements of a democratic society. Examples of this were already seen in Hong Kong when protesters started wearing masks and using lasers to avoid being 'caught' by facial recognition cameras.

[32] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512.
[33] Cambridge Analytica, Netflix Documentary: The Great Hack
[34] UN Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/73/348
[35] Algorithms and Human Rights, Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, Council of Europe, 2018

iii. AI & Equality, Non-discrimination and Solidarity

Prohibition of Discrimination (art. 14 ECHR, Protocol 12)

36. One of the most reported impacts of AI on human rights is the impact on the prohibition of discrimination and the right to equal treatment. As noted earlier, in many cases, AI has shown to perpetuate and amplify and possibly enshrine discriminatory or otherwise unacceptable biases. Also, AI can enlarge the group of impacted people, when it groups them based on shared characteristics.[36] Moreover, these data-driven systems obscure the existence of biases, marginalising the social control mechanisms that govern human behaviour.

37. As an example, Amazon's recruitment AI favoured men over women, because it was trained on profiles of successful Amazon employees, which happened to be men. The AI-system did not simply filter out women, it looked at characteristics of successful employees such as typical wordings and phrasing and filtered out CV's that did not show these characteristics.[37]

38. Going back to the workings of present-day AI, where the systems merely look for correlations based on shared characteristics with other 'cases', all kinds of unacceptable biases can easily surface. The problem with these systems is that, even if they would excel at identifying patterns, e.g. typical phrases used by successful employees, the system has no understanding of the meaning of the phrases, let alone that it will be able to understand the meaning of success, or even grasp what an employee is. It will only be able to provide a label to a specific pattern.

39. Contrary to popular belief, not all biases are the result of low-quality data. The design of any artefact is in itself an accumulation of biased choices, ranging from the inputs considered to the goals set to optimize for; does the system optimize for pure efficiency, or does it take the effect on workers and the environment into account? Is the goal of the system to find as many potential fraudsters as possible, or does it avoid flagging innocent people? All these choices are in one way or another driven by the inherent biases of the person(s) making them. In short, suggesting that we can remove all biases in (or even with) AI is wishful thinking.[38]

iv. AI & Social and Economic Rights AI in and around the Workplace

40. AI can have major benefits when used for hazardous, heavy, exhausting, dirty, unpleasant, repetitive or boring work. AI systems are however also increasingly being used to monitor and track workers, distribute work without human intervention and assess and predict worker potential and performance in hiring and firing situations. These applications of AI could jeopardize the right to just conditions of work, safe and healthy working conditions, dignity at work as well as the right to organize (art. 2 and 3 ESC). If workers are constantly monitored by their employers, they might become more hesitant to organize (art. 5). AI-systems that assess and predict performance of workers could jeopardize the right to equal opportunities and equal

[36] Affinity Profiling and Discrimination by Association in Online Behavioural Advertising, Wachter 2020

[37] Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.

[38] First Analysis of the EU Whitepaper on AI, Virginia Dignum, Catelijne Muller, Andreas Theodorou, 2020.

treatment in matters of employment and occupation without discrimination on the grounds of sex (art. 20 ESC), especially when these systems enshrine biases within the data or of their creator.

41. There is a risk of loss of necessary skills when more and more work and decisions that were previously performed or taken by humans are taken over by AI-systems. This could not only lead to a less skilled workforce, it also raises the risk of systemic failure, where only a few humans are capable of working with AI-systems and reacting to events where these systems fail.

42. While it is unknown if, and if so how many jobs will be lost or gained as a result of AI, in the disruptive transformation period**,** a mismatch between vulnerable labour forces and required skills could lead to technological unemployment.[39]

v. AI & Democracy

43. AI can have (and likely already has) an adverse impact on democracy, in particular where it comes to: (i) social and political discourse, access to information and voter influence, (ii) inequality and segregation and (iii) systemic failure or disruption.

Social and political and social discourse, access to information and voter influence

44. Well-functioning democracies require a well-informed citizenry, an open social and political discourse and absence of opaque voter influence.

45. This requires a well-informed citizenry. In information societies citizens can only select to consume a small amount of all the available information. Search engines, social media feeds, recommender systems and many news sites employ AI to determine which content is created and shown to users (information personalization). If done well, this could help citizens to better navigate the flood of available information and improve their democratic competences, for instance by allowing them to access resources in other languages through translation tools.[40] However, if AI determines which information is shown and consumed, what issues are suppressed in the flood of online information and which are virally amplified, this also brings risks of bias and unequal representation of opinions and voices.

AI-driven information personalisation is enabled by the constant monitoring and profiling of every individual. Driven by commercial or political motives this technologically-enabled informational infrastructure of our societies could amplify hyper-partisan content one is likely to agree with and provide an unprecedented powerful tool for individualised influence.[41] As a consequence it may undermine the shared understanding, mutual respect and social cohesion required for democracy to thrive. If personal AI predictions become very powerful and effective, they may even threaten to undermine the human agency and autonomy required for

[39] Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. W. W. Norton & Company.
[40] Schroeder, R. (2018). Social Theory after the Internet. UCL Press.
[41] Wu, T. (2016). The attention merchants: The epic scramble to get inside our heads (First edition); Zuboff, S. (2019). The age of surveillance capitalism: The fight for the future at the new frontier of power.

meaningful decisions by voters.[42]

46. Thirdly, AI can undermine a fair electoral process. Political campaigns or foreign actors can use (and have been using) personalised advertisements to send different messages to distinct voter groups without public accountability in the agora.[43] However, it should be noted that it remains uncertain exactly how influential micro-targeted advertisement is.[44] AI can also be used to create and spread misinformation and deep fakes, in the form of text, pictures, audio or video. Since these are hard to identify by citizens, journalists or public institutions, misleading and manipulating the public becomes easier and the level of truthfulness and credibility of media and democratic discourse may deteriorate.

Inequality and segregation

47. AI is widely expected to improve the productivity of economies. However, these productivity gains are expected to be distributed unequally with most benefits accruing to the well-off. Similarly, data and design choices, combined with a lack of transparency of black box algorithms have shown to lead to a perpetuating unjust bias against already disadvantaged groups in society, such as women and ethnic minorities.[45] AI could lead to inequality and segregation and thus threaten the necessary level of economic and social **equality** required for a thriving democracy.

Systemic risks

48. AI decisions that previously only humans were able to make, create new challenges for the security and resilience of societal systems. In particular, if decisions that previously were made by many decentralised actors are replaced by few centralised AI-driven systems, the systemic risks increase, where only a failure of few centralised systems is enough to potentially create catastrophic results.

Financial markets illustrate how new systemic risks emerge when different AI agents interact at superhuman speeds, as the rise of financial flash crashes have demonstrated.[46] When critical energy infrastructures, transport systems and hospitals increasingly depend on automated decisions of artificial agents this introduces new vulnerabilities in the form of a single point of failure with widespread effects. Once efficient systems of critical infrastructure are introduced, they are harder to replace or back-up if they break down.[47]

49. A particular danger to international security and peace lies in seeing the development of AI as a competitive race. AI will not only lead to undesirable side effects, but also empower malicious actors ranging from cybercriminals to totalitarian states in their desire to control populations.

[42] Taddeo, M., & Floridi, L. (2018b). How AI can be a force for good. Science, 361(6404), 751–752
[43] Bradshaw, S., & Howard, P. (2019). Social Media and Democracy in Crisis. Oxford University Press.
[44] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61- million-person experiment in social influence and political mobilization. Nature, 489(7415), 295–298.
[45] Eubanks, V. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor (First Edition). St. Martin's Press; O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy (First edition).
[46] Wellman, M. P., & Rajan, U. (2017). Ethical Issues for Autonomous Trading Agents. Minds and Machines, 27(4), 609–624.
[47] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Éigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.

Digital power concentration

50. Many AI-applications are developed and deployed by only a handful of large private actors, sometimes referred to as the Big Five, GAFAM or even Frightful Five.[48] If too much political power is concentrated in a few private hands which prioritise shareholder- value over the common good, this can threaten the authority of democratic states.

vi. AI & Rule of Law

51. Public institutions are held to a higher standard when it comes to their behaviour towards individuals and society, which is reflected in principles such as justification, proportionality and equality. AI can increase the efficiency of institutions, yet on the other it can also erode the procedural legitimacy of and trust in democratic institutions and the authority of the law.

52. Courts, law enforcement and public administrations could become more efficient, yet at the cost of being more opaque and less human agency, autonomy and oversight.[49]

53. Similarly, whereas previously courts were the only ones to determine what counts as illegal hate speech, today mostly private AI systems determine whether speech is taken down by social media platforms.[50] These AI systems de facto compete for authority with judges and the law and In general, AI can contribute to developing judicial systems that operate outside the boundaries and protections of the rule of law.

54. Automated online dispute resolutions provided by private companies are governed by the terms of service rather than the law that do not award consumers the same rights and procedural protections in public courts.[51]

55. The European Commission for the Efficiency of Justice already in 2018 outlined 5 principles for the use of AI in the judiciary in the “European Ethical Charter on the use of AI in the judicial systems and their environment”. The High-Level Expert Group on AI has called for public bodies to be held to the 7 Requirements for Trustworthy AI when developing, procuring or using AI. Similar principles and requirements should be imposed on law enforcement agencies.

[48] Google, Facebook, Microsoft, Apple and Amazon. See: Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180089. Webb, A. (2019). The Big Nine: How the tech titans and their thinking machines could warp humanity.

[49] Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. Philosophy & Technology, 29(3), 245–268.

Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman.

[50] Cohen, J. E. (2019). Between truth and power: The legal constructions of informational capitalism.

[51] Susskind, J. (2018). Future politics: Living together in a world transformed by tech.

56. However, AI can not only threaten the rule of law, it could also strengthen it.[52] If developed and used responsibly, it can empower agencies to identify **corruption** with the state.[53] Similarly, AI can either be used to detect and defend against cyberattacks.[54]

IV. How to address the impact of AI on Human Rights, Democracy and Rule of Law?

57. The impact of AI on human rights, democracy and the rule of law has been receiving more attention lately, most prominently, in the recent Whitepaper on AI of the European Commission. How to address the impact, however, remains uncertain territory to date. This Chapter describes possible strategies that can be followed within the existing framework of human rights, democracy and the rule of law. These strategies are not exhaustive and should help move the discussion towards the next phase.

i. Putting human rights in an AI context

58. Many AI developers, deployers and users (public and private) seem to be unaware of the (potential) impacts of AI on Human Rights. As a first step, an iteration or (re)articulation exercise in which existing Human Rights of the ECHR are 'translated' to an AI context, is very useful and could be done by means of a Framework Convention.

ii. Measures for compliance, accountability and redress

To properly address the impact of AI on existing human rights, democracy and the rule of law, certain existing compliance, accountability and redress mechanisms could be further developed, and new mechanisms could be introduced. What is important however, is that the use of AI is often hidden or unknown, making it difficult or impossible to know whether there is an impact on human rights, democracy and the rule of law in the first place. Measures for compliance, accountability and redress should thus start with the obligation of transparency about the use of AI systems, which can impact human rights, democracy or the rule of law. This includes an AI registry, which then specifies the risk class and required amount of transparency and accountability for a particular application.

59. Compliance could then start with what has recently been described as a new culture of “Human Rights, Democracy and Rule of Law by design”.[55] In such a culture, developers, deployers and uses of AI, from the outset would reflect on how the technology might affect human rights, democracy and the rule of law and adjust the technology or its use accordingly. This could be underpinned by a (legal) obligation to perform an AI Human Rights, Democracy and Rule of Law Impact Assessment.

[52] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 1–10.
[53] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. Knowledge-Based Systems, 128, 139– 152.
[54] Taddeo, M., & Floridi, L. (2018a). Regulate artificial intelligence to avert cyber arms race. Nature, 556(7701), 296– 298.
[55] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence.

60. Such a new culture would need to include the obligation to account for the appropriate structure to be put in place, but also for the outcomes of the AI Human Rights, Democracy and Rule of Law Impact Assessment as well as the design and governance decisions based thereon.

61. Redress in light of AI impact on human rights entails access to justice and effective remedy. As far as access to justice goes, it might be too soon to determine whether this is sufficiently guaranteed when it comes to AI and human rights impact. Only just recently have we seen the first couple of judgements by domestic courts on the (potential) impact of AI on one particular human right, the right to privacy of art. 8 ECHR.[56]

62. More importantly however, access to justice is challenged when many AI-applications are developed and deployed by only a handful of large private actors. These companies dominate both the development of AI as well as the (eco)systems AI operates in and on. While states are obliged to protect individuals and groups against breaches of human rights perpetrated by other actors, appreciation of non-state actors' influence on human rights has steadily grown.[57] As these large tech companies have now become operators that are capable of determining and perhaps altering our social and even democratic structures, the impact of their AI(-use) on human rights becomes more prevalent. In this respect, AI might serve as a good opportunity and think of a structure that would legally oblige private actors to comply with human rights and to grant access to justice if they fail to do so.[58] The basic question is whether to a) accept the private power of AI companies and to make sure they use it responsibly, or to b) challenge it and try to reassert the power of the state.

63. When it comes to an effective remedy, AI is a topic where, as Sheldon also observed, remedies are 'not only about making the victim whole; they express opprobrium to the wrongdoer from the perspective of society as a whole' and thus 'affirm, reinforce, and reify the fundamental values of society'.[59] The European Court of Human Rights has stressed in its *Broniowski* judgment, that international law requires that 'individual and general redress (…) go hand in hand'.[60]

64. To determine an effective remedy in case of a human rights violation as a result of AI, one thus needs to look at **both** individual and general remedies. Moreover, because AI has a myriad of applications, ranging from surveillance and identification, to profiling, nudging and decision making, remedies need to be tailored towards those different applications. Proper remedies should include cessation of unlawful conduct and guarantees of non-repetition, where states could for example be obliged to adopt and implement enforceable legislation to protect human rights from future AI impacts. The obligation to repair the injury or damage caused by the violation, either to an individual or to a community, should exist. For some AI applications just ensuring an effective remedy might not be sufficient to address the human

[56] A Dutch court has considered a law that allowed public institutions to potentially use AI to predict fraud with social benefits in violation of the right to a private life of art. 8 ECHR, the UK High Court in Cardiff accepted that facial recognition affects art. 8 ECHR, as it enables the extraction of "intrinsically private" information, but it considered the use lawful for proportionality reasons. A French Court considered the use of facial recognition in schools in violation with art. 8 ECHR.
[57] Business and Human Rights, A Handbook for Legal Practitioners, Claire Methven O'Brien, Council of Europe
[58] This means going beyond merely referring to the Recommendation CM/Rec(2016)3 on human rights and business of the Committee of Ministers of the Council of Europe (and the UN Guiding Principles on Business and Human Rights)
[59] Dinah Shelton, 'The Right to Reparations for Acts of Torture: What Right, What Remedies?', 17(2) *Torture* 96 (2007), at 96
[60] Broniowski v. Poland, ECHR

rights impact of that application. More far- reaching measures, such as a ban or restrictive use might be necessary (see Chapter IV).

iii. Protecting democracy, democratic structures and the rule of law

65. To prevent systemic failure or disruption due to centralisation of AI-driven decision-making processes in vital structures, distributed decision-making processes, rather than centralised should be implemented to prevent risk of catastrophic failure. These processes should have proper structures of human oversight built in.[61]

66. Human oversight helps ensure that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), or human- in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system.

67. To address the risk of inequality, governments need to actively halt the use of AI applications that increase inequality.

68. Preventing election influence or public manipulation through AI-driven personalised information is not an easy task. Regulations for online campaigning, either for the (social) media platforms or for political parties, could be considered. Obviously, this raises questions regarding the freedom of speech. Keeping humans in/on the loop and in command (see above) could help detect and eliminate undesirable voter influencing.

69. A crucial leverage in ensuring responsible use of AI in public services is public procurement. If the legally binding requirements for **public** procurement are updated to include criteria such as fairness, accountability and transparency in AI this can serve two purposes. On the one hand, it ensures that governments strictly only use systems that are compatible with the rule of law, but also creates economic incentives for the private sector to develop and use systems that comply with the principles of the rule of law. Furthermore, the use of AI in government should be subject to oversight mechanisms, including court orders and ombudspersons for complaints.

[61] Ethics Guidelines for Trustworthy AI, High Level Expert Group on AI, 2019

V. What if current human rights, democracy and the rule of law fail to adequately protect us?

i. Question Zero

70. Due to the invasiveness of some AI-applications or uses, there might be situations in which our current framework of human rights, democracy and the rule of law fails to adequately or timely protect us and where we might need to pause for reflection and find the appropriate answer to what one could consider “question zero”: Do we want to allow this particular AI-system or use and if so, under what conditions? Answering this question should force us to look at the AI-system or use from all perspectives, which could result in several ‘solutions’:

- A particular AI-system or use is put under a moratorium, (temporarily or indefinitely) banned or put under restrictions (''Red Lines'')
- New Human Rights are introduced as safeguards against the 'new' adverse impact of AI
- Existing Human Rights are adapted to allow for responsible development and use of AI
- A particular AI-system or use is made subject to a specific democratic oversight-mechanism
- Private owners of powerful AI-systems are obliged to align their AI development and governance structures with the interests of those affected by the system and society at large, which could include measures to involve relevant parties (such as workers, consumers, clients, citizens, policy makers)

71. First and foremost, 'AI impact' is to be considered both at individual and at societal/collective level whereas AI can impact both the individual as well as larger parts of our collective society. Secondly, context, purpose, severity, scale and likelihood of the impact is important to determine the appropriate and proportionate action. For AI applications that generate unacceptable risks or pose threats of harm or systemic failure that are substantial, a precautionary and principle-based regulatory approach should be adopted. For other AI applications a risk-based approach could be more appropriate.

ii. Red Lines

72. Red lines could be drawn for certain AI-systems or uses that are considered to be too impactful to be left uncontrolled or unregulated or to even be allowed. These AI- applications could give rise to the necessity of a ban, moratorium and/or strong restrictions or conditions for exceptional and/or controlled use:

- Indiscriminate use of facial recognition and other forms of biometric recognition either by state actors or by private actors
- AI-powered mass surveillance (using facial/biometric recognition but also other forms of AI-tracking and/or identification such as through location services, online behaviour, etc.)
- Personal, physical or mental tracking, assessment, profiling, scoring and

nudging through biometric and behaviour recognition

- AI-enabled Social Scoring
- Covert AI systems and deep fakes
- Human-AI interfaces

73. Exceptional use of such technologies, such as for national security purposes or medical treatment or diagnosis, should be evidence based, necessary and proportionate and only be allowed in controlled environments and (if applicable) for limited periods of time.

iii. Some adapted or new human rights

74. In addition to Red Lines-measures, the following adapted or new Human Rights could be considered (non-exhaustive):

- A right to human autonomy, agency and oversight over AI
- A right to transparency/explainability of AI outcomes, including the right to an explanation of how the AI functions, what logic it follows, and how its use affects the interests of the individual concerned, even if the AI-system does not process personal data, in which case there is already a right to such information under GDPR.[62]
- A separate right to physical, psychological and moral Integrity in light of AI-profiling, affect recognition
- A strengthened right to privacy to protect against AI-driven mass surveillance
- Adapting the right to data privacy to protect against indiscriminate, society-wide online tracking of individuals, using personal and non-personal data (which often serves as a proxy for personal identification)

75. Diverging from these rights in exceptional circumstances such as for security purposes should only be allowed under strict conditions and in a proportionate manner.

iv. Future scenarios

76. Extrapolating into the future with a longer time horizon, certain critical long-term concerns can be hypothesized and are being researched, necessitating a risk-based approach in view of possible unknown unknowns and “black swans”. While some consider that Artificial General Intelligence, Artificial Consciousness, Artificial Moral Agents, Super-intelligence can be examples of such long-term concerns (currently non- existent), many others believe these to be unrealistic. Nevertheless, close monitoring of these developments is necessary in order to determine whether ongoing adaptations to our human rights, democracy and rule of law systems are necessary.

[62] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence.

CHAPTER II. AI Ethics Guidelines: European and Global Perspectives[63]

Marcello Ienca[64] and Effy Vayena[65]

I. Executive Summary

In recent years, private companies, research institutions and public- sector organizations have issued principles, guidelines and other soft law instruments for the ethical use of artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. The aim of this report is mapping the relevant corpus of soft law documents and other ethical-legal frameworks developed by governmental and non- governmental organisations globally with a twofold aim.

First, we want to monitor this ever-evolving spectrum of non-mandatory governance instruments.

Second, we want to prospectively assess the impact of AI on ethical principles, human rights, the rule of law and democracy. The report employs an adapted and pre-validated scoping review protocol to provide a comprehensive and up-to-date overview of current soft law efforts. We reviewed a total of 116 documents published inter alia by governmental agencies, non-governmental organisations, academic institutions and private companies.

Our analysis identifies five prominent clusters of ethical principles and assesses their role in the current governance discourse. *Ex negativo*, our analysis reveals existing blind spots and interpretative gaps in the current soft law landscape.

Furthermore, we establish a link between ethical principles and human rights, with special focus on the rights and freedoms enshrined in the European Convention on Human Rights (ECHR) to assess the extent to which the protection of human rights is integral in current non- mandatory governance frameworks.

Finally, we provide empirically-informed policy implications to inform scientists, research institutions, funding agencies, governmental and inter-governmental organisations and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI.

[63] Report prepared for the CAHAI. Strasbourg, 15 June 2020, named CAHAI(2020)07-fin

[64] Chair of Bioethics, Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich.

[65] Chair of Bioethics, Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich.

II. Key findings

- An increasing number of governmental and nongovernmental organisations (incl. private companies and academic organisations) are developing ethics guidelines or other soft law instruments on AI.

- These soft law documents are being primarily developed in Europe, North America and Asia. The global south is currently underrepresented in the landscape of organisations proposing AI ethics guidelines.

- Current AI ethics guidelines tend to agree on some generic principle but they sharply disagree over the details of what should be done in practice. Furthermore, no single ethical principle is common to all of the 116 documents on ethical AI we reviewed.

- We found growing agreement around the following ethical principles: transparency, justice, non-maleficence, responsibility, and privacy. Ethical considerations regarding sustainability, dignity and solidarity appear significantly underrepresented.

- Most guidelines agree that AI should be *transparent* to avoid potential problems. But it is not clear whether transparency should be achieved through publishing source code, the underlying databases or some other means.

- Slightly more than half of reviewed soft law documents explicitly recommend the promotion of human rights —or warn against their violation— when designing, developing and deploying AI systems.

- Regular expressions built from the codes reveal significant variations in theme coverage among documents produced within member countries of the Council of Europe (CoE) compared to documents produced elsewhere. Compared to the rest of the world, soft law documents produced within countries that are members of the Council of Europe appear to emphasize the ethical principles of solidarity, trust and trustworthiness. In contrast, they appear to refer more sporadically to the principles of beneficence and dignity.

- The principles of privacy, justice and fairness showed the least variation across CoE-member countries, CoE-observer countries and the rest of the world, hence the highest degree of cross- geographical and cross-cultural stability.

III. Key policy implications

- Soft law instruments issued by governmental and nongovernmental organisations (incl. private companies and academic organisations) are useful tools to exert practical influence on public decision making over AI and steering the development of AI systems for social good and in abidance of ethical values and legal norms. However, soft law approaches should not be considered substitutive of mandatory governance. Due to conflict of interest, self-regulation efforts by private AI actors are at particular risk of being promoted to bypass or obviate mandatory governance by governmental and intergovernmental authorities.

- In order to ensure inclusiveness, cultural pluralism and fair participation to collective decision making on AI, the development of soft law documents by organisations located in currently underrepresented global regions, especially Africa and South America, should be promoted.

- The convergence of current soft law instruments around five generic ethical principles such as transparency, justice, non-maleficence, responsibility, and privacy reveals five priority areas of oversight and possible intervention by mandatory governance authorities at both the governmental and intergovernmental level.

- In order to be translated into effective governance, these ethical principles should be conceptually clarified. Policy makers have the duty to resolve semantic ambiguities and conflicting characterisations of these principles.

- The sharp disagreement of current soft law documents on the interpretation and practical implementation of these principles indicates that mandatory governance solutions are likely subject to public disagreement, hence require a transparent process of democraticdeliberation.

- Underrepresented ethical considerations such as those regarding sustainability, dignity and solidarity need to be further scrutinized to avoid importing into mandatory governance the same conceptual gaps and normative blind spots of soft law.

- As nearly half of reviewed soft law documents do not explicitly recommend the promotion— or warn against the violation— of human rights when designing, developing and deploying AI systems, greater focus on the human rights implications of AI is urgently needed.

- Member countries of the Council of Europe are well-positioned to steer the international governance of AI towards the promotion of human rights.

IV. Introduction

Artificial Intelligence (AI) is the study and development of computer systems able to perform tasks normally believed to require human intelligence. Typically, computer systems are deemed intelligent (hence called 'intelligent agents' or 'intelligent machines') when they have the ability to perceive their environment and take autonomous actions directed towards successfully achieving a goal. Historically observed, although general reflections on mechanical reasoning have populated the scientific and philosophical literature since ancient times, the field of AI in the narrow sense originated in the 1940s as a consequence of concomitant advances in mathematical logic (e.g. the Church-Turing thesis), information theory, neurobiology and cybernetics. The field of AI encompasses a variety of complex computational approaches that render or mimic cognitive functions such as learning, memory, reasoning, vision, and natural language processing. The most common of these approaches is called machine learning (ML) and involves the development of algorithms that perform tasks in absence of explicit instructions from human operators. Unlike conventional computer programs, ML algorithms build mathematical models based on training data and rely exclusively on inference and pattern identification to make autonomous predictions and decisions[1]. Today, AI is a major catalyzer of technological transformation. At the dawn of the 2020s, AI systems are embedded in an uncountable number of systems and devices regularly used by humans such as mobile phones, social media, cars, airplanes, analytic software, email communication systems, home appliances etc. AI is integral to a broad variety of human activities including (but not restricting to) telecommunication, transportation, manufacturing, healthcare, banking, insurance, law enforcement and the military.

Due to its technological novelty, capacity for autonomous action and general-purposive nature, AI holds potential for transforming human societies at greater pace and in greater magnitude compared to any other technology. The transformative potential of AI has been deemed "revolutionary" by experts[2], with authors referring to AI development as an "ongoing revolution" that "will change almost every line of work"[3]. For this reason, it is paramount and urgent to assess the implications of AI for core principles and values of human life, the future of human societies and the systems of rules that govern those societies, first and foremost democracy and the rule of law[4-6].

In recent years, several governmental and intergovernmental organizations as well as non- state actors have issued principles, guidelines, recommendations, governance frameworks or other soft law instruments for AI. Soft law instruments are normative documents that are not legally binding or enforceable but of persuasive nature which can have practical influence on decision making in a manner that is comparable to that of binding regulations (hard laws). The aim of these instruments is steering the development of AI for social good and in abidance of ethical values and legal norms. However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. Furthermore, due to the rapid proliferation of AI-related soft law documents and the large diversity of their issuers, it is hard to keep track and make sense of this ever-evolving body of non- mandatory governance in a comprehensive and rigorous manner.

This report provides a comprehensive mapping and meta-analysis of the current corpus of principles and guidelines on ethical AI. This analysis will inform scientists, research institutions, funding agencies, governmental and intergovernmental organizations and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI. Furthermore, it will discuss how these ethical principles, moral customs, and recommended social practices can be translated into mandatory governance, especially internationally binding legal instruments.

Particular attention is devoted in this report to examining the nexus between AI governance and human rights and providing a prospective assessment of the impact of AI technology on human

rights and freedoms [7-9]. Human rights are rights inherent to all human beings, regardless of race, sex, nationality, ethnicity, language, religion, or any other status[10]. These rights describe both moral principles and legal norms in municipal and international law to which a person is inherently entitled as a human being. They are, therefore, inalienable, inviolable and universal. They are inalienable as they are not subject to being taken away by anyone; inviolable, as they should not be infringed under any circumstance; universal, as they are applicable everywhere and at every time. Human rights and freedoms are protected by international conventions. In the European space, a fundamental instrument is the European Convention on Human Rights (ECHR) which was drafted in 1950 by the Council of Europe and entered into force on 3rd September 1953. The ECHR enshrines a set of basic rights and freedoms that should be protected, making a legal commitment to abide by standards of behaviour that respect those rights and freedoms[11].

V. Methodology

In February 2020, we conducted a scoping review of the existing corpus of soft law instruments related to AI. Scoping review methods allow to synthesize and map the literature in a certain domain in an exploratory manner, hence are particularly suitable for screening and assessing complex or heterogeneous areas of research. Given the absence of a unified database for soft law instruments, we developed a protocol for discovery and filtering, adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. The protocol, which was pilot-tested and calibrated prior to data collection, consisted of three sequential and iteratively linked phases: screening, eligibility assessment and content analysis. This methodology is designed to provide a formal and evidence-based procedure to map, monitor and iteratively assess the soft law governance efforts in the area of AI.

i. Screening

In the screening phase, we combined retrospective screening of existing repositories with purposive and unstructured web search. First, we screened the following four data repositories and textual sources to retrieve relevant entries related to soft law documents on AI:

- European Union Agency for Fundamental Rights (December 11, 2019), AI Policy Initiatives List.
- Fjeld & Nagy (January 2020), Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, Harvard Berkman Klein Center Research Publication.
- Jobin, Ienca & Vayena (September 2019). The Global Landscape of AI Ethics Guidelines.
- Nature Machine Intelligence.
- Wong & CASBS (June 2019). Fluxus Landscape: An Expansive View of AI Ethics and Governance.

As according to the Arksey and O'Malley framework for scoping reviews[12], structured database search was complemented with unstructured grey literature search to identify soft law instruments that might have eluded. Entries were assessed for eligibility (see 1.2) and, wherever eligibility was confirmed, included manually into the final synthesis and admitted to the second phase.

Finally, we reviewed the list of "top-45 AI companies" compiled in May 2019 by Datamation, a US-based computer science magazine focused on technology analysis. Each of the 45 AI actors ranked in this list was screened independently by accessing their websites and searching for AI ethics or policy statements manually and via keyword search. Finally, unstructured web search was performed to retrieve information that might have remained undetected through our search strategy. Eligible entries were included manually in the final synthesis and admitted to the second phase.

ii. Eligibility Assessment

In the eligibility assessment phase, we screened all retrieved entries to assess their eligibility to be included into the final synthesis. Decisions on eligibility were guided by the inclusion and exclusion criteria listed in Table 1.

Table 1- Eligibility Criteria

Screening	
	- Types: websites, written articles and other documents published online or parts thereof, such as dedicated web pages, blog posts, institutional reports and declarations, as well as references contained within;
Sources included:	- Issuers: private sector for profit organizations (companies, corporations, holdings etc., including private sector alliances); academic and research institutions (universities, professional societies, science foundations etc.); national governmental agencies (ministries, data protection authorities, competition authorities etc.; non-governmental organisations including non-profit organisations and charities.
	- Language: English, German, French, Spanish, Dutch, Italian and Greek (the languages spoken by the researchers).
Sources excluded:	- Types: videos, images and audio/podcasts (except written descriptions), books, journalistic articles, academic articles, syllabi, legislation, official standards, conference summaries;
	- Issuers: intergovernmental and supranational organisations.
	- Language: others than those above.

Eligibility	
Sources included:	- which refer to “artificial intelligence” and/or “AI”, either explicitly in their title or within their description (example: Google: “AI Principles”); or
	- which do not contain the above reference in their title but mention “robot”, “robotics”, “big data”, “machine learning” instead and reference AI or artificial intelligence explicitly as being part of robots and/or robotics; or
	- which do not contain the above reference in their title but are thematically equivalent (by referring to “algorithms”, “predictive analytics”, “cognitive computing”, “machine learning”, “big data”, “deep learning”, “autonomous” or “automated” instead.
	AND
	- which describes a principle, guideline, standard (including “ethics/ethical”, “principles”, “tenets”, “declaration”, “policy”, “guidelines”, “values” etc.), internal strategy (e.g. creation of advisory board) or other type of initiative.
	AND
	- which is expressed in normative or prescriptive language (i.e. with modal verbs or imperatives such as "responsible", "fair", "trust/trustworthy" etc.); or
	- which is principle- or value-based (i.e. indicating a preference and/or a commitment to a certain ethical vision or course of action).
	- which reference actions/visions/commitments/courses of action that apply to the actor enunciating them or other private sector actors.
Sources excluded:	- websites and documents about robotics that do not mention artificial intelligence as being part of robots/robotics; and
	- websites and documents about data or data ethics that do not mention artificial intelligence as being part of data;
	- websites and documents about AI ethics directly aimed at non-private sector actors (e.g. consulting for the public sector)
	- websites and documents about ethics whose primary focus is not AI (e.g. business ethics).

iii. Content Analysis

In the second phase, entries included in the final synthesis were assessed using an expanded version of a previously validated content analysis protocol developed by the authors[13,14]. This protocol involves both a quantitative and a qualitative analysis. At the quantitative level, entries were classified according to instrument type, issuer, and geographic provenience of the issuer. Furthermore, relative frequencies of relevant quantitative data were measured and visually charted. Finally, we performed full-text screening with the assistance of a keyword search plugin to identify documents that made explicit reference to human rights. Documents included in this category made explicit reference to either preserving and promoting human rights or preventing their violation when designing, developing or deploying AI applications. These documents were differentiated from those that did not mention human rights or did so but without making any explicit normative statement about their promotion or non- violation in the context of AI.

At the qualitative level, thematic content analysis was conducted to identify recurrent thematic patterns related to the following domains: (i) ethical principles and values, and (ii) human rights. This thematic content analysis was conducted manually by the researchers with qualitative software assistance (NVivo/MAXQDA for Mac). Emerging thematic patterns were analyzed in-depth, coded, and clustered into pre-defined ethical categories based on the ethical matrix developed by Jobin, Ienca and Vayena (2019).

Given the size of the final synthesis database, this manual thematic analysis was complemented with an automated analysis via natural language processing (NLP). We retrieved the documents and web contents automatically, where possible, using Python wget package and added the rest manually. Next, we built regular expressions[15] from the codes resulting from the qualitative analysis protocol developed by Jobin, Ienca & Vayena[13]. The regular expressions of the codes belonging to the same theme were joined together into one regular expression by ‘or' statements (‘|’). To ensure comprehensiveness and inclusion, the original English codes were translated into the following languages: German, French, Spanish, Italian and Dutch. To determine the theme coverage, we checked for the occurrence of the theme regular expressions in the documents, i.e., we determined the theme to be present in the guideline if the theme's regular expression had at least one match. Finally, we grouped the results by member type and normalised by the total number of guidelines within each group. Variations between the ethical principles and values raised within the 47 Member States of the Council of Europe were compared with, respectively, principles and values raised within Observer States as well as the rest of the world.

iv. Normative ethical and policy analysis

In the fourth and last phase, empirically informed normative ethical and policy analysis was conducted. The aim of this conclusive study component is transferring the preliminary findings of the previous study phases from the descriptive to the normative-prescriptive level. During this phase, we performed three sequential theoretical steps. First, we assessed the results of our content analysis to identify which ethical principles and values are most common and recurrent across the corpus of documents under analysis. As previous research has shown significant interpretative variation within recurrent thematic clusters[13], we complemented the assessment of relative thematic frequencies with a detailed appraisal of their interpretation. This appraisal was instrumental to evaluating which interpretations of the principles are the most effective, hence should be adopted and pursued by global actors. Second, we assessed our review data to identify which principles and values are less frequent or missing in the current landscape of AI ethics guidelines. This second step was instrumental to identifying possible blind spots in international soft law initiatives and, consequently, making normative recommendations on how to overcome these ethical gaps. Third and finally, we advanced normative recommendations on core ethical principles and values that require prioritization in

international AI governance. This conclusive part was instrumental to informing future normative ethical frameworks and delineating a roadmap for international policy on AI, ethics and human rights. To this purpose, we provided a reader- friendly visual summary of the study findings and a toolbox for future monitoring and evaluation (e.g. indicators) at the interface between AI, ethics and human rights.

VI. Findings

Our search identified 116 documents containing soft law documents on AI issued by non-intergovernmental organisations until February 2020. Data reveal a significant increase over time in the number of publications, with 93.9% having been released since 2016. The peak in the number of soft law documents published internationally was reached in 2018 and experienced a non-negligible decrease in the subsequent year (see Figure 1).

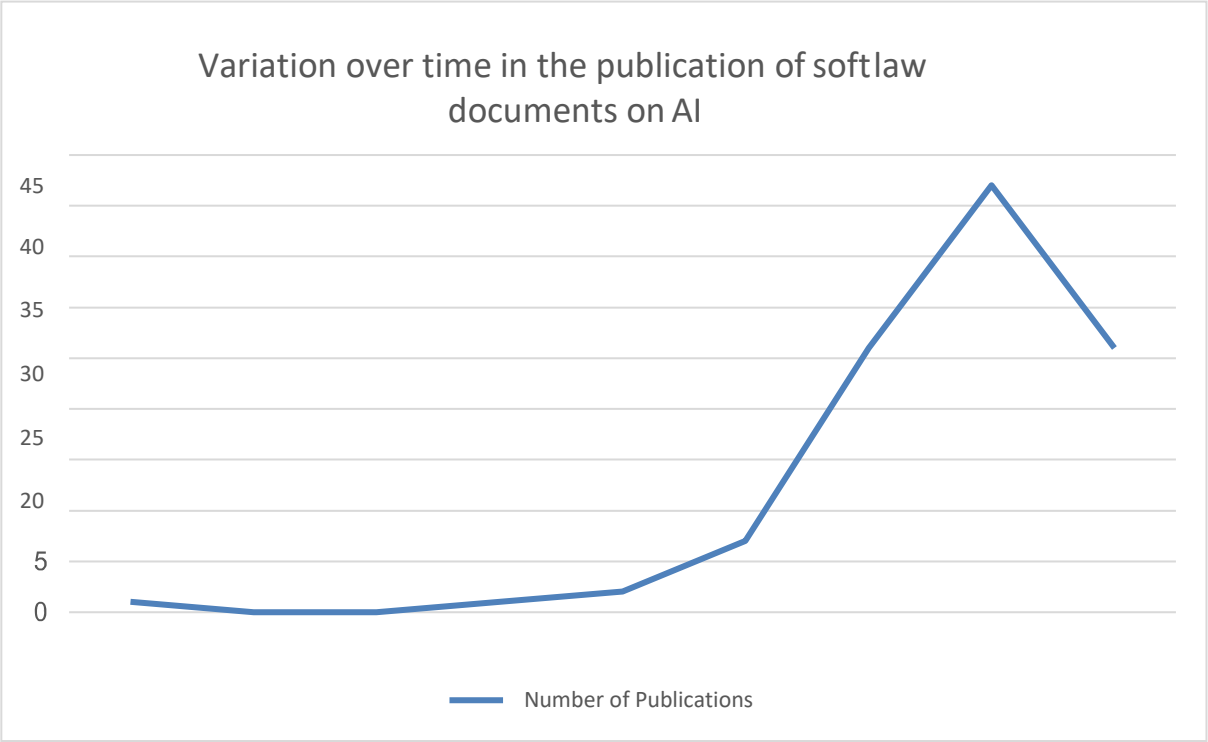


Figure 1- Variation over time in the publication of soft law documents on AI

Data breakdown by type of issuing organization shows that most documents were produced by governmental agencies (n=39), followed by private companies and private sector alliances (n=36), academic and research institutions including science foundations, professional societies and research alliances (n=28) as well as non- governmental organisations (NGOs) including non-profit organisations (NPOs) and charities (n=13). A detailed distribution of issuing organisations by type is provided in Figure 2.

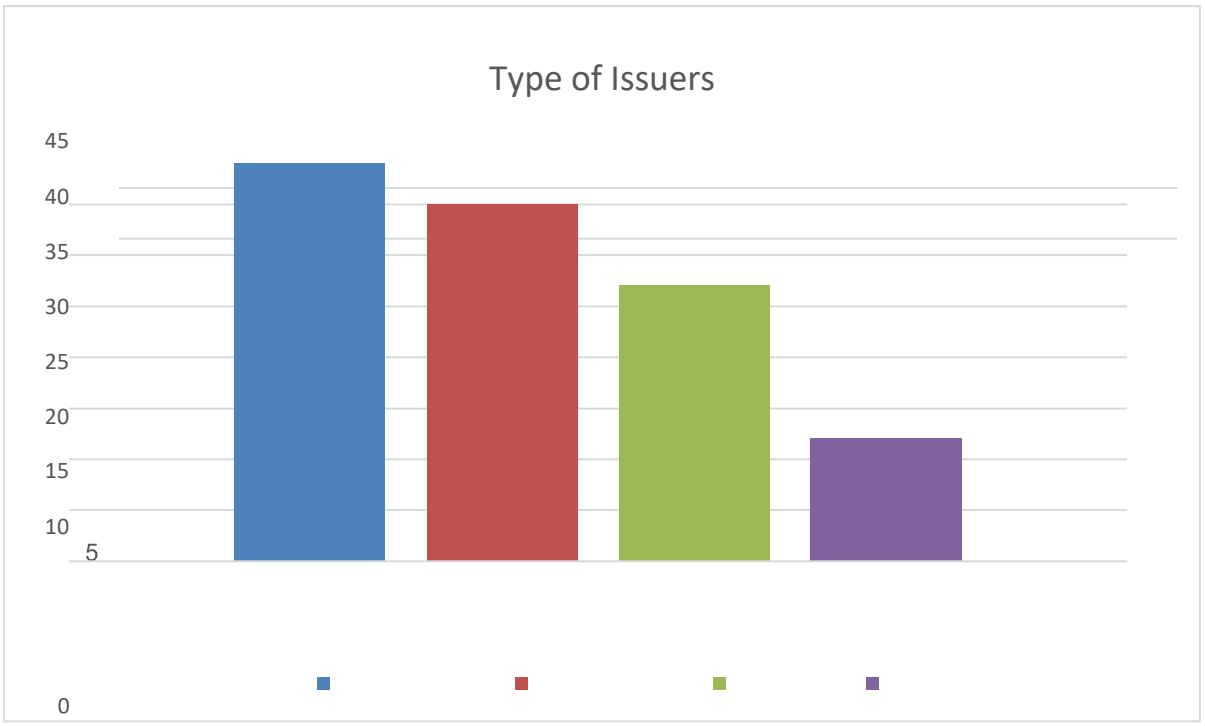


Figure 2- Types of issuing organisations

Data breakdown by geographic distribution of issuing organisations shows that 46% (n=53.5) of soft law documents are issued by organisations based in member countries of the Council of Europe. 32% (n=37.5) by organisations based in observer countries of the Council of Europe. 21% (n=25) by organisations based in countries that are neither members nor observers of the Council of Europe. Overall, data show a prominent representation of issuing organisations based in economically developed countries, with the USA (*n* = 29.5; 25.2%) and the UK (*n* = 17.5; 16%) together accounting for more than one third of all ethical AI principles. Other countries include, in descending order, Germany (n=8), Japan (n=6), Finland (n=4), Belgium, China, France and The Netherlands (n=3), India, Italy, Singapore and Spain (n=2), Australia, Austria, Czech Republic, Iceland, Lithuania, Malta, Mexico, New Zealand, Norway, Russia, South Korea, Sweden, Switzerland, UAE, and the Vatican (n=1). Thirteen documents were issued by international organisations or organisations that could not be ascribed to any specific country. African and South-American countries are not represented independently from international organizations. A visual overview of the geographic distribution of issuing organisations is presented in Figure 3:

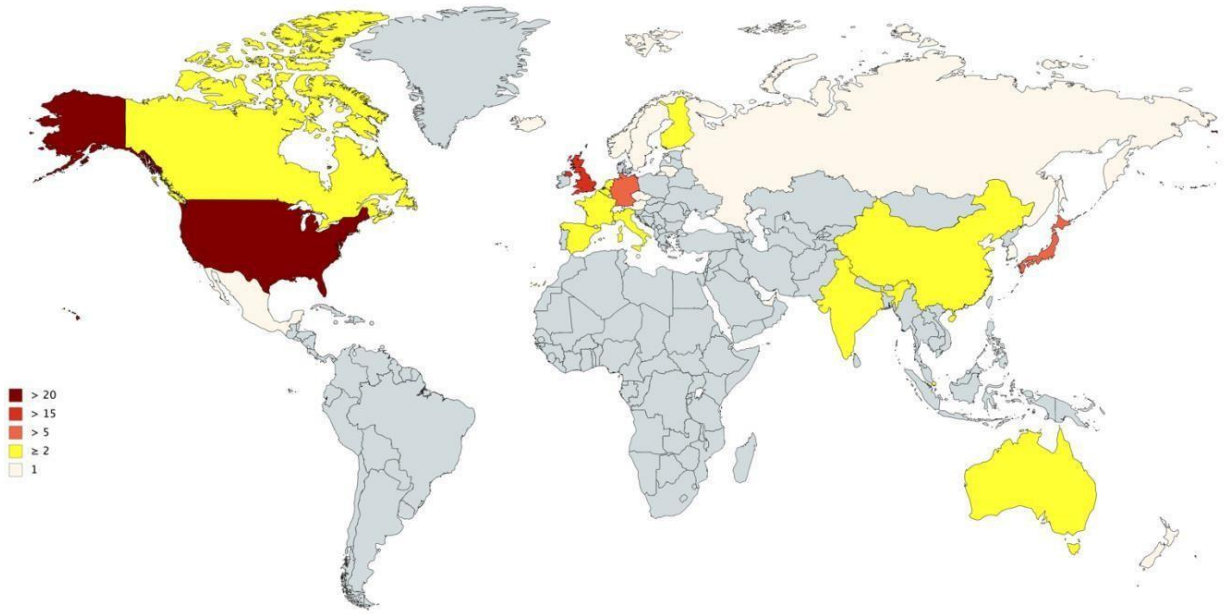


Figure 3- Geographic distribution of soft law documents by country of issuing organisation

More than half of the documents (n=62) make explicit reference to promoting, respecting or preventing the violation of human rights. Of these documents, 31 are issued by organisations based in member countries of the Council of Europe, 14 by organisations based in Observer countries and 17 in non-members non-observer countries. Documents issued by organisations based in member countries of the Council of Europe make reference to human rights in 57.9% of cases. Documents from non-CoE member countries make reference to human rights in 49.6% of cases. This reveals that the human rights implications of Artificial Intelligence are more frequently addressed by organisations based in member countries of the Council of Europe compared to the rest of the world.

Our thematic content analysis retrieved a variety of ethically relevant codes, which could all be consistently allocated to the eleven overarching ethical clusters identified by Jobin, Ienca & Vayena (2019)[13]. These are, by decreasing order of frequency of the sources in which they were featured: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity. A detailed frequency representation of ethical principles and associated codes is presented in Table 2.

Ethical principle	Number of documents	Included codes
		Transparency, explainability,
		explicability, understandability,
Transparency	101/116	interpretability, communication,
		disclosure, showing
		Justice, fairness, consistency,
		inclusion, equality, equity, (non-)bias,
		(non-)discrimination, diversity,
Justice and fairness	97/116	plurality, accessibility, reversibility,
		remedy, redress, challenge, access
		and distribution, impartiality
		Non-maleficence, security, safety,
		harm, protection, precaution,
Non-maleficence	84/116	prevention, integrity (bodily or
		mental), non- subversion
		Responsibility, accountability,
Responsibility	79/116	liability, acting with integrity
		Privacy, personal or private
Privacy	74/116	information, confidentiality
		Benefits, beneficence, well-
Beneficence	58/116	being, peace, social good,
		common good
		Freedom, autonomy, consent,
Freedom and autonomy	48/116	choice, self-determination, liberty,
		empowerment
Trustworthiness	41/116	Trust, trustworthiness
		Sustainability, environment (nature),
Sustainability	20/116	energy, resources (energy)
Dignity	20/116	Dignity
Solidarity	10/116	Solidarity, social security, cohesion

Table 2- Frequency of ethical themes and associated codes

No single ethical principle appears to be common to the entire corpus of documents, although there is an emerging convergence around the following principles: transparency, justice and fairness, non- maleficence, responsibility and privacy. These principles are referenced in nearly two thirds of all the sources. Nonetheless, further thematic analysis reveals the persistence of significant semantic and conceptual divergences in both how the 11 ethical principles are interpreted and the specific recommendations or areas of concern derived from each.

Regular expressions built from the codes reveal significant variations in theme coverage among documents produced within member countries of the Council of Europe (CoE) compared to documents produced elsewhere. Compared to documents produced in CoE observer countries, soft law documents produced within member countries of the Council of Europe appear to emphasize the following ethical principles: transparency, sustainability, freedom and autonomy, trust/trustworthiness and solidarity (see Figure 4). In contrast, they appear to refer more sporadically to the principles of justice, beneficence, and dignity. Compared to documents produced in the rest of the world (non-member non-observer countries), soft law documents produced within member countries of the Council of Europe appear to emphasize the principles of trust/trustworthiness and solidarity while addressing all other principles less frequently. The principles of privacy, justice and fairness showed the least variation, hence the highest degree of cross-geographical and cross-cultural stability.

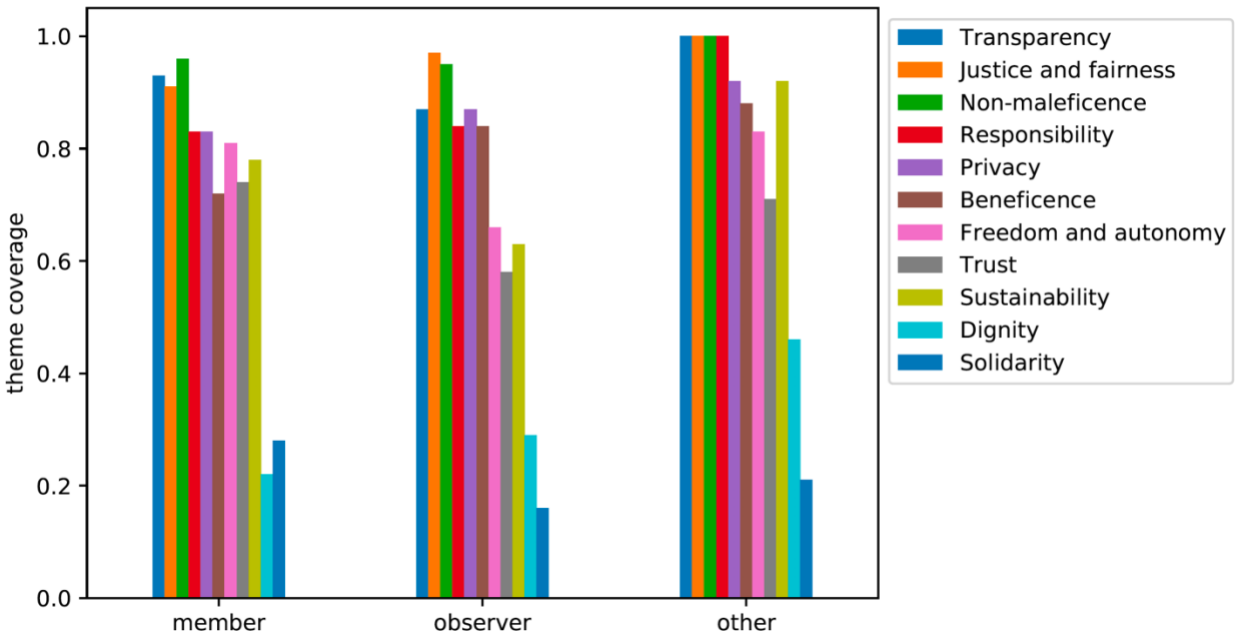


Figure 4- Variations in theme coverage across documents produced within member countries
of the Council of Europe (CoE) vs documents produced in the rest of the world.

A detailed thematic evaluation of the afore listed is presented in the following.

Transparency: Featured in 101 out of 116 sources, transparency is the most prevalent ethical principle in the current soft law spectrum. Thematic analysis reveals significant variation in relation to the interpretation and justification of calls for transparency. This variation is observed to cause obvious divergences in the implementation strategies proposed to achieve transparency in relation to AI. References to transparency can be clustered into two

main thematic families: (1) transparency of algorithms and data processing methods, (2) transparencies of human practices related to the design, development and deployment of AI systems. Calls for transparency of type 1 typically involve the promotion of methodological approaches to “explainable AI”, that is AI systems whose outputs and decisions can be understood by human experts. These methods and techniques contrast with "black box" approaches to machine learning where the steps through which an AI system arrived at a specific decision are unintelligible to human experts including the system’s designers. While private companies, especially private AI actors, tend to reduce transparency to interpretability and explainability through technical solutions —such as, among others, layerwise relevance propagation (LRP) and local interpretability—governmental bodies such as national data protection officers emphasise the importance of oversight methods such as audits. Calls for transparency of type 2 do not focus on interpretable algorithms but on the transparency of human practices related to data and AI such as disclosing relevant information to data subjects, avoiding secrecy when deploying AI strategies and forbidding conflicts of interest between AI actors and oversight bodies. Calls for transparency of this type are more common among governmental actors and NGOs.

Justice, fairness, and equity: Justice is mainly expressed in terms of fairness and prevention (or mitigation) of algorithmic biases that can lead to discrimination. Fears that AI might increase inequality and cause discrimination appear less common in soft law documents issued within the private sector compared to governmental bodies and academia. Documents disagree on how to achieve justice and fairness in AI. Some sources focus on respecting diversity and favouring inclusion and equality both when designing AI systems (especially when compiling the training datasets) and when deploying them in the society. Others sources call for a possibility to appeal or challenge decisions, predicating it on the right to redress and remedy. Fair access to the benefits of AI is also a commonly recurring theme. Documents issued by governmental actors place particular emphasis on AI’s impact on the labour market, and the need to address democratic or societal challenges. We identified five main non- mutually-exclusive implementation strategies for preserving and promoting justice and fairness in AI:

i. Via technical solutions such as standards and best practices;

ii. By raising public awareness of existing rights and regulation;

iii. Via better testing, monitoring and auditing of AI systems;

iv. By developing or strengthening the rule of law and the right to appeal, recourse, redress, or remedy;

v. Via systemic changes and processes such as governmental action and oversight, a more interdisciplinary workforce, as well as better inclusion of civil society or other relevant stakeholders in an interactive manner.

While solutions II-V appeared to be the preferred solution among governmental agencies (especially data protection officers), solutions of type I appeared more common among private AI actors.

Non-maleficence: References to non-maleficence occur significantly more often than references to beneficence and encompass general calls for safety and security or state that AI should never cause foreseeable or unintentional harm. Some documents focus on specific risks or potential harms, especially the risk of intentional misuse via cyberwarfare and malicious hacking. The most common sources of harm mentioned in the documents are social discrimination, privacy violation, and bodily or psychological harm. Soft law documents focused on harm mitigation often call for both technical solutions and mandatory governance interventions at the level of AI research, design, as well as technology development and deployment. Technical solutions include in-built data quality evaluations or security and privacy by design frameworks, though others advocate for establishing industry standards. Proposed governance strategies include active cooperation across disciplines and

stakeholders, compliance with existing or new legislation, and the need to establish oversight processes and practices, notably tests, monitoring, audits and assessments by internal units, customers, users, independent third parties, or governmental entities. Some sources explicitly mention co- optation for military purposes —the so-called dual use problem— as a primary area of AI deployment requiring governance intervention.

Responsibility and accountability: References to developing 'responsible AI' are widespread. Nonetheless, the notions of responsibility and accountability are rarely defined. Diverse actors are named as being responsible and accountable for AI's actions and decisions. These include AI developers, designers, and the entire industry sector. Further disagreement emerged on whether AI should be held accountable in a human-like manner or whether humans should always be the only actors who are ultimately responsible for technological artefacts.

Privacy: Privacy is widely regarded as a value to uphold and a right to be protected. While privacy considerations are frequently addressed in current AI guidelines, there is no consensus on which unique challenges, if any, are raised by advances in AI compared to other data- intensive technologies. Thematic analysis reveals that most documents refer to privacy in general terms, without establishing any explicit nexus between the capabilities of AI and novel privacy challenges. Although poorly characterized, the privacy problem of AI is often presented in association with issues of data protection and data security. Proposed strategies to preserve privacy in AI can be clustered into three categories: (A) technical solutions such as differential privacy, secure multiparty computation and homomorphic encryption; (B) public engagement solutions such as raising awareness among users and data subjects, and (C) regulatory approaches solutions such as better defining the requirements for legal compliance (especially data protection regulation) or even creating new laws and regulations to accommodate the unique of AI.

Beneficence: While promoting good (*beneficence* in ethical terms) is often mentioned, it is rarely defined, though notable exceptions mention promoting human well-being and flourishing, peace and happiness, creating socio-economic opportunities and favouring economic prosperity. Similar uncertainty concerns the actors that should benefit from AI: private sector issuers tend to highlight the benefit of AI for customers, while academic and governmental sources typically argue that AI should benefit 'everyone', 'humanity' and 'society at large'. Strategies for the promotion of good include aligning AI with human values, minimizing power concentration and using AI capabilities for the promotion of human rights.

Freedom and autonomy: Soft law documents link AI to the preservation or promotion of several freedoms and liberties. These notably include freedom of expression, informational self- determination, the right to privacy and personal autonomy. This latter notion is generally referred to as a positive freedom, specifically the freedom to flourish, to decide for oneself and to self-determine one's own course of action. A minority of documents, however, refer to autonomy as a negative freedom, such as a freedom from technological experimentation, manipulation or surveillance. Proposed solutions to preserve freedom and autonomy in AI include pursuing transparent and explainable AI, raising AI literacy, ensuring informed consent or, conversely, actively refraining from collecting and spreading data in absence of informed consent.

Trust and trustworthiness: Slightly more than one in three soft law documents call for trustworthy AI research and technology or for the promotion of a culture of trust among scientists and engineers. Some documents, however, explicitly warning against excessive trust in AI, arguing that trust can only occur among peers and should not be delegated to AI. Suggestions for building or sustaining trust include education, reliability, accountability, processes to monitor and evaluate the integrity of AI systems over time and tools and techniques ensuring compliance with norms and standards.

Sustainability: Sustainability is sporadically mentioned, typically in relation to protecting the

environment or even improving the planet's ecosystem and biodiversity. Some documents demand AI systems to process data sustainably and increase their energy efficiency to minimize ecological footprin[4,7]. A smaller portion of document focuses on social sustainability, that is ensuring accountability in relation to potential job losses and expand opportunities for innovation.

Dignity: While dignity remains undefined in existing guidelines, soft law documents specify that it is a prerogative of humans but not of robots. References to dignity are strongly intertwined with the protection and promotion of human rights. It is argued that AI should not diminish or destroy but respect, preserve or even increase human dignity. Dignity is believed to be preserved if it is respected by AI developers in the first place and promoted through new legislation, through governance initiatives, or through government-issued technical and methodological guidelines.

Solidarity: Solidarity is the least recurring ethical theme and it is mostly referenced in relation to the implications of AI for the labour market. Sources call for a stronger social safety net to cope with the long-term implications of AI for human labour. They underline the need for redistributing the benefits of AI in order not to threaten social cohesion[6,5] and respecting potentially vulnerable persons and groups. Lastly, there is a warning of data collection and processing practices focused on individuals which may undermine solidarity in favour of 'radical individualism'.

i. Limitations

This study has several limitations. First, from a bibliographic perspective, guidelines and soft-policy documents are an instance of grey literature, hence not indexed in conventional scholarly databases. Therefore, their retrieval is inevitably less replicable and unbiased compared to systematic database search of peer-reviewed literature. Following best practices for grey literature review, this limitation has been mitigated by developing a discovery and eligibility protocol which was pilot-tested prior to data collection. Although search results from search engines are personalized, the risk of personalization influencing discovery has been mitigated through the broadness of both the keyword search and the inclusion of results. A language bias may have skewed our corpus towards English results. We minimised this limitation by including entries written in the following languages (besides English): German, French, Italian, Spanish and Dutch. Keywords and codes in the afore-listed languages were translated into English and included in the analysis. Our content analysis presents the typical limitations of qualitative analytic methods. Following best practices for content analysis, this limitation has been mitigated by developing an inductive coding strategy which was conducted independently by two reviewers to minimize subjective bias. Finally, given the rapid pace of publication of AI guidance documents, there is a possibility that new policy documents were published after our search was completed. To minimize this risk, continuous monitoring of the literature was conducted in parallel with the data analysis and until 1st March 2020.

ii. Discussion and Normative Ethical Analysis

We found a rapid increase in the number and variety of soft law documents on AI, demonstrating the increasing active involvement of the international community in non-mandatory governance in this technological domain. Organisations issuing AI guidelines, principles and other soft law instruments come from a wide range of sectors. In particular the nearly equivalent proportion of documents issued by the public (i.e. governmental organisations) and the private sector (companies and private sector alliances) indicates that the ethical challenges of AI concern both public entities and private enterprises. However, there is significant divergence in the solutions proposed to meet the ethical challenges of AI, with public actors prioritizing technical solutions such as explainable and interpretable AI over mandatory regulation and in-depth ethical reflection. Further, the relative underrepresentation of geographic areas such as Africa and South America indicates that the international debate over ethical AI may not be happening globally in equal measures. More economically developed countries (MEDCs) are shaping this debate more than others, which raises concerns about neglecting local knowledge, cultural pluralism and global fairness. These findings confirm the uneven geographic representation and distribution of AI ethics actors observed in previous studies[13]. Compared to previous studies, however, our review reveals that novel actors from previously unrepresented countries are now participating in international non- mandatory governance. These include actors from AI superpowers, that is global-leading AI countries such as China, as well as middle income countries from previously unrepresented world regions such as Russia and Mexico.

The proliferation of soft-law efforts can be interpreted as a governance response to advanced research into AI, whose research output and market size have drastically increased in recent years[16]. Our analysis shows the emergence of an apparent cross-stakeholder convergence on promoting the ethical principles of transparency, justice, non-maleficence, responsibility, and privacy. Nonetheless, our thematic analysis reveals substantive divergences in relation to four major factors: (i) how ethical principles are interpreted, (ii) why they are deemed important, (iii) what issue, domain or actors they pertain to, and (iv) how they should be implemented. Furthermore, unclarity remains as to which ethical principles should be prioritized, how conflicts between ethical principles should be resolved, who should enforce ethical oversight on AI and how researchers and institutions can comply with the resulting guidelines. These findings suggest the existence of a gap at the cross-section of principles formulation and their implementation into practice which can hardly be solved through technical expertise or top- down approaches.

Although no single ethical principle is explicitly endorsed by all existing guidelines, transparency, justice and fairness, non-maleficence, responsibility and privacy are each referenced in more than half of all guidelines. This focus could be indicating a developing convergence on ethical AI around these principles in the global policy landscape. In particular, the prevalence of calls for transparency, justice and fairness points to an emerging moral priority to require transparent processes throughout the entire AI continuum (from transparency in the development and design of algorithms to transparent practices for AI use), and to caution the global community against the risk that AI might increase inequality if justice and fairness considerations are not adequately addressed. Both these themes appear to be strongly intertwined with the theme of responsibility, as the promotion of both transparency and justice seems to postulate increased responsibility and accountability on the side of AI makers and deployers.

It has been argued that transparency is not an ethical principle per se, but rather “a proethical condition for enabling or impairing other ethical practices or principles”[17]. This characterization of transparency as a proethical condition for other principle is detectable in IBM’s Supplier’s Declaration of Conformity (SDoC) that helps to provide information about the four key pillars of trustworthy AI. The allegedly pro-ethical nature of transparency might partly explain its higher prevalence compared to other ethical principles. It is notable that

current guidelines place significant value in the promotion of responsibility and accountability, yet few of them emphasize the duty of all stakeholders involved in the development and deployment of AI to act with integrity. This mismatch is probably associated with the observation that existing guidelines fail to establish a full correspondence between principles and actionable requirements, with several principles remaining uncharacterized or disconnected from the requirements necessary for their realization.

As codes related to non-maleficence outnumber those related to beneficence, it appears that, for the current AI community, the moral obligation to preventing harm takes precedence over the promotion of good. This fact can be partly interpreted as an instance of the so-called negativity bias, i.e. a general cognitive bias to give greater weight to negative entities[18,19], a hypothesis emphasized by cognitive psychologist Steven Pinker in a recent in-depth analysis of the Scientific Foresight Unit (STOA) of the European Parliament[20]. This negative characterization of ethical values is further emphasized by the fact that existing guidelines focus primarily on how to preserve privacy, dignity, autonomy and individual freedom *in spite of* advances in AI, while largely neglecting whether these principles could be actively promoted through responsible innovation in AI.

The issue of trust in AI, while being addressed by less than one third of all sources, tackles a critical ethical dilemma in AI governance: determining whether it is morally desirable to foster public trust in AI. While several sources, especially those produced within the private sector, highlight the importance of fostering trust in AI through educational and awareness-raising activities, a smaller number of sources contend that trust in AI may actually diminish scrutiny and undermine some societal obligations of AI producers[21]. This possibility would challenge the dominant view in AI ethics that building public trust in AI is a fundamental requirement for ethical governance[22]. In relation to trust, we observed to additional conceptual challenges. First, conceptual clarity on the meaning and dynamics of trust seems lacking across the current documents. Most sources failed to specify the trustor and the trustee of the trusting relationship they described, hence neglect that “trust" is a relational and highly complex which involves at least two actors, which trust each other to do, or not to do, a certain activity. This relationship is affected by a wide range of framing factors, for example culture, belief systems, contexts, as well as traits of the actors within the trust relationship. These contextual factors seemed to be neglected in the current literature. Most importantly, the trait "trustworthiness" and the relational construct "trust" appeared frequently conflated or used interchangeably by the AI actors we reviewed. This conflation does not only lead to conceptual confusion but may also foster false hopes among AI users and policy makers. Trust and trustworthiness are different concepts, and trustworthiness does not lead per se to a trusting relationship. Further governance work in this area should clarify this crucial conceptual distinction and demand greater clarification about the requirements of a trusting relationship.

The relative thematic underrepresentation of sustainability and solidarity suggests that these topics might be currently flying under the radar of the mainstream ethical discourse on AI. The underrepresentation of sustainability-related principles is particularly problematic in light of the fact that the deployment of AI requires massive computational resources which, in turn, require high energy consumption[23]. The environmental impact of AI, however, does not only involve the negative effects of high-footprint digital infrastructures, but also the possibility of harnessing AI for the benefit of ecosystems and the entire biosphere. This latter point, highlighted in a report by the World Economic Forum though not in the AI guidelines by the same institution, requires wider endorsement to become entrenched in the ethical AI narrative[24]. The ethical principle of solidarity is sparsely referenced, typically in association with the development of inclusive strategies for the prevention of job losses and unfair sharing of burdens. Little attention is devoted to promoting solidarity through the emerging possibility of using AI expertise for solving humanitarian challenges, a mission that is currently being pursued, among others, by intergovernmental organisations such as the United Nations Office for Project Services (UNOPS) or the World Health Organization (WHO) and private

companies such as Microsoft. As the humanitarian cost of anthropogenic climate change is rapidly increasing[25], the principles of sustainability and solidarity appear strictly intertwined though poorly represented compared to other principles.

While numerical data indicate an emerging convergence around the promotion of some ethical principles, in-depth thematic analysis paints a more complicated picture, as there are critical differences in *how* these principles are interpreted as well as what requirements are considered to be necessary for their realization. Results show that different and often conflicting measures are proposed for the practical achievement of ethical AI. For example, the need for ever larger, more diverse datasets to "unbias" AI appears difficult to conciliate with the requirement to give individuals increased control over their data and its use in order to respect their privacy and autonomy. Similar contrasts emerge between the requirement of avoiding harm at all costs and that of balancing risks and benefits. Furthermore, it should be noted that risk-benefit evaluations will lead to different results depending on whose well-being it will be optimized for by which actors. If not resolved, such divergences and tensions may undermine attempts to develop a global agenda for ethical AI.

Despite a general agreement that AI should be ethical, significant divergences emerge within and between guidelines for ethical AI. Furthermore, uncertainty remains regarding how ethical principles and guidelines should be implemented. These challenges have implications for science policy, technology governance and research ethics. At the policy level, they urge increased cooperative efforts among governmental organisations to harmonize and prioritize their AI agendas, an effort that can be mediated and facilitated by inter-governmental organisations. While harmonization is desirable, however, it should not come at the costs of obliterating cultural and moral pluralism over AI. Therefore, a fundamental challenge for developing a global agenda for AI is balancing the need for cross-national harmonization over the respect for cultural diversity and moral pluralism. This challenge will require the development of deliberative mechanisms to adjudicate disagreement concerning the values and implications of AI advances among different stakeholders from different global regions. At the level of technology governance, harmonization is typically implemented in terms of standardizations. Efforts in this direction have been made, among others, by the Institute of Electrical and Electronics Engineers (IEEE) through the "Ethically Aligned Designed" initiative[26]. Finally, soft governance mechanisms such as Independent Review Boards (IRBs) will be increasingly required to assess the ethical validity of AI applications in scientific research, especially those in the academic domain. However, AI applications by governments or private corporations will unlikely fall under their oversight, unless significant expansions to the IRBs' purview are made.

Overall, our findings indicate that the international community does not agree on what constitutes ethical AI and what requirements are necessary for its achievement. Nonetheless, signs of convergence are noticeable around the notions of transparency, non-maleficence, responsibility, and privacy. Enriching the current ethical AI discourse through a better appraisal of critical yet underrepresented ethical principles such as human dignity, solidarity and sustainability is likely to result into a better articulated ethical landscape for AI. Furthermore, shifting the focus from principle- formulation to translation into practice is desirable. A global agenda for ethical AI should balance the need for cross-national and cross-domain harmonization over the respect for cultural diversity and moral pluralism. Overall, our review provides a useful starting point for understanding the inherent diversity of current principles and guidelines for ethical AI and outlines the challenges ahead for the global community.

iii. Policy Implications

The plethora of international efforts to produce soft law documents on AI provides valuable proxy information about how humanity will react to the many governance challenges posed by AI. The international community seems to converge on the importance of transparency, non- maleficence, responsibility, and privacy for the development and deployment of ethical AI. However, enriching the current ethical AI discourse through a better appraisal of critical yet underrepresented ethical principles such as human dignity, solidarity and sustainability is likely to result into a better articulated ethical landscape for artificial intelligence. Furthermore, shifting the focus from principle-formulation to translation into practice must be the next step. A global agenda for ethical AI should balance the need for cross-national and cross-domain harmonization over the respect for cultural diversity and moral pluralism.

These findings have implications for public policy, technology governance and research ethics. At the policy level, greater intra-stakeholder cooperation is needed to mutually align different AI ethics agendas and seek procedural convergence not only on the ethical principles but also on their implementation. While global consensus might be desirable, it should not come at the costs of obliterating cultural and moral pluralism and might require the development of deliberative mechanisms to adjudicate disagreement among stakeholders from different global regions. Such efforts can be mediated and facilitated by inter-governmental organisations such as the Council of Europe. Furthermore, they could be complemented by bottom-up approaches involving all relevant stakeholders on an equal footing.

Policy interventions in this arena should clarify how AI ethics guidelines relate to existing national and international regulation. In spite of AI's alleged sociotechnical uniqueness, soft law documents on AI do not operate in an ethical-legal vacuum. In contrast, ethics guidelines and other soft law instruments will ultimately have to operate in a context already heavily populated by rules, including hard law (mandatory governance). Failure to consider the context of those rules could undermine the import of the principles into actionable and effective international governance. An example of that is transparency, the most widely recurring ethical principle. In spite of its frequent occurrence, the principle of transparency is typically referred without an explicit link to the underlying binding regulation. Today, institutions that use AI technology are already subject to numerous transparency rules under existing legal systems such as the Fair Credit Reporting Act in the United States and the specific practical requirements on data controllers and processors as outlined in Articles 12-14 of the EU General Data Protection Regulation (GDPR). Similarly, clarifying the distinction between "trust" and "trustworthiness" is a critical task for policy makers.

Besides integrating hard and soft law, an additional challenge is translating ethics principles into practice and seeking harmonization between divergent AI ethics codes. At the level of technology governance, promising attempts to harmonization have been pursued through standardization initiatives such as those led by the Institute of Electrical and Electronics Engineers (IEEE), i.e. the world's largest technical professional organization dedicated to advancing technology innovation. The IEEE is pursuing both AI ethics efforts for general-purpose autonomous and intelligent systems, under the framework of the "Ethically Aligned Designed" initiative[26], as well as domain-specific ones such as the "Neurotechnologies for Brain-Machine Interface Standards Roadmap" developed by the IEEE Standards Association.

Another policy implication regards research oversight. Research ethics mechanisms such as Independent Review Boards (IRBs) will be increasingly required to assess the ethical validity of AI applications in scientific research, especially those in the academic domain. However, AI applications by governments or private corporations will unlikely fall under their oversight, unless significant expansions to the IRBs' purview are made.

Overall, the thematic variety and informational richness of the documents we analysed suggests that soft law instruments issued by governmental and nongovernmental organisations (incl. private companies and academic organisations) are useful tools to exert practical influence on public decision making over AI. If adequately conceptualized, designed and drafted, soft law initiatives hold potential for steering the development of AI systems for social good and in abidance of ethical values and legal norms. However, soft law approaches should not be considered substitutive of mandatory governance. Self-regulation efforts by private AI actors are at particular risk of being promoted to bypass or obviate mandatory governance by governmental and intergovernmental authorities. This risk has been emphasised by the German philosopher Thomas Metzinger, a member of the EU High-Level Expert Group on AI, who observed how a significant portion of the AI ethics discourse is shaped by the private sector[20].

The uneven geographic representation of issuing organisations of AI ethics guidelines requires close monitoring and reflection by international, especially inter-governmental, organisations. In order to ensure inclusiveness, cultural pluralism and fair participation to collective decision making on AI, the development of soft law documents by organisations located in currently underrepresented global regions, especially Africa and South America, should be promoted. Intergovernmental organisations such as the Council of Europe can play a crucial role in the establishment of international platforms of mutual exchange and debate on AI ethics and governance.

The convergence of current soft law instruments around five generic ethical principles such as transparency, justice, non-maleficence, responsibility, and privacy reveals five priority areas of oversight and possible intervention by mandatory governance authorities at both the governmental and intergovernmental level. Prioritizing the realisation of these principles could facilitate the establishment of a core set of norms based on widely agreed ethical precepts. Furthermore, their wide acceptance across both private and public actors is likely to ensure higher degrees of compliance. That being said, in order to be translated into effective governance, these ethical principles should be conceptually clarified. Policy makers have the duty to resolve semantic ambiguities and conflicting characterisations of these principles. The sharp disagreement of current soft law documents on the interpretation and practical implementation of these principles indicates that mandatory governance solutions are likely subject to public disagreement, hence require a transparent process of democratic deliberation.

In parallel, the relative underrepresentation of ethical considerations such as those regarding sustainability, dignity and solidarity needs to be further scrutinized to avoid importing into mandatory governance the same conceptual gaps and normative blind spots of soft law. Mandatory governance should complement and fill the gaps of non-mandatory approaches rather than mirroring the same blind spots of the soft law. To adequately address the sustainability and solidarity challenges of AI, a greater cooperation between environmental protection agencies, ministries of labour and employment as well as ministries of technology and innovation might be required.

As nearly half of reviewed soft law documents do not explicitly recommend the promotion—or warn against the violation— of human rights when designing, developing and deploying AI systems, greater focus on the human rights implications of AI is urgently needed. Member countries of the Council of Europe are well-positioned to steer the international governance of AI towards the promotion of human rights. The human rights implications of AI should be thoroughly investigated at various levels: First, it should be investigated at the level of rights and obligations in the philosophical sense, as they operate independently of legal enactment as justified moral norms. Second, it should be assessed at the level of international human rights law. In this regard, European Convention on Human Rights (ECHR) can pivotal role in international doctrinal research and deliberation on AI. Adherence to the convention is a critical requirement to ensure

the socially responsible development and adoption of a new technology. It is therefore of paramount importance to assess the impact of the sociotechnical transformation induced by AI on the fundamental rights and freedoms postulated in the ECHR.

This impact assessment should have a twofold goal:

(i) evaluating if and how AI will affect or pose new risks for human rights and freedoms;

(ii) (ii) evaluating if and how the responsible development of AI and public deliberation in its regard can contribute to the promotion of those rights and freedoms.

It should be underscored that since technologies are not developed in a vacuum but within a social-historical context of human practices, customs and norms, effective impact assessment strategies should not look at AI in abstraction but contextually to current practices and norms[27].

Finally, it is important to investigate the interface between AI and human rights not only from a high-level perspective, but also and foremost by looking at the human rights salience of specific domains of applications of AI such as *inter alia* robotics[8,28], big data[29,30], autonomous weapons[31,32] and brain-computer interface.

Acknowledgments

The authors would like to thank Dr. Anna Jobin, Karolina Ignatiadis and Manuel Schneider whose work has contributed to the realization of this report.

References

1. Michie D, Spiegelhalter DJ, Taylor C. Machine learning. *Neural and Statistical Classification.* 1994;13.
2. Appenzeller T. The AI revolution in science. *Science.* 2017;357:16-17.
3. Harari YN. Reboot for the AI revolution. *Nature News.* 2017;550(7676):324.
4. Helbing D, Frey BS, Gigerenzer G, et al. Will democracy survive big data and artificial intelligence? In: *Towards Digital Enlightenment.* Springer; 2019:73-98.
5. Livingston S, Risse M. The Future Impact of Artificial Intelligence on Humans and Human Rights. *Ethics & International Affairs.* 2019;33(2):141-158.
6. Nemitz P. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 2018;376(2133):20180089.
7. Ashrafian H. Intelligent robots must uphold human rights. *Nature.* 2015;519(7544):391-391.
8. Van Est R, Gerritsen J, Kool L. Human rights in the robot age: Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality–Expert report written for the Committee on Culture. *Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE)(Rathenau Institute), retrieved from:* https://www/ *rathenau nl/sites/default/files/2018-02/Human% 20Rights% 20in% 20the% 20Robot% 20Age-Rathenau% 20Instituut-2017 pdf (January 5, 2019).* 2017.
9. Raso FA, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L. Artificial Intelligence & Human Rights: Opportunities & Risks. *Berkman Klein Center Research Publication.* 2018(2018-6).
10. Assembly UG. Universal declaration of human rights. *UN General Assembly.* 1948;302(2).
11. Mowbray A. The European Convention on Human Rights. In: *International Human Rights Law.* Routledge; 2016:287-304.

12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International journal of social research methodology.* 2005;8(1):19-32.
13. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence.* 2019;1(9):389-399.
14. Ienca M, Ferretti A, Hurst S, Puhan M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. *PloS one.* 2018;13(10):e0204937.
15. Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish H. Regular expression learning for information extraction. Paper presented at: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing2008.
16. Shoham Y, Perrault R, Brynjolfsson E, et al. The AI Index 2018 annual report. *AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA.* 2018.
17. Turilli M, Floridi L. The ethics of information transparency. *Ethics and Information Technology.* 2009;11(2):105-112.
18. Rozin P, Royzman EB. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review.* 2001;5(4):296-320.Ito TA, Larsen JT, Smith NK, Cacioppo JT. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of personality and social psychology.* 1998;75(4):887.
19. Peter J. Bentley, Miles Brundage, Olle Häggström, Thomas Metzinger. Should we fear artificial intelligence? *European Parliamentary Research Service.* 2018:Scientific Foresight Unit (STOA).
20. Bryson J. No one should trust artificial intelligence. *Science & Technology: Innovation, Governance, Technology.* 2018;11:14.
21. Winfield AF, Jirotka M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 2018;376(2133):20180085.
22. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:190602243.* 2019.
23. Forum WE. Harnessing Artificial Intelligence for the Earth.
24. Scheffran J, Brzoska M, Kominek J, Link PM, Schilling J. Climate change and violent conflict. *Science.* 2012;336(6083):869-871.
25. IEEE. Ethically aligned design. *IEEE Standards v1.* 2016(Global Initiative).
26. Rasmussen T. *Social theory and communication technology.* Routledge; 2019.
27. Liu H-Y, Zawieska K. From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. *Ethics and Information Technology.* 2017:1-13.
28. Mantelero A. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review.* 2018;34(4):754-772.
29. Vayena E, Tasioulas J. 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.* 2016;374(2083):20160129.
30. Heyns C. Human rights and the use of autonomous weapons systems (AWS) during domestic law enforcement. *Hum Rts Q.* 2016;38:350.
31. Asaro P. On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross.* 2012;94(886):687-709.
32. Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy.* 2017;13(1):5.

CHAPTER III. Analysis of international legally binding instruments. Final report[66]

Alessandro Mantelero[67]

I. Executive Summary

The latest wave of Artificial Intelligence (AI) development is having a growing transformative impact on society and raises new questions in different fields, from predictive medicine to media content moderation, from the quantified self to judicial systems, without overlooking the issues of environmental impact.

An analysis of the international legally binding instruments is thus the obligatory starting point to define the existing legal framework, identify its guiding values and verify whether this framework and its principles properly address all the issues raised by AI.

With a view to preserving the harmonisation of the existing legal framework in the field of human rights, democracy and the rule of law, this study aims to contribute to the drafting of future AI regulation by building on the existing binding instruments, contextualising their principles and providing key regulatory guidelines for **a future legal framework**.

The theoretical basis of this approach relies on the assumption that the general principles provided by international human rights instruments should underpin all human activities, including AI-based innovation. Moreover, only the human rights framework can provide **a universal reference for AI regulation**, while other realms (e.g. ethics) do not have the same global dimension, are more context-dependent and characterised by a variety of theoretical approaches.

The analysis of the existing binding legal instruments contained in this document is not limited to a harmonising study, extracting common values and principles from a given set of rules on AI. A more articulated investigation is carried out in different stages.

After an initial sector-specific analysis to map and identify key guiding principles in four core areas (data protection, health, democracy and justice), these principles are contextualised in the light of the changes to society produced by AI. In so doing, we benefit from the existing non-binding instruments that provide more granular applications of the principles enshrined in international legal instruments, in some cases also providing specific guidance on AI.

This **contextualisation of the guiding principles and legal values** provides a more refined and elaborate formulation of them, considering the specific nature of AI products and services, and helps better address the challenges arising from AI. This makes it possible to formulate **an initial set of provisions for future AI regulation** focusing on the most challenging issues in each sector examined.

Considering the large number of documents adopted by several international and intergovernmental bodies and given the parallel ongoing study on ethical instruments carried

[66] Report prepared for the CAHAI, Strasbourg, 15 June 2020, named CAHAI(2020)08-fin

[67] Associate professor of Private Law and Data Ethics & Data Protection, Polytechnic University of Turin (Politecnico di Torino). The opinions expressed in this analysis do not necessarily reflect the position of the CAHAI or the Council of Europe.

out by CAHAI, this document focuses on the legally binding instruments, plus the non-binding instruments adopted to implement them.

The study is divided into two parts. The first one identifies the scope and methodology of this analysis, while the second presents the results of the sectoral analysis on guiding principles.

In the sector-specific analysis, the first two key areas examined are **health** and **data protection**. The intersection between these two realms is interesting in view of this study's focus, given the large number of AI applications concerning healthcare data and the common ground between the two fields. This is reflected in several provisions

of the Oviedo Convention and Convention 108+, as well as by the non-binding instruments. Moreover, individual self-determination plays a central role in both the field of data protection and biomedicine, and the challenges of AI – in terms of the complexity and opacity of treatments and processing operations – are therefore particularly relevant and share common concerns.

The fourth and the fifth sections are centred on **democracy** and **justice**. Here the field of investigation is wider and there are no comprehensive legal instruments that can provide specific sectoral principles, such as Convention 108+ or the Oviedo Convention. The analysis is therefore more closely focused on high-level principles and their contextualisation with a more limited elaboration of key guiding provisions compared with the previous sections.

The last section provides an overview of the guiding principles identified and suggests a harmonisation framework pointing out the existing correlations and common ground between these principles and, at the same time, highlighting the unique contributions of each sector to future AI regulation.

The main objective of this study is not to add a new list of guiding principles to those already provided by a variety of bodies and entities, but to achieve a different result in methodological and substantive terms.

First, **the analysis carried out and the solution proposed have their roots and build on human rights and freedoms**, adopting a concrete approach centred on existing international legal instruments. Other studies are often sector-specific and have a different set of normative references (national or regional) or adopt a theoretical approach enunciating principles or referring to human rights in a general and abstract manner. Although these works do enhance the legal and ethical debate on AI, their impact in terms of contribution to the regulatory framework is limited and not specifically contextualised in the framework of the Council of Europe's standards on human rights, democracy and the rule of law.

Second, the result of this analysis of the legally binding instruments, including the non-binding instruments adopted to implement them, is not merely a list of principles however accurate that may be. **Identifying common guiding principles is important but not sufficient to provide a roadmap for future AI regulation**. Transparency, accountability, human oversight and many other principles already listed in several charters on AI are abstract concepts without a proper contextualisation.

The main contribution of this study is to furnish precisely this **contextualisationwith regard to the legal framework and to AI challenges**. If this document succeeds in suggesting **concrete and effective ways to formulate and codify these guiding principles with regard to AI** and concretely **embed the Council of Europe's standards on human rights, democracy and the rule of law in the outline of the future AI regulation**, it will have achieved its goal in helping to frame the relationship between humans and AI from a legal standpoint.

II. Scope and Methodology

Just as with the Internet, electricity and steam power, Artificial Intelligence (AI) comprise a range of different technologies having a broad impact on a variety of human activities and society.

In this context, many different legal instruments can assume importance in regulating AI applications. At the same time, these legal instruments were adopted in a pre-AI era and this might reduce their effectiveness in providing an adequate and specific response to the new challenges of AI.

An analysis of the international legally binding instruments is thus the obligatory starting point to define the existing legal framework, identify its guiding values and verify whether this framework and its principles properly address all the issues raised by AI, with the view to preserving the harmonisation of the existing legal framework in the field of human rights, democracy and the rule of law.

This approach does not set out to create a completely new and comprehensive reference framework, as the regulation should focus on what changes AI will bring to society, not on reshaping all areas where AI can be applied.[68] This targeted approach is made possible by building on the existing binding instruments, contextualising their guiding principles and providing key regulatory guidelines for a future legal framework for AI, which can cover areas that are not presently regulated by the existing binding instruments.

In this regard, it is important to highlight the difference between the existing legally binding instruments and other documents, such as soft law instruments or ethical charters on AI. Legally binding instruments pre-existed the current AI spring. They were not drafted with AI in mind and do not provide a specific set of rules for this field, while soft law and ethics documents on AI do provide a specific focus, albeit from different perspectives.

Analysis of the existing binding legal instruments is not therefore limited to a harmonising study (i.e. extracting common values and principles from a given set of rules on AI), but requires a more articulated process, in which harmonisation is just one of several stages. The process can be divided into three separate stages: (i) mapping and identification of key principles, (ii) contextualisation, and (iii) harmonisation.

i. The scenario

The latest wave of AI development is having a growing transformative impact on society and rises new question in different fields, from predictive medicine to media content moderation, from the quantified self to judicial systems, without overlooking the issues of environmental impact.

The rapid evolution of applied AI over the last few years has been incompatible with a specific legal response in terms of international legally binding instruments focused on AI. This is why we have seen the development of two different operating strategies to address these issues: (i) a significant effort in interpreting the existing legal framework in the light of AI related issues (see for example the ongoing debate on the GDPR provisions on transparency and automated decision-making); (ii) the use of non-binding rules to contextualise the principles provided by the existing binding instruments (e.g. T-PD(2019)01 Guidelines on Artificial Intelligence and Data Protection; CEPEJ. 2019. European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment[69]).

[68] See, for example, the EU approach to interstitial regulation of e-commerce.
[69] European Commission for the Efficiency of Justice (CEPEJ). 2018. European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment.

Future regulation of AI should therefore build on these efforts, focusing both on the guiding principles and values deriving from the existing binding instruments and on their related non-binding implementations, which in some cases already contemplate the new AI scenario.

ii. Research focus and methodology

The main aim of this study is to define the key principles for the future regulation of AI through an analysis of the existing legal framework. The methodology is therefore necessarily deductive, extracting these principles from the variety of regulations concerning the fields where AI solutions can potentially be adopted.

The theoretical basis of this approach relies on the assumption that the general principles provided by international human rights instruments should underpin all human activities, including AI-based innovation.[70] Moreover, only the human rights framework can provide a universal reference for AI regulation, while other realms (e.g. ethics) do not have the same global dimension, are more context-dependent and characterised by a variety of theoretical approaches.

Against this background, many questions arise, such as: when should an AI system make a decision? Which criteria should the system apply? Who is accountable for decisions that may negatively affect individuals and society? Around these and many other emerging questions, the existing regulations need to be reconsidered.

To provide a harmonised regulatory framework to address the challenges of AI, common and high-level guidance on the principles and values to be enshrined should be derived from international charters of human rights (e.g. Universal Declaration of Human Rights, International Covenant on Civil and Political Rights, Convention for the Protection of Human Rights and Fundamental Freedoms, Charter of Fundamental Rights of the European Union).

The guiding principles must be considered within the AI-driven transformative scenario, which in many cases will require their adaptation. These principles remain valid, but their operation should be reconsidered in the light of the social and technical changes induced by AI (e.g. freedom of choice in the event of so-called black boxes). This will deliver a more contextualised and granular application of the principles so that they can provide a concrete contribution to the shape of future AI regulation.

To conduct this study, we need to start by defining the main areas of investigation, considering both the potential impacts of AI and the fields of action of the Council of Europe. In this regard four key areas have been selected: data, health, democracy and justice.

iii. Analysis and expected results

The study takes a top-down approach with a view to contributing to the future AI regulatory framework, to be implemented by additional binding and non-binding instruments, rather as happened in the field of biomedicine. The expected result is a set of provisions concerning the investigated areas and key common guiding principles, based on a comprehensive analysis of the entire corpus of the binding instruments, including the non-binding tools adopted.

[70] See also Committee of Ministers. 2020. Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems.

First stage: Mapping and identification of key principles. Guiding principles will be identified in the different investigated areas. The first stage of the analysis is based on the different subjects, as binding instruments are sector-specific and not rights- based. The following two tables provide a first example of this mapping exercise based on a preliminary overview of the data protection and justice realms to identify the guiding principles for future regulation of AI.

Figure 1: Data protection

Binding instruments	Convention 108+ Convention on Cybercrime
Impacted areas	Decision-making systems Group privacy and collective dimension Profiling
Related non-binding instruments	CoE. 2019. Guidelines on the data protection implications of artificial intelligence CoE. 2017. Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data CoE. 2010. Recommendation on the protection of individuals with regard to automatic processing of personal data in the context of profiling [under revision] UNESCO. 2019. Preliminary Study on a Possible Standard-Setting Instrument on the Ethics of Artificial Intelligence OECD. 2019. Recommendation of the Council on Artificial Intelligence 40th International Conference of Data Protection and Privacy Commissioners. 2018
Guiding principles and legal values	Accountability Risk-based approach Precautionary principle Data quality & security Transparency Fairness Contextual approach Role of experts Participation/Inclusiveness Freedom of choice/Autonomy Human control/oversight Awareness Literacy Responsible innovation Cooperation between supervisory authorities

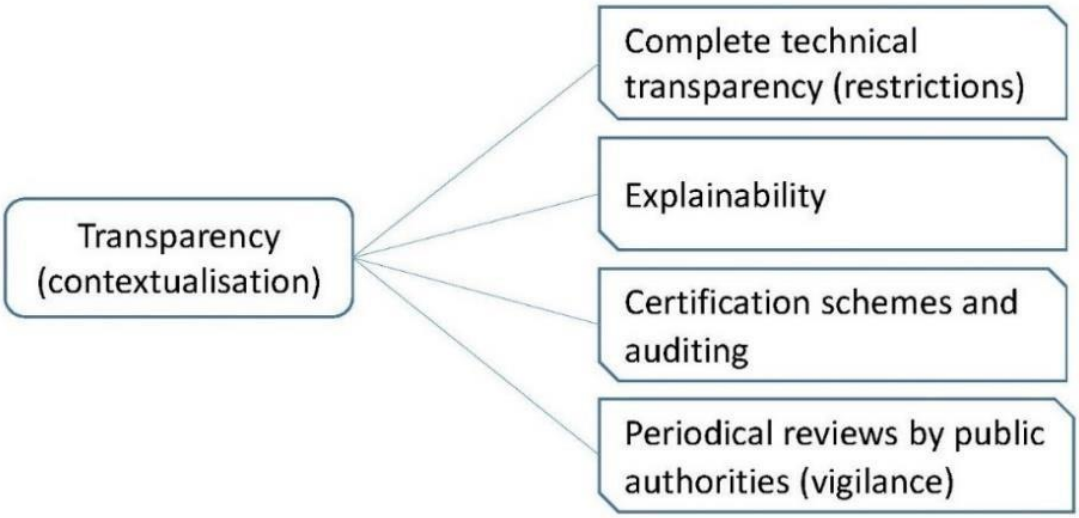
Figure 2: Justice

Binding instruments	Universal Declaration of Human Rights International Covenant on Civil and Political Rights International Convention on the Elimination of All Forms of Racial Discrimination Convention on the Elimination of All Forms of Discrimination against Women Convention for the Protection of Human Rights and Fundamental Freedoms Charter of Fundamental Rights of the European Union
Impacted areas	Processing of judicial decisions and data Predictive policing
Related non- binding instruments	CEPEJ. 2019. European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment
Guiding principles and legal values	Non-discrimination Data quality & security Transparency Impartiality Fairness Contextual approach Freedom of choice/ Independence of judges (decision- making process) Human control/oversight Guarantees of the right to a fair trial

Second stage: Contextualisation. The guiding values identified in the mapping exercise should be contextualised in the light of the changes to society produced by AI. This phase will benefit from the existing non-binding instruments that provide more granular applications of the principles enshrined in the binding instruments, in some case also providing specific guidance on AI.

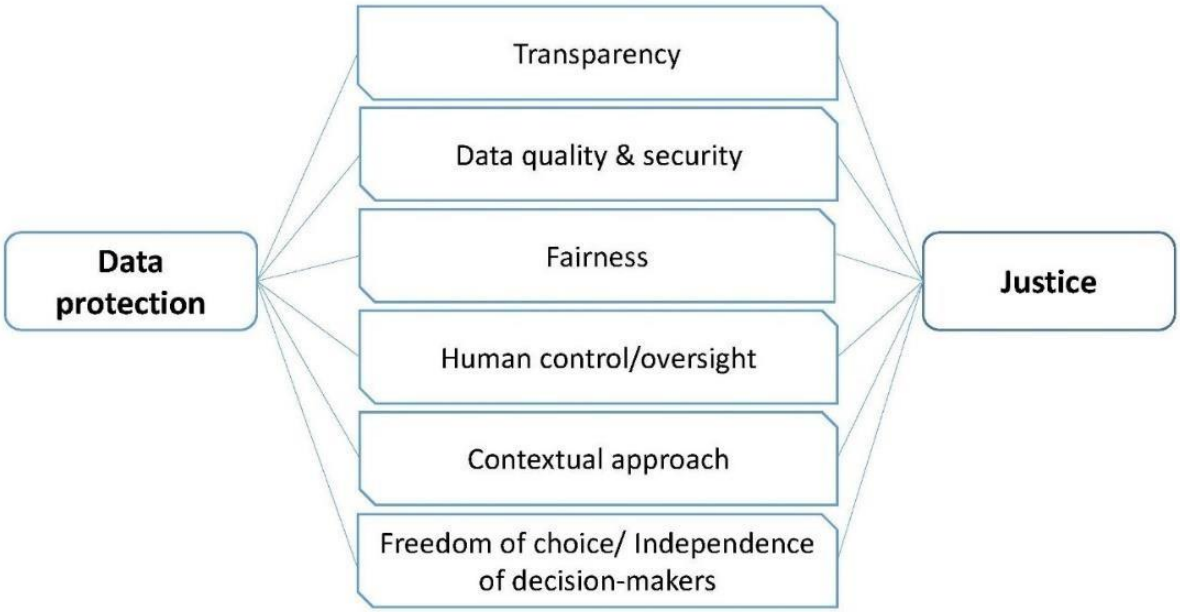
This contextualisation of the guiding principles and legal values will provide a more refined and elaborate formulation of them, considering the specific nature of AI products and services. At this stage, it will therefore be possible to formulate an initial set of provisions for future AI regulation focusing on the most challenging issues in each sector.

Figure 3: Context-specific implementation of the transparency principle



Third stage: Harmonisation (cross-sectoral analysis). Based on the sector- specific analysis carried out in this study, a list of key guiding principles common to the different realms will be drawn up in the last section (Figure 4). These shared principles will then be the cornerstone of the common core of the future provisions on AI.

Figure 4: Common guiding values in the field of data protection and justice



III. Analysis

Considering the large number of documents adopted by several international and intergovernmental bodies and given the parallel ongoing study on ethical instruments carried out by CAHAI, this part focuses on the legally binding instruments, including the non-binding instruments adopted to implement them. Ethical guidelines are therefore not considered at this stage, and documents concerning future regulatory strategies (e.g. white papers) are only taken into account as background information.

This Part is divided into six sections followed by some concluding considerations. The first section presents a general overview of the existing instruments adopted by the Council of Europe and the main underlying principles/values. This helps to define the potential core principles of future AI regulation and its coherence with the existing framework.

The second and third sections focus on two key and related areas: health and data protection. The intersection between these two realms is interesting in view of this study's focus, given the large number of AI applications concerning healthcare data and the common ground between the two fields. This is reflected in several provisions of the Oviedo Convention and Convention 108+, as well as by the non-binding instruments.[71] Moreover, individual self-determination plays a central role in both the field of data protection and biomedicine, and the challenges of AI – in terms of the complexity and opacity of treatments and processing operations – are therefore particularly relevant and share common concerns.

The fourth and the fifth sections are centred on democracy and justice. Here the field of investigation is broader and there are no general legal instruments that can provide sector-specific principles, such as Convention 108+ or the Oviedo Convention. The analysis is therefore focused on high-level principles and their contextualisation, resulting in a more limited elaboration of key guiding provision than in previous sections.

Section 6 provides a general overview of the guiding principles identified and suggests a harmonisation framework that highlights both the existing correlations between these principles and the unique contribution of each sector to future AI regulation.

As highlighted by comments received during the monitoring exercise described in the next section, AI technologies impact on a variety of sectors and raise issues concerning a large body of regulatory instruments.[72] This initial study is therefore a starting point focused on the four core areas mentioned. However, despite its limited scope, the results validate the methodology proposed and provide a number of pointers towards future provisions in AI regulation.

i. General overview

As AI impacts on a variety of situations[73] dealt with by different binding instruments covering several areas, we need to conduct an evidence-based analysis to identify key principles and common values to be considered for future regulation.

An initial monitoring exercise was carried out in this light between 12 and 28 February 2020, involving the different branches of the Council of Europe to benefit from the sector-specific expertise of the various units that have operated over the years in a range of fields relating to human rights, democracy, and the rule of law.

Using a survey based on open-ended questions, the different units interviewed were asked to provide information on the following areas: (i) binding instruments, (ii) impacted areas (applications), (iii) related non-binding instruments, (iv) guiding principles and legal values, and (v) missing principles/issues. Thanks to the positive commitment of the different areas,

[71] See Recommendation CM/Rec(2019)2 on the protection of health-related data.
[72] See Annex 1.
[73] See also UNESCO, 2019.

it was possible to collect a variety of different types of information.

From a methodological point of view, the structure of this preliminary survey based on open-ended questions necessarily affects the results of the quantitative analysis. The main limitations regard the use of different and partially overlapping general categories, as well as differing levels of granularity and specificity of the answers.

Nevertheless, by aggregation in macro-areas and focusing on similarities (i.e. frequency) in the principles and values identified, we were able to achieve some perspective in the results, and the exercise provided a more detailed map of the available non-binding instruments adopted by the Council of Europe that can help to establish a legal framework for future regulation (see Annex 1).

With regard to the impacted areas (see Annex 2), the exercise suggests focusing future AI regulation along two main axes: the use of AI and the development of AI. In both cases, different human rights and fundamental freedoms are potentially affected or can play an important role in shaping future AI scenarios.[74]

Regarding the use of AI, there are four main areas of application and consequent regulation: predictive analysis and decision support systems, automated decision- making systems, evidence collection/computer forensics, and content generation.

The first two areas are well known and debated, as they cover an extremely wide range of applications (see Annex 2). Nevertheless, the distinction between decision support and autonomous decision-making systems is crucial in terms of value oriented-design and the role of human beings in the decision-making process: the differing nature of these two types of systems will necessarily require different procedural and substantive safeguards in AI regulation.

The last two areas are sector-specific but should be considered separately since they do not concern the decision-making process directly but do provide the evidence that underpins it (evidence collection and computer forensics) or affect the creation processes (content generation). In these cases, the main issues seem to be different and more focused on the procedural aspects and their coherence with traditional (i.e. non-AI-based) approaches.

Although most of the existing literature and guidelines focus on AI systems and their potential consequences, an important impact of AI on human rights and fundamental freedoms is also related to the development of AI and the provision of AI services. In this respect, future AI regulation should carefully consider the issues relating to working conditions of the people involved in the whole AI product and service supply chain.[75]

The second block of information provided by the monitoring exercise concerns the guiding principles and legal values that should underpin the future development and use of AI (see Annex 3). Here, the diversity of notions employed by the units surveyed suggests an aggregation of principles and values. The result of this analysis made it possible to group the guiding principles and values around a number of key elements which emerged in terms of distribution (frequency):

Non-discrimination (15)

Diversity, inclusion and pluralism (13) Privacy and Data Protection (11)

Transparency (9)

Equality (8)

Access to justice, fair trial (7) Human control (7)

Impartiality (6)

Access to information (5) Security (5)

[74] See Council of Europe-Committee of experts on internet intermediaries (MSI-NET), 2018.
[75] See also Crawford and Joler, 2018.

Fairness (5)

Participation (5) Freedom of choice (5)

Freedom of expression and of creation (5)

Accountability (3) Competence and capacity (2) Independence (3)

Individual autonomy (3)

Cultural cooperation (2)

Sustainability and Long-term Orientation (2)

Despite the limitations of the analysis mentioned, it is clear that the first three principles are seen as key elements in the future regulation of AI and will therefore be its main focus. This is further confirmed by the second set of principles/values, which is closely related to the first: transparency and human control are important factors in non-discrimination and data protection, while access to justice is a general condition for addressing any potential infringement of these values. Similarly, though more substantively, equality is linked in various ways to the first three main values/principles. The other values/principles, addressing various specific concerns of AI implementation, differ more widely.

This exercise made it possible to identify a first list of guiding principles of AI regulation, already codified in binding and non-binding legal instruments, but in need of contextualisation in the field of AI. In the sector-specific analysis this contextualisation, based on an in-depth analysis of international legally binding instruments, will be achieved by assessing any potential gaps in the existing regulatory framework, sector-by-sector.

As AI is a cross-sector technology, it is expected that the results of this analysis may suggest similar regulatory interventions in other areas, as outlined in the part on methodology.[76] Once the sector-specific analysis is completed, all these potential interventions will be systematised to avoid overlaps and aggregating them into a coherent framework based on key values.

ii. Data Protection

In the past decade, the international regulatory framework in the field of data protection has seen significant renewal. Legal instruments shaped on the basis of principles defined in the 1970s and 1980s[77] no longer responded to the changed socio-technical landscape created by the increasing availability of bandwidth for data transfer, data storage and computational resources (cloud computing), the progressive datafication of large parts of our life and environment (IoT), and large- scale and predictive data analysis based on Big Data and Machine Learning.

In Europe, the main responses to this change have been the modernised version of Convention 108 (Convention 108+) and the GDPR. A similar redefinition of the regulatory framework has been, or is being, carried out in other international contexts
– such as the OECD[78] – or by individual countries.

However, given the rapid development of the last wave of AI development, these new binding instruments fail to directly address some AI-specific challenges and several non-binding instruments have been adopted to bridge this gap, as well as future regulatory strategies under discussion.[79]

For the purposes of this study, the following non-binding legal instruments were therefore

[76] See above Part I.
[77] See also Mayer-Schönberger, 1997; González Fuster, 2014.
[78] See OECD. 2013. Recommendation of the Council concerning Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data, C(80)58/FINAL, as amended on 11 July 2013 by C(2013)79.
[79] See European Commission. 2020. Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, COM(2020) 64 final; European Commission. 2020. White Paper on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final; European Commission. 2020. A European strategy for data, COM(2020) 66 final.

analysed:[80] T-PD(2019)01, Guidelines on Artificial Intelligence and Data Protection [GAI]; T-PD(2017)1, Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data; Recommendation CM/Rec(2019)2 of the Committee of Ministers of the Council of Europe to member States on the protection of health-related data;[81] Recommendation CM/Rec(2010)13 of the Committee of Ministers of the Council of Europe to member States on the protection of individuals with regard to automatic processing of personal data in the context of profiling; UNESCO. 2019. Preliminary Study on a Possible Standard- Setting Instrument on the Ethics of Artificial Intelligence [UNESCO];[82] OECD. 2019. Recommendation of the Council on Artificial Intelligence [OECD]; 40th International Conference of Data Protection and Privacy Commissioners. 2018 [ICDPPC]. Declaration on Ethics and Data Protection in Artificial Intelligence.

These instruments differ in nature: while those adopted by the Council of Europe define different specific requirements and provisions, the others are mainly principles- based, setting out several principles but without, or only partially, providing more detailed guidance in terms of specific requirements. The following paragraphs illustrate the key principles derived from these different instruments and how they can be contextualised within the AI scenario.

Several of these principles classed in the field of personal data protection (e.g. data quality), can be extended to non-personal data, mainly in regard to the impact of the use of non-personal data (e.g. aggregated data) on individual and groups in the context of decision making processes (e.g. mobility data or energy consumption data).

Primacy of the human being

AI systems shall be designed to serve mankind and any creation, development and use of AI systems shall fully respect human rights, democracy and the rule of law.[83]

Human control

AI applications should allow meaningful control by human beings over their effects on individuals and society.[84]

Transparency and expandability

Every individual shall have a right to be informed appropriately when she or he is interacting directly with an AI system, providing adequate and easy-to-understand information on the purpose and effects of this system, including the existence of automated decisions, in order to verify continuous alignment with the expectation of individuals, to enable overall human control on such systems and to enable those adversely affected by an AI system to challenge its outcome.[85]

Every individual shall also have a right to obtain, on request, knowledge of the reasoning underlying an AI-based decision process where the results of such process are applied to him or her.[86] Moreover, States shall promote scientific research on explainable artificial intelligence and best practices for transparency and auditability of AI systems.[87]

[80] See Annex 4.
[81] This Recommendation has replaced Recommendation No. R(97)5 on the protection of medical data. See also Rec(2016)8 on the processing of personal health-related data for insurance purposes, including data resulting from genetic tests and its Explanatory Memorandum.
[82] Despite the reference to ethics in the title, the purpose of the study is described as follows: “This document contains the preliminary study on the technical and legal aspects of the desirability of a standard-setting instrument on the ethics of artificial intelligence and the comments and observations of the Executive Board thereon”.
[83] See CM/Rec(2019)2; ICDPPC; GAI, paras. I.1 and II.1; UNESCO. See also GDPR, Recital no. 4.
[84] See GAI, para. I.6.
[85] See ICDPPC, CM/Rec(2019)2, OECD, UNESCO. See also Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems.
[86] See Convention 108+; GAI, para. II.11.
[87] See ICDPPC.

Precautionary approach

When the potential risks of AI applications are unknown or uncertain, AI development shall be based on the precautionary principle.[88]

Risk management

AI developers, manufacturers and service providers should assess and document the possible adverse consequences of AI applications on human rights and fundamental freedoms, and adopt appropriate risk prevention and mitigation measures from the design phase (human rights by-design approach) and during their entire lifecycle.[89] Adverse consequences include those due to the use of de-contextualised data and de-contextualised algorithmic models.[90]

AI developers, manufacturers, and service providers should consult competent supervisory authorities when AI applications have the potential to significantly impact the human rights and fundamental freedoms of individuals.[91]

Risk of re-identification

Suitable measures should be introduced to guard against any possibility that anonymous and aggregated data may result in the re-identification of the data subjects.[92]

Data quality and minimisation

AI developers should critically assess the quality, nature, origin and amount of personal data used, reducing unnecessary, redundant or marginal data during AI development and training phases, and monitoring the model's accuracy as it is fed with new data. The use of synthetic data may be considered as one possible solution to minimise the amount of personal data processed by AI applications.[93]

Role of experts

AI developers, manufacturers and service providers are encouraged to set up and consult independent committees of experts from a range of fields, as well as engage with independent academic institutions, which can contribute to designing human rights-based, ethically and socially-oriented AI applications, and to detecting potential bias. Such committees may play an especially important role in areas where transparency and stakeholder engagement can be more difficult due to competing interests and rights, such as in the fields of predictive justice, crime prevention and detection.[94]

Appropriate mechanisms should be put in place to ensure the independence of these committees of experts.[95]

[88] See GAI, para. II.2.
[89] 22 See GAI, paras II.2 and II.3; ICDPPC; OECD; UNESCO. See also Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems.
[90] See GAI, para, II.5.
[91] See GAI, para. III.5.
[92] See CM/Rec(2010)13.
[93] See GAI para. II.4; OECD. See also CM/Rec(2020)1.
[94] See also below Section II.3.
[95] See GAI, paras II.6 and II.7; ICDPPC. See also Article11, UNESCO. Declaration on the Human Genome and Human Rights (11 November 1997).

Participation and democratic oversight on AI development

Participatory forms of risk assessment, based on the active engagement of the individuals and groups potentially affected by AI applications, shall be developed. Individuals, groups, and other stakeholders should be informed and actively involved in the debate on what role AI should play in shaping social dynamics, and in decision- making processes affecting them.[96]

Derogations can be introduced for public interest, where proportionate in a democratic society and with adequate safeguards.

Human oversight

AI products and services shall be designed in a manner that ensures the right of individuals not to be subject to a decision significantly affecting them based solely on the automated processing of data, without having their views taken into consideration. AI products and services shall enable overall human control over them.[97]

In addition, the role of human intervention in AI-based decision-making processes and the freedom of human decision makers not to rely on the result of the recommendations provided using AI should be preserved.[98]

Algorithm vigilance

AI developers, manufacturers, and service providers shall adopt forms of algorithm vigilance that promote the accountability of all relevant stakeholders by assessing and documenting the expected impacts on individuals and society in each phase of the AI system lifecycle on a continuous basis, to ensure compliance with human rights, the rule of law and democracy.[99] Governments should provide regular reports about their use of AI in policing, intelligence, and security.[100]

Freedom of choice

In order to enhance users' trust, AI developers, manufacturers and service providers are encouraged to design their products and services in a manner that safeguards users' freedom of choice over the use of AI, by providing feasible alternatives to AI applications.[101]

Right to object

The right to object should be ensured in relation to AI systems based on technologies that influence the opinions and personal development of individuals.[102]

Interoperability

Interoperability between AI systems shall be implemented in full compliance with the principles of lawfulness, necessity and proportionality, putting in place appropriate safeguards for human rights, democracy and the rule of law.[103]

[96] See GAI, paras. II.7 and III.8; ICDPPC. See also CM/Rec(2020)1.
[97] See Convention 108+; GAI para. II.8; ICDPPC; UNESCO.
[98] See GAI para. III. 4.
[99] See GAI para. II.10; OECD; ICDPPC. See also CM/Rec(2020)1.
[100] See UNESCO.
[101] See GAI, para. II.9.
[102] See GAI, para. II.12. See also below Section II.4.
[103] See CM/Rec(2019)2.

Cooperation

Cooperation shall be encouraged between supervisory authorities with competence related to AI.[104]

Digital literacy, education, and professional training

Policy makers should invest resources in digital literacy and education to increase data subjects' awareness and understanding of AI applications and their effects. They should also encourage professional training for AI developers to raise awareness and understanding of the potential effects of AI on individuals and society. They should support research in human rights-oriented AI.[105]

Scientific research integrity

Where a data subject withdraws from a scientific research project, the withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Personal data should be destroyed or anonymised in a manner which does not compromise the scientific validity of the research and the data subject should be informed accordingly.[106]

iii. Health

The European regulatory framework for healthcare is characterised by a few Council of Europe binding instruments and a number of sector-specific instruments adopted at EU level, according to the different nature, scope and regulatory remit of these two entities.

The European Convention on Human Rights, as well as Convention 108+ and the European Social Charter, lay down several general provisions on health protection and related rights. However, these provisions and principles already set out in other general instruments at international level,[107] find a broader and more sector-specific contextualisation in the Oviedo Convention.

The Oviedo Convention – the only multilateral binding instrument entirely focused on biomedicine – and its additional protocols is therefore the main reference point to identify the key principles in this field,[108] which need further elaboration and, where necessary, amplification to regulate AI applications. Furthermore, the Convention is complemented by two non-binding instruments: the Recommendation on health data[109] and the Recommendation on research on biological materials of human origin.[110] The first of these two recommendations illustrates the close link between biomedicine (and healthcare more generally) and data processing, which will be discussed further below.

Most of the existing regulation on health focuses on medical treatment, research (including medical trials) and medical devices/products. AI has a potential impact on all these areas, given its application in precision medicine,[111] diagnosis, and medical devices and services.

[104] See ICDPPC; GAI, para. III.6.
[105] See ICDPPC; OECD; GAI, para. III.9; UNESCO. See also CM/Rec(2020)1.
[106] See Convention 108+; CM/Rec(2019)2.
[107] Some of the general principles enshrined in this convention have been affirmed in previous international human rights instruments, such as the International Covenant on Civil and Political Rights, the International Covenant on Economic, Social and Cultural Rights, and the Convention on the Rights of the Child of 20 November 1989.
[108] See Andorno, 2005; Seatzu, 2015.
[109] See Recommendation CM/Rec(2019)2 on the protection of health-related data
[110] See Recommendation CM/Rec(2016)6 on research on biological materials of human origin.
[111] See Azencott, 2018; Ferryman and Pitcan, 2018.

Although the Oviedo Convention and the related non-binding instruments were adopted in a pre-AI era, they provide specific safeguards regarding self- determination, human genome treatments, and research involving human beings, which are unaffected by AI application in this field and require no changes.

Nevertheless, self-determination in the field of biomedicine faces the same challenges as already discussed in data processing. Notwithstanding the different nature of consent to medical treatments and consent to data processing, the high level of complexity and, often, a certain degree of obscurity of AI applications can undermine the effective exercise of individual autonomy in both cases.[112]

Against this background, the main contribution of the Oviedo Convention to future AI regulation does not concern the sector-specific safeguards it provides, but consists in the important set of general principles and values that can be extrapolated from it to form a building block of future AI regulation.

The Council of Europe's main contribution in the field of medicine concerns the following eight areas: human dignity, primacy of the human being, professional standards, general rule on informed consent, private life and the right to information, non-discrimination, protection of persons undergoing research, and public debate. The contribution of this Convention to the debate on the future regulation of AI goes beyond biomedicine since several provisions, centred on the right balance between technology and human rights, can be extended generally beyond the field of AI, as described in the following paragraphs.[113]

Primacy of the human being

In a geo-political and economic context characterised by competition in AI development, the primacy of the human being should generally be affirmed as a key element of the European approach: better performances of AI-based systems and their efficiency should not override the interests and welfare of human beings. The application of this principle should cover both the development (e.g. systems developed violating human rights and freedoms) and the use of AI systems.[114]

Equitable access to health care

The equitable access principle can be extended to access to the benefits of AI. This entails the adoption of appropriate measures to tackle the risks concerning the digital divide, discrimination, marginalisation of vulnerable persons or cultural minorities, and limitations to the access to information.[115]

Professional standards

AI development therefore embraces several areas of expertise and, where the development of AI systems can impact on individuals and society, it must be carried out in accordance with relevant professional obligations and standards of each area of expertise involved. The professional standards and skills required shall be based on the current state of the art.[116]

States shall encourage professional training to raise awareness and understanding of AI and its potential effects on individuals and society. They should support research in human rights-oriented AI. States shall also cooperate in defining common educational programmes and

[112] See above Section II.2.

[113] Human dignity and informed consent are not included in the table as the first is a value common to the instruments adopted by the Council of Europe in the area of human rights, democracy and the rule of law and informed consent is a principle that is also relevant in the context of data processing.

[114] See also Oviedo Convention, Article 2.

[115] See also Oviedo Convention, Article 3.

[116] See also Oviedo Convention, Article 4; Recommendation CM/Rec(2019)2 on the protection of health-related data.

common standards for professionals who deal with AI and society.

In using AI in the healthcare sector special attention shall be paid to the patient's confidence in his or her doctor and mutual trust, which shall not be compromised by the use of AI.

Protection of persons not able to consent and of persons not able to consent to research

Respect for the principle of beneficence should be considered a requirement where, given the complexity or opacity of AI-based treatments, individual consent suffers from several limitations and cannot be the exclusive basis for treatment.[117]

Private life and right to information

According to Article 10 of the Oviedo Convention, AI health applications shall guarantee the right to information and respect the wishes of individuals not to be informed, except where compliance with an individual's wish not to be informed constitutes a serious risk for the health of others.[118]

Non-discrimination

The principle of non-discrimination in the field of health should be complemented by forbidding any form of discrimination against a person or group based on predictions of future health conditions.[119]

Role of experts

The experience of ethics committees in the field of biomedicine should be considered, introducing multidisciplinary committees of experts in the assessment of AI applications.[120]

Public debate

Fundamental questions raised by the developments of AI shall be subject of appropriate public discussion in the light, in particular, of relevant social, economic, ethical and legal implications, and that their possible application is made the subject of appropriate consultation.[121]

***

These considerations show that the existing legal framework on biomedicine provides important principles and elements that can be extended to future AI regulation, even beyond the health sector.

On the other hand, a series of shortcomings created by the impact of AI remain unresolved in the following areas.

a) **Decision-making systems** [Contextual approach, Fairness, Data quality, Human control/oversight]

[117] See also Oviedo Convention, Articles 6 and 17.
[118] See also Oviedo Convention, Article 10.
[119] See also Oviedo Convention, Article 11.
[120] See also Oviedo Convention, Article 16.
[121] See also Oviedo Convention, Article 28.

In recent years a growing number of AI applications have been developed and used in the medical sector for diagnosis, using both analytics and ML solutions. Large-scale data pools are created, and predictive analytics is used to try and arrive at solutions for clinical cases based on existing knowledge and practices. Likewise, ML applications in image recognition look like they may provide increased cancer detection capability. In addition, in the field of the precision medicine, large-scale collection and analysis of multiple data sources (medical data but also non-medical data, such as air and housing quality) are used to develop individualised insights into health and disease.

The use of clinical data, medical knowledge and practices, as well as non-medical data, is not in itself new in medicine and public health studies. However, the scale of data collection, the granularity of the information gathered, the complexity (and in some case opacity) of data processing, and the predictive nature of the results of analysis raise concerns about the potential weakness of decision-making systems.

Most of these issues are not limited to health sector, as potential biases (including lack of diversity and the exclusion of outliers and smaller populations), data quality, decontextualization, the context-based nature of data labelling and the re-use of data[122] are common to many cases of AI application and concern data in general.[123] In line with the methodology adopted,[124] the existing guidance in the field of data protection[125] can also be applied in this case and the data quality aspects extended to non-personal data.

b) **Self-determination** [Freedom of choice/Autonomy, Awareness]

The opacity of AI applications and the transformative use of data in large-scale data analysis undermine the traditional notion of consent in both data processing[126] and medical treatment, suggesting the adoption of new schemes – such as broad[127] or dynamic consent – which, however, could only contribute in part to solving this problem.

c) The doctor-patient relationship

Several factors concerning AI-based diagnosis – such as the loss of knowledge that cannot be encoded in data,[128] over-reliance on AI in medical decisions, effects of local practices on training datasets, and potential deskilling in the medical sector[129] – may affect the care-patient relationship[130] and should be evaluated when adopting AI in this field.

d) **Risk management** [Risk-based approach, Accountability]

The field of medical devices[131] represents an interesting case study in terms of risk management, considering the significant consequences that the use of these devices can

[122] Ferryman and Pitcan, 2018, 19-20 ("Because disease labels, such as sepsis, are not clear cut, individual labels may be used to describe very different clinical realities" and "these records were not designed for research, but for billing purposes, which could be a source of systematic error and bias").
[123] See above Section II.2.
[124] See e.g. above Figure 4: Common guiding values in the field of data protection and justice.
[125] See Consultative Committee of the Convention of the Protection of Individuals with Regard to Automatic Processing of Personal Data. 2017. Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data. T-PD(2017)1; Consultative Committee of the Convention of the Protection of Individuals with Regard to Automatic Processing of Personal Data.
2019. Guidelines on Artificial Intelligence and Data Protection. T-PD(2019)01. See also the related preliminary studies: Mantelero, A. 2019; Rouvroy, A. 2016.
[126] See also Recommendation CM/Rec(2019)2 on the protection of health-related data.
[127] See also Convention 108+. Explanatory Report, 43 ("In the context of scientific research it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose") and Recommendation CM/Rec(2019)2 on the protection of health-related data, 15.6 ("As it is not always possible to determine beforehand the purposes of different research projects at the time of the collection of data, data subjects should be able to express consent for certain areas of research or certain parts of research projects, to the extent allowed by the intended purpose, with due regard for recognised ethical standards").
[128] See Caruana et al., 2015.
[129] See Cabitza, Rasoini, and Gensini, 2017.
[130] See also WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects, 9th July 2018, https://www.wma.net
[131] See also European Commission, 2014.

have on individuals. The European Union has already adopted a risk-based classification model[132] based on progressive safeguards according to the class of risk of each device (from conformity assessment procedures under the sole responsibility of the manufacturer or the intervention of a notified body, to inspection by a notified body and, in the cases of highest risk, the requirement of prior authorization before being placed on the market).

A model based on such progressive safeguards could be generalised for future AI regulation and also adopted outside the field of medical devices, focusing on the impact on human rights and fundamental freedoms. However, the classification of AI products/services is more difficult, given their variety and different fields of application: several sector-specific classifications should be introduced, or general criteria adopted based on risk assessments procedures.

In addition, specific provisions on AI vigilance and the adoption of the precautionary principle in AI development, as discussed above,[133] can help to address these challenges.

iv. Democracy

Democracy covers an extremely wide array of societal and legal issues,[134] most of them likely to be implemented with the support of ICT[135]. In this scenario, AI can play an important role in the present and future development of digital democracy in a information society.

Compared to the other areas examined (data protection and health), the broad dimension of this topic makes it difficult to identify a single binding sector-specific legal instrument for reference. Several international instruments deal with democracy and its different aspects, starting with the UN Declaration of Human Rights and the International Covenant on Civil and Political Rights. Similarly, in the European context, key principles for democracy are present in several international sources.

Based on Article 25 ICCPR, we can identify two main areas of intervention: (i) participation[136] and good governance, and (ii) elections. Undoubtedly, it is difficult or impossible to draw a red line between these fields as they are interconnected in various ways. AI can have an impact on all of them: participation (e.g. citizens engagement, participation platforms), good governance (e.g. e-government, decision- making processes, smart cities), pre-electoral phase (e.g. financing, targeting and profiling, propaganda), elections (e.g. prediction of election results, e-voting), and the post-election period (e.g. electoral dispute resolution).

As in any classification, this distinction is characterised by a margin of directionality. It is worth pointing here out that this is a functional classification based on different AI impacts, with no intention to provide a legal or political representation of democracy and its different key elements. The relationship between participation, good governance, and elections can therefore be considered from different angles and shaped in different ways, unifying certain areas or further subdividing them.

Participation is expressed both through taking part in the democratic debate and through the electoral process, but the way that AI tools interact with participation in these two cases differs and there are distinct international legal instruments specific to the electoral process.

[132] See Directive 93/42/EEC.
[133] See above Section II.2.
[134] See e.g. Council of Europe. Directorate General of Democracy – European Committee on Democracy and Governance. 2016. The Compendium of the most relevant Council of Europe texts in the area of democracy
[135] See e.g. Directorate General of Democracy and Political Affairs – Directorate of Democratic Institutions. 2009. Project «Good Governance in the Information Society», CM(2009)9 Addendum 3. Indicatives Guides and Glossary relating to Recommendation Rec(2009) 1 of the Committee of Ministers to member states on electronic democracy (e-democracy), prepared by The Council of Europe's Ad hoc Committee on E-Democracy (CAHDE); Additional Protocol to the European Charter of Local Self-Government on the right to participate in the affairs of a local authority, 2009, Article 2.2.iii.
[136] For a more detailed analysis see Faye Jacobsen, 2013. See also Maisley, 2017.

Participation and good governance

The right to participate in public affairs (Article 25 Covenant) is based on a broad concept of "public affairs",[137] which includes public debate and dialogue between citizens and their representatives, with a close link to freedom of expression, assembly and association.[138] In this respect, AI is relevant from two different perspectives: as a means to participation and as the subject of participatory decisions.

Considering AI as a means, technical and educational barriers can undermine the exercise of the right to participate. Participation tools based on AI should therefore consider the risks of under-representation and lack of transparency in participative processes (e.g. platforms for the drafting of bills). At the same time, AI is also the subject of participatory decisions, as they include decisions on the development of AI in general and its use in public affairs.

AI-based participative platforms (e.g. Consul,[139] Citizenlab,[140] Decidim[141]) can make a significant contribution to the democratic process, facilitating citizen interaction, prioritising of objectives, and collaborative approaches in decision-making[142] on topics of general interests at different levels (neighbourhood, municipality, metropolitan area, region, country).[143] As these platforms are used in a social environment and collect information, the same aspects already discussed with regard to data protection, including security, can be recalled here by extending the guidelines discussed in the previous section on data to these applications.

However, other more specific issues arise in relation to AI tools for democratic participation (including those for preventing and fighting corruption[144]), which are associated with the following four main areas: transparency, **accountability**, **inclusiveness**, and **openness**. In this regard, the general principles set out in international binding instruments have an important implementation in the Recommendation CM/Rec(2009)1 of the Committee of Ministers to member states on electronic democracy (e-democracy), which provides a basis for further elaboration of the guiding principles in the field of AI with regard to democracy.

Transparency is a requirement for the use of technological applications for democratic purposes.[145] This principle is common to the fields analysed above, data and healthcare. However, transparency is a context-based notion. While in these fields transparency is closely related to self-determination, here it takes on a broader meaning. In a democratic process, transparency is not only a requirement for citizens' self-determination with respect to a technical tool, but is also a component of the democratic participatory process.[146] Transparency no longer has an individual dimension but assumes a collective dimension as a guarantee of the democratic process.

[137] See UN Office of the High Commissioner for Human Rights. 1996. General Comment No. 25: The right to participate in public affairs, voting rights and the right of equal access to public service (Art. 25). CCPR/C/21/Rev.1/Add.7.
[138] See also UN Office of the High Commissioner for Human Rights. 1981. CESCR General Comment No. 1: Reporting by States Parties, para 5 ("facilitate public scrutiny of government policies with respect to economic, social and cultural rights and to encourage the involvement of the various economic, social and cultural sectors of society in the formulation, implementation and review of the relevant policies").
[139] See <https://consulproject.org/en/>, accessed 29.12.2019.
[140] See <https://www.citizenlab.co/>, accessed 29.12.2019.
[141] See <https://decidim.org/>, accessed 29.12.2019.
[142] See also Council of Europe. Guidelines for civil participation in political decision making. CM(2017)83-final. Adopted by the Committee of Ministers on 27 September 2017 at the 1295th meeting of the Ministers' Deputies.
[143] See also Recommendation CM/Rec(2009)2 on the evaluation, auditing and monitoring of participation and participation policies at local and regional level.
[144] See United Nations Convention against Corruption, 2003, Article 13.
[145] See Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), para 6.
[146] See also Guidelines for civil participation in political decision making. CM(2017)83-final, IV.

In this context, the use of AI-based solutions for e-democracy must be transparent in respect of their logic and functioning (e.g. content selection in participatory platforms) providing clear, easily accessible, intelligible and updated information about the AI tools used.[147]

Moreover, the implementation of this notion of transparency should also consider the range of different users of these tools, adopting an **accessible** approach[148] from the early stages of the design of AI tools. This is to ensure effective transparency with regard to vulnerable and impaired groups, giving added value to accessibility in this context.

Transparency and accessibility are closely related to the nature of the architecture used to build AI systems. **Open source and open standards**[149] can therefore contribute to democratic oversight of the most critical AI applications.[150] There are cases where openness is affected by limitations, due to the nature of the specific AI application (e.g. crime prevention). In these cases, auditability, as well as certification schemes, play a more important role than they already do in relation to AI systems in general.[151]

In the context of AI applications to foster democratic participation, an important role can be also played by **interoperability**[152] as it facilitates integration between different services/platforms for e-democracy and at different geographical levels. This aspect is already relevant for e-democracy in general,[153] and should therefore be extended to the design of AI-based systems.

Another key principle in e-democracy, as in the data and health sectors, is **accountability**. Unlike the previous principles examined, accountability does not take on a different meaning here, and therefore does not seem to require a sector-specific implementation in the context of AI, other than its general application.

Finally, given the role of media in the context of democratic participation and in line with Recommendation CM/Rec(2016)4 of the Committee of Ministers of the Council of Europe,[154] AI applications must not compromise the confidentiality and security of communications and protection of journalistic sources and whistle-blowers.[155]

In addressing the different aspects of developing AI solutions for democratic participation, a first consideration is that a democratic approach is incompatible with a techno-determinist approach. AI solutions to address societal problems should therefore be the result of an inclusive process. Hence, values such as the protection of minorities, pluralism and diversity should be a necessary consideration in the development of these solutions.

From a democratic perspective, the first question we should ask is: do we really need an AI-based solution to a given problem as opposed to other options,[156] considering the potential

[147] See Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), para. 6 ("facilitates and enhances access, accessibility […] by using, where feasible, transparent […] means") and Appendix to Recommendation CM/Rec(2009)1, para. P.57. See also Recommendation CM/Rec(2016)5 on Internet freedom. Appendix, paras 2.1.3 and 3.2.
[148] See also Recommendation CM/Rec(2018)4 on the participation of citizens in local public life, Appendix, para. B.IV.
[149] See also Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), para. 6 and Appendix, para P.54.
[150] See also Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), Appendix, para. G.58.
[151] 84 It is worth to underline that auditing and certification schemes play an important role also in cases of open source AI architecture, as this nature does not imply per se absence of bias or any other shortcomings. See also Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), Appendix, paras P.55 and G.57 ("E-democracy software should either be open source software that can be inspected or, alternatively, be certified by an independent body").
[152] See also Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), Appendix, paras P. 56, G.56, 59 and 60.
[153] See also Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), para. 6.
[154] See Recommendation CM/Rec(2016)4 on the protection of journalism and safety of journalists and other media actors, Appendix, para. 2; Council of Europe, Parliamentary Assembly. 2019. Resolution 2254 (2019)1. Media freedom as a condition for democratic elections.
[155] See also Parliamentary Assembly, Resolution 2300 (2019)1, Improving the protection of whistle- blowers all over Europe; Recommendation CM/Rec(2014)7 on the protection of whistleblowers.

[156] See also Recommendation CM/Rec(2020)1, Appendix, para. 5.7.

impact of AI on rights and freedoms? If the answer to this question is yes, the next step is to examine **value-embedding** in AI development.[157]

The proposed AI solutions must be designed from a human rights-oriented perspective, ensuring full respect for human rights and fundamental freedoms, including the adoption of **assessment tools and procedures** for this purpose.[158] In the case of AI applications with a high impact on human rights and freedoms, such as electoral processes, legal compliance should be **prior assessed**. In addition, AI systems for public tasks should be **auditable** and, where not excluded by competing prevailing interests, audits should be publicly available.

Another important aspect to be considered is the **public-private partnership** that frequently characterises AI services for citizens, weighing which is the best choice between in-house and third-party solutions, including the many different combinations of these two extremes. In this regard, when AI solutions are fully or partially developed by private companies, **transparency of contracts** and clear **rules on access and use of citizens' data** have a critical value in terms of democratic oversight.

Restrictions on access and use of citizens' data are not only relevant from a data protection perspective (principles of data minimisation and purpose limitation) but more generally with regard to the bulk of data generated by a community, which also includes non-personal data and aggregated data. This issue should be considered as a component of democracy in the digital environment, where the **collective dimension** of the digital resources generated by a community should entail forms of citizen control and oversight, as happens for the other resources of a territory/community (e.g. the environment).

The considerations already expressed above on openness as a key element of democratic participation tools should be recalled here, given their impact on the design of AI systems. Furthermore, the design, development and deployment of these systems should also consider the adoption of an environmentally friendly and sustainable strategy.[159]

Finally, it is worth noting that while AI-design is a key component of these systems, design is not neutral. Values can be embedded in technological artefacts,[160] including AI systems. These values can be chosen intentionally and, in the context of e- democracy, this must be based on a democratic process. But values may also be unintentionally embedded into AI solutions, due to the cultural, social and gender composition of AI developer teams. For this reason, **inclusiveness** has an added value here, in terms of inclusion and diversity[161] in AI development.

With regard to good governance,[162] the principles discussed for e-democracy can be repeated here.[163] This is the case with smart cities and sensor-based environmental management, where open, transparent and inclusive decision-making processes play a central role. Similarly, the use of AI to supervise the activities of local authorities,[164] for

[157] See also Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes (Adopted by the Committee of Ministers on 13 February 2019 at the 1337th meeting of the Ministers' Deputies, para. 7.
[158] See Recommendation CM/Rec(2009)1, paras 5 and 6, and Appendix to Recommendation CM/Rec(2009)1, para. G.67. See also above Section II.2 on data and the role of the committees of experts and A Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment' (2018) 34 Computer Law & Security Review 754.
[159] See also Recommendation CM/Rec(2009)1, Appendix, para. P. 58.
[160] See also P-P Verbeek, 2011, 41-65.
[161] See also P-P Verbeek, 2011, 41-65.
[162] See Recommendation CM/Rec(2009)1 on electronic democracy (e-democracy), Appendix, para. P.4 ("[…] good governance, which is the efficient, effective, participatory, transparent and accountable democratic exercise of power in electronic form, and includes informal politics and non-governmental players").
[163] See also Recommendation Rec(2004)15 on electronic governance ("e-governance"); Council of Europe. 2008. The 12 Principles of Good Governance enshrined in the Strategy on Innovation and Good Governance at local level, endorsed by a decision of the Committee of Ministers of the Council of Europe in 2008.
97 See also Privacy International, 2017.
[164] See also Recommendation CM/Rec(2019)3 on supervision of local authorities' activities, Appendix, Guidelines on the improvement of the systems of supervision of local authorities' activities, paras 4 and 9.

auditing and anticorruption purposes,[165] should be based on **openness** (open source software), **transparency** and **auditability**.

More generally, AI can be used in government/citizen interaction to automate citizen' inquiries and information requests.[166] However, in these cases, it is important to guarantee the right to know we are interacting with a machine[167] and to have a human contact point. Moreover, access to public services must not depend on the provision of data that is unnecessary and not proportionate to the purpose.

Special attention should also be paid to the potential use of AI in human-machine interaction to implement nudging strategies.[168] Here, due to the complexity and obscurity of the technical solutions adopted, AI can increase the passive role of citizens and negatively affect the democratic decision-making process. Otherwise, an active approach based on conscious and active participation in community goals should be preferred and better managed by AI participation tools. Where adopted, nudging strategies should still follow an evidence-based approach.

Finally, the use of AI systems in governance tasks raises challenging questions about the relationship between human decision-makers and the role of AI in the decision- making process.[169] These issues are more relevant with regard to the functions that have a high impact on individual rights and freedoms, as in the case of jurisdictional decisions. For this reason, concerns about transparency (including explainability) of AI reasoning and the relationship between the use of AI and the freedom of decision- makers will be analysed in Section 5.

Elections

As in other areas, the impact of AI on electoral processes is broad and concerns the pre-election, election, and post-election phases in different ways. However, an analysis focused on the stages of the electoral process does not adequately highlight the different ways in which AI solutions interact with it.

The influence of AI is therefore better represented by the following distinction: AI for the electoral process (e-voting, predictions of results, and electoral dispute resolution) and AI for electoral campaigns (micro-targeting and profiling, propaganda and fake news). While in the first area AI is mainly a technological improvement of an existing process, in the field of electoral campaigning AI-based profiling and propaganda raise new concerns that are only partially addressed by the existing legal framework. In addition, several documents have emphasised the active role of states in creating an enabling environment for freedom of expression.[170]

As regards the technological implementation of e-democracy (e-voting, prediction of results, and electoral dispute resolution), some of the key principles mentioned with regard to

[165] See also Savaget, Chiarini and Evans, 2019, discussing the Brazilian case of the 'Operação Serenata de Amor' (OSA).

[166] See Mehr. 2017.
[167] See also GAI 2.11.
[168] See, *ex multis*, Sunstein, 2015a; Sunstein, 2015a; Sunstein and Thaler, 2003; Thaler and Sunstein, 2008.
[169] See also Calo and Citron, 2020, Forthcoming.

[170] See Recommendation CM/Rec(2018)1 on media pluralism and transparency of media ownership; Joint Declaration on "Fake News," Disinformation and Propaganda, The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information (3 March 2017). See also Recommendation CM/Rec(2016)5 on Internet freedom, Appendix, paras 1.5, 2.1 and 3; European commission for Democracy trough law (Venice Commission). 2019. Joint Report of the Venice Commission and of the Directorate of Information society and Actions Against Crime of the Directorate General of Human Rights and Rule of Law (DGI) on Digital Technologies and Elections, para. 151.E; OSCE, 2020. See also Bychawska-Siniarska, 2017.

democratic participation are also relevant here. **Accessibility**,[171] **transparency**,[172] **openness**,[173] **risk management and accountability** (including the adoption of certification and auditing procedures)[174] are fundamental elements of the technological solutions adopted in these stages of the electoral process.

As regards AI for campaigning (micro-targeting and profiling, propaganda and fake news), some of the issues raised concern the processing of personal data in general. The principles set out in Convention 108+ can therefore be applied and properly contextualised.[175]

More specific and new responses are needed in the case of propaganda and disinformation.[176] Here the existing binding and non-binding instruments do not set

specific provisions, given the novelty of the disinformation based on new forms of communication, such as social networks, which differ from traditional media[177] and often bypass the professional mediation of the journalists.

However, general principles, such as the **principle of non-interference** by public authorities on media activities to influence elections,[178] can be extended to these new forms of propaganda and disinformation. Considering the use of AI to automate propaganda, future AI regulation should extend the scope of the general principles of non-interference to AI-based systems used to provide false, misleading and harmful information. In addition, to prevent such interference, states[179] and social media providers should adopt a **by-design approach** to increase their resilience to disinformation and propaganda.

Similarly, the obligation to cover election campaigns in a **fair, balanced and impartial** manner[180] should entail obligations for media and social media operators regarding the transparency of the logic of the algorithms used for content selection,[181] ensuring pluralism and diversity of voices,[182] including critical ones.[183]

Moreover, states and intermediaries should promote and facilitate access to tools to detect disinformation and non-human agents, as well as support independent research on the impact of disinformation and projects offering fact-checking services to users.[184]

Given the important role played by advertising in disinformation and propaganda, the criteria used by AI-based solutions for political advertising should be **transparent**,[185] **auditable** and

---

[171] See Recommendation CM/Rec(2017)5 on standards for e-voting, Appendix I, E-voting Standards, paras 1 and 2.
[172] See Recommendation CM/Rec(2017)5, Appendix I, para. 32. See also Council of Europe. Directorate General of democracy and Political Affairs – Directorate of Democratic Institutions. 2011. Guidelines on transparency of e-enabled elections.
[173] See Recommendation CM/Rec(2017)5, Appendix I, para. 35.
[174] See Recommendation CM/Rec(2017)5, Appendix I, paras 36, 37, 38, 39 and 40.
[175] Recommendation CM/Rec(2010)13 on the protection of individuals with regard to automatic processing of personal data in the context of profiling and its ongoing review, see Council of Europe, Consultative Committee of the Convention of the Protection of Individuals with Regard to Automatic Processing of Personal Data. 2019. Profiling and Convention 108+: Suggestions for an update. T- PD(2019)07BISrev.
[176] See Manheim and Kaplan, 2019; European Commission - Networks, Content and Technology- Directorate-General for Communication, 'A Multi-Dimensional Approach to Disinformation Report of the Independent High Level Group on Fake News and Online Disinformation' (2018). See also *Stoll v. Switzerlad* [GC], no.69698/01, § 104.
[177] See also Recommendation CM/Rec(2011)7 on a new notion of media.
[178] See Recommendation CM/Rec(2007)15 on measures concerning media coverage of election campaigns, para. I.1.
[179] See also Joint Declaration on "Fake News," Disinformation and Propaganda, para. 2.c.
[180] See Recommendation CM/Rec(2007)15, para. II.1.
[181] See also Joint Declaration on "Fake News," Disinformation and Propaganda; Recommendation CM/Rec(2018)2 on the roles and responsibilities of internet intermediaries, Appendix, paras. 2.1.3 and
2.3.5 ("Due to the current limited ability of automated means to assess context, intermediaries should carefully assess the human rights impact of automated content management, and should ensure human review where appropriate. They should take into account the risk of an overrestrictive or too lenient approach resulting from inexact algorithmic systems, and the effect these algorithms may have on the services that they provide for public debate").
[182] See also EU Code of Practice on Disinformation, 2018.
[183] See also Recommendation CM/Rec(2016)4, Appendix, para. 15.
[184] See also Joint Declaration on "Fake News," Disinformation and Propaganda, para. 4.e; European commission for Democracy trough law. 2019, para. 151.D.
[185] See also Council of Europe. Parliamentary Assembly. Resolution 2254 (2019)1. Media freedom as a condition for democratic elections, paras 9.2 and 11.1; European commission for Democracy trough law (Venice Commission). 2019. Joint Report of

provide **equal conditions** to all the political parties and candidates.[186]

In addition, intermediaries should review their advertising models to ensure that they do not adversely affect the **diversity of opinions and ideas**.[187]

v. Justice

As in the previous section, the field of justice is a broad domain and analysing the whole spectrum of the consequences of AI on justice and its related effects on democracy would be too ambitious. In line with the scope of this study, this section sets out to describe the main challenges associated with the use of AI and the principles which, based on international legally binding instruments, can contribute to its future regulation.

Justice differs from data protection and health in the absence of specific and dedicated binding instruments, such as Convention 108+ and the Oviedo Convention. This analysis is therefore more centred on the contextualisation of general guiding principles than on specific legal instruments.

This exercise is facilitated by the European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment, adopted by the CEPEJ in 2019, which directly addresses the relationship between justice and AI. Although this non-binding instrument is classed as an ethical charter, to a large extent it concerns legal principles enshrined in international instruments.

Guiding principles for the development of AI in the field of justice can be derived from the following binding instruments: the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, the Convention for the Protection of Human Rights and Fundamental Freedoms, the International Convention on the Elimination of All Forms of Racial Discrimination, the Convention on the Elimination of All Forms of Discrimination against Women, and the Convention for the Protection of Human Rights and Fundamental Freedoms.[188]

Given the range of types and purposes of operations in this field and the various professional figures and procedures involved, this section makes a functional distinction between two areas: (i) judicial decisions and alternative dispute resolutions (ADRs) and (ii) crime prevention/prediction. Before analysing and contextualising the key principles relating to these two areas, we should offer some general observation, which may also apply to the action of the public administration as a whole.[189]

First of all, it is worth noting that – compared to human decisions, and more specifically judicial decisions – the logic behind AI systems does not resemble legal reasoning. Instead they simply execute codes based on a data-centric and mathematical/statistical approach.

In addition, error rates for AI are close to, or lower than, the human brain in fields such as image labelling, but more complicated decision-making tasks have higher error rates. This is the case with legal reasoning in problem solving.[190] At the same time, while a

---

the Venice Commission and of the Directorate of Information society and Actions Against Crime of the Directorate General of Human Rights and Rule of Law (DGI) on Digital Technologies and Elections, paras 151.A and 151.B.

[186] See also Recommendation CM/Rec(2007)15, para. II.5.

[187] See also Joint Declaration on "Fake News," Disinformation and Propaganda, para. 4.e.

[188] See also, with regard to the EU area, the Charter of Fundamental Rights of the European Union.

[189] See above Section II.4.

[190] See Dupont et al., 2018, 148 ("Deep Learning has no natural way to deal with hierarchical structure, which means that all the available variables are considered on the same level, as 'flat' or non- hierarchical. This presents a major hurdle when decisions carry a heavy moral or legal weight that must supersede other features"). See also Osoba and Welser, 2017, 18 ("Another angle on the problem is that judgments in the space of social behavior are often fuzzy, rather than well-defined binary criteria […]. We are able to learn to navigate complex fuzzy relationships, such as governments and laws, often relying on subjective evaluations to do this. Systems that rely on quantified reasoning (such as most artificial agents) can mimic the

misclassification of an image of a cat may have limited adverse effects, an error rate in legal decisions has a high impact on rights and freedom of individuals.

It is worth pointing out that the difference between errors in human and machine decision-making has an important consequence in terms of scale: while human error affects only individual cases, poor design and bias in AI inevitably affect all people in the same or similar circumstances, with AI tools being applied to a whole series of cases. This may cause group discrimination, adversely affecting individuals belonging to different categories.

Given the textual nature of legal documents, natural language processing (NLP) plays an important role in AI applications for the justice sphere. This raises several critical issues surrounding commercial solutions developed with a focus on the English- speaking market, making them less effective in a legal environment that uses languages other than English.[191] Moreover, legal decisions are often characterised by implicit unexpressed reasoning, which may be amenable to expert systems, but not by language-based machine learning tools. Finally, the presence of general clauses requires a prior knowledge of the relevant legal interpretation and continual updates which cannot be derived from text mining.

All these constraints suggest a careful and more critical adoption of AI in the field of justice than in other domains and, with regard to court decisions and ARDs, suggest following a distinction between cases characterised by routinely and fact-based evaluations and cases characterised by a significant margin for legal reasoning and discretion.[192]

Court decisions and ADRs

Several so-called Legal Tech AI products do not have a direct impact on the decision- making processes in courts or alternative dispute resolutions (ADRs), but rather facilitate content and knowledge management, organisational management, and performance measurement.[193] These applications include, for example, tools for contracts categorisation, detection of divergent or incompatible contractual clauses, e-discovery, drafting assistance, law provision retrieval, assisted compliance review. In addition, some applications can provide basic problem-solving functions based on standard questions and standardised situations (e.g. legal chatbots).

Although AI has an impact in such cases on legal practice and legal knowledge that raises various ethical issues,[194] the potential adverse consequences for human rights, democracy and the rule of law are limited. To a large extent, they are related to inefficiencies or flaws of these systems.

In the case of content and knowledge management, including research and document analysis, these flaws can generate incomplete or inaccurate representations of facts or situations, but this affects the meta-products, the results of a research tool that need to be interpreted and adequately motivated when used in court. Liability rules, in the context of product liability, for instance, can address these issues.

In addition, bias (poor case selection, misclassification etc.) affecting standard text- based computer-assisted search tools for the analysis of legislation, case-law and literature, ca be countered by suitable **education and training** of legal professionals and the **transparency** of AI systems (i.e. description of their logic, potential bias and limitations) can reduce the negative consequences.

effect but often require careful design to do so. Capturing this nuance may require more than just computer and data scientists."). See also Cummings et al., 2018, 13

[191] See Council of Bars & Law Societies of Europe, 2020, 29.

[192] See the following Section on the distinction between codified justice and equitable justice.

[193] See European Commission for the Efficiency of Justice (CEPEJ). 2018. European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment, Appendix II.

[194] See also Nunez, 2017.

Transparency should also characterise the use by courts of AI for legal research and document analysis. Judges must be transparent as to which decisions depend on AI and how the results provided by AI are used to contribute to the arguments, in line with the **principles of fair trial and equality of arms**.[195]

Finally, transparency can play an important role with regard to legal chatbots based on AI, making users aware of their logic and the resources used (e.g. list of cases analysed). Full transparency should also include the sources used to train these algorithms and access to the database used to provide answers. Where these databases are private, third party **audits** should be available to assess the quality of datasets and how potential biases have been addressed, including the risk of under- or over-representation of certain categories (**non-discrimination**).

Further critical issues affect AI applications designed to automate alternative dispute resolution or to support judicial decision. Here, the distinction between codified justice and equitable justice[196] suggests that AI should be circumscribed for decision-making purposes to cases characterised by routine and fact-based evaluations. This entails the importance to carry out further research on the classification of the different kind of decisional processes to identify those routinised applications of legal reasoning that can be demanded to AI, preserving in any case human overview that also guarantees legal creativity of decision-makers.[197]

Regarding equitable justice, as the literature points out,[198] its logic is more complicated than the simple outcome of individual cases. Expressed and unexpressed values and considerations, both legal and non-legal, characterise the reasoning of the courts and are not replicable by the logic of AI. ML-based systems are not able to perform a legal reasoning. They extract inferences by identifying patterns in legal datasets, which is not the same as the elaboration of legal reasoning.

Considering the wider context of the social role of courts, jurisprudence is an evolving system, open to new societal and political issues. AI path-dependent tools could therefore stymie this evolutive process: the deductive and path-dependent nature of certain AI-ML (Machine Learning) solutions can undermine the important role of human decision-makers in the evolution of law in practice and legal reasoning.

Moreover, at the individual level, path-dependency may also entail the risk of "deterministic analyses",[199] prompting the resurgence of deterministic doctrines to the detriment of doctrines of individualisation of the sanction and with prejudice to the principle of rehabilitation and individualisation in sentencing.

In addition, in several cases, including ADR, both the mediation between the parties' demands and the analysis of the psychological component of human actions (fault, intentionality) require emotional intelligence that AI systems do not have.

These concerns are reflected in the existing legal framework provided by the international legal instruments. The Universal Declaration of Human Rights (Articles 7 and 10), the ICCPR

[195] See also European Commission for the Efficiency of Justice (CEPEJ). 2018. European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment.
[196] See Re and Solow-Niederman, 2019, 252-254 ("Equitable justice entails both reflection on the values set in place by the legal system and the reasoned application of those values, in context […] Codified justice refers to the routinized application of standardized procedures to a set of facts […] In short, codified justice sees the vices of discretion, whereas equitable justice sees its virtues").
[197] See also Clay, 2019. In this regard, for example, a legal system that provides compensation for physical injuries on the basis of the effective patrimonial damages could be automatised, but it will not be able to reconsidered the foundation of the legal reasoning and extend compensation to non- personal and existential damages.
[198] See Re and Solow-Niederman, 2019.
[199] See European Commission for the Efficiency of Justice (CEPEJ). 2018. European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment, 9.

(Article 14), the Convention for the Protection of Human Rights and Fundamental Freedoms (Article 6) and also the Charter of Fundamental Rights of the European Union (Article 47) stress the following key requirements with regard to the exercise of judicial power: equal treatment before the law, impartiality, independence and competency. AI tools do not possess these qualities and this limits their contribution to the decision-making process as carried out by courts.

As stated by the European Commission for the Efficiency of Justice, "the neutrality of algorithms is a myth, as their creators consciously or unintentionally transfer their own value systems into them". Many cases of biases regarding AI applications confirm that these systems too often – albeit in many cases unintentionally – provide a partial representation of society and individual cases, which is not compatible with the principles of **equal treatment before the law** and **non-discrimination**.[200] **Data quality** and other forms of quality **assessment** (impact assessment, audits, etc.) can reduce this risk[201]but, given the degree of potentially affected interests in the event of biased decisions, the risks remain high in the case of equitable justice and seem disproportionate to the benefits largely in terms of efficiency for the justice system.[202]

Further concerns affect the **principles of fair trial and of equality of arms**,[203] when court decisions are based on the results of proprietary algorithms whose training data and structure are not publicly available.[204] A broad notion of **transparency** might address these issues in relation to the use of AI in judicial decisions, but the transparency of AI – a challenging goal in itself – cannot address the other structural and functional objections cited above.

In addition, data scientists can shape AI tools in different ways in the design and training phases, so that were AI tools to became an obligatory part of the decision- making process, governments selecting the tools to be used by the courts could potentially indirectly interfere with the **independence** of the judges.

This risk is not eliminated by the fact that the judge remains free to disregard AI decisions, providing a specific motivation. Although **human oversight** is an important element,[205] its effective impact may be undermined by the psychological or utilitarian (cost-efficient) propensity of the human decision-maker to take advantage of the solution provided by AI.[206]

Crime prevention

The complexity of crime detection and prevention has stimulated research in AI applications to facilitate human activities. In recent years, several solutions[207] and a growing literature have been developed in the field of predictive policing, which is a proactive data-driven approach to crime prevention. Essentially, the available solutions pursue two different goals: to predict where and when crimes might occur or to predict who might commit a crime.[208]

---

[200] See also CEPEJ, 2018.
[201] See also CEPEJ, 2018.
[202] See also Recommendation CM/Rec(2020)1, Appendix, para. 11.
[203] See also CEPEJ, 2018, Appendix I, para. 138.
[204] See also CEPEJ, Appendix I, para. 131 ("the lack of transparency in the algorithm operation processes designed by private companies (which claim intellectual property) was another cause for concern. If we take into account the fact that they take their source data from the state authorities themselves, their lack of accountability to citizens poses a major democratic problem […] an example of this is when ProPublica revealed the flaws in the COMPAS algorithm following the owner company's refusal to share it").
[205] See also CEPEJ, 2018.
[206] See also Mantelero, 2019 ("the supposedly reliable nature of AI mathematics-based solutions can induce those taking decisions on the basis of algorithms to place trust in the picture of individuals and society that analytics suggest").
[207] See Završnik, 2019; European Union Agency for Fundamental Rights, 2018, 98-100; Osoba and Welser, 2017.
[208] For a taxonomy of predictive methods, see Perry et al., 2013, who identifies the following four categories: methods for predict crimes (focused on places and times of crimes), method for predicting offenders (focused on individuals), methods for predicting perpetrators' identities (focused on individuals), and methods for prediction victims of crimes (focused on groups and, in some cases, on individuals).

These two purposes have a distinct potential impact on human rights and freedom, which is more pronounced when AI is used for individual predictions. However, in both cases, we can repeat here the considerations about the general challenges related to AI (obscurity, intellectual property rights, large-scale data collection[209], etc.) discussed in the previous sections and partially addressed by transparency, **data quality**, **data protection**, **auditing** and the other measures. It is worth noting that the role of **transparency**[210] in the judicial context could be limited so as not to frustrate the deterrent effect of these tools. Full transparency could therefore be replaced by auditing and oversight by independent authorities.

Leaving aside the organisational aspects regarding the limitation of police officers' self-determination in the performance of their duties, the main issues with regard to the use of AI to predict crime on geographic and temporal basis concern the impact of these tools on the **right to non-discrimination**.[211] Self-fulfilling bias, community bias[212] and historical bias[213] can produce forms of stigmatisation for certain groups and the areas where they typically live.

Where data analysis is used to classify crimes and infer evidence on criminal networks, proprietary solutions raise issues in terms of respect for the **principles of fair trial and of equality of arms** with regard to the collection and use of evidence. Moreover, if the daily operations of policy departments are guided by predictive software, this raises a problem of **accountability** of the strategies adopted, as they are partially determined by software and hence by software developer companies, rather than the police.

A sharper conflict with human rights arises in the area of predictive policing tools that use profiling to support individual forecasting. Quite apart from the question of data processing and profiling,[214] these solutions can also adversely affect the principle of **presumption of innocence**,[215] procedural **fairness**, and the right to **non- discrimination**.[216]

While non-discrimination issues could be partially addressed, the remaining conflicts seem to be more difficult to resolve. From a human rights standpoint and in terms of proportionality (including the right to respect for private and family life[217]), the risk of prejudice to these principles seems high and not adequately countered by the evidence of benefits for individual and collective rights and freedoms.[218] In the light of future AI regulation, this should urge careful consideration of these issues, taking into account the distinction between the technical possibilities of AI solutions and their concrete benefits in safeguarding and enhancing human rights and freedoms.

Finally, from a wider and comprehensive human rights perspective, the focus on crime by data-driven AI tools drives a short-term factual approach that underrates the social issues that are often crime-related and require long-term social strategies involving the effective enhancement of individual and social rights and freedoms.[219]

---

[209] See also Recommendation Rec(2001)10 on the European Code of Police Ethics, Appendix, para. 42.
[210] See also Barrett, 2017, 361-62.
[211] See European Union Agency for Fundamental Rights, 2018, 10.
[212] See also Barrett, 2017, 358-59 ("For some, the goal of collective safety merits a unilateral sacrifice of some degree of individual rights in this particular context. But that calculus must change if the sacrifice is not collective, but instead confined to minority groups, or becomes fundamentally arbitrary by virtue of an unacceptable degree of error.")
[213] See Bennett Moses and Chan, 2018.
[214] See above Section II.2.
[215] See also Recommendation Rec(2001)10 on the European Code of Police Ethics, Appendix, para. 47.
[216] See also Recommendation Rec(2001)10 on the European Code of Police Ethics, Appendix, para. 49.
[217] See van Brakel and De Hert, 2011, 183. See also *Szabó and Vissy v Hungary* [2016] European Court of Human Rights Fourth Section. Application no. 37138/14.
[218] See Meijer and Wessels, 2019.
[219] See also Rosenbaum, 2006, 245–266

IV. Harmonisation of the principles identified

The previous sections identified **several guiding principles for the future regulation of AI**. These principles were **contextualised with regard to the challenges associated with AI** in the various areas examined, but it is worth looking at the existing level of harmonisation between these principles.

The findings of this study indicate that in a limited number of cases **there are common principles** (the primacy of the human being, individual self-determination, non-discrimination, human oversight). This is due to several factors.

First, **some principles are sector specific**. This is the case, for instance, of the independence of the judges or the principles of fair trial and of equality of arms, which concern justice alone.[220]

Second, some guiding principles are the same in different areas, but with **different nuances** in each context. This is true for transparency, which is often regarded as pivotal in AI regulation, but takes on different meanings in different regulatory contexts.

In the fields of health and personal data, transparency relates to the information given to individuals about the treatment concerning them, with particular attention to the process and related risks and with a strong connotation of individual self- determination. But transparency is also relevant in data protection to control the exercise of power over data in the hands of public and private entities. This different face of transparency is then considered with regard to AI applications for democratic participation and good governance. Then again, in the context of justice, transparency has a more complex significance being vital to safeguard fundamental rights and freedoms (e.g. use of AI in the courts), but also requiring limitations to avoid prejudicing competing interests (e.g. crime detection and prevention in predictive policing).

We can therefore conclude that transparency is a guiding value, but we must go beyond a mere claim to transparency as a key principle for AI regulation. As with other key principles (such as participation, inclusion, democratic oversight, and openness), a proper contextualisation is necessary, adopting provisions that take into account the different contexts in which they operate.

Third, some principles are different, but belong to the **same conceptual area**, assuming various nuances in the different contexts. This is the case with accountability and guiding principles on risk management in general. Here the level of detail and related requirements can be more or less elaborate. For instance, in the field of data protection there are several provisions implementing these principles with a significant degree of detail, whereas in the case of democracy and justice these principles are less developed with regard to data-intensive applications such as AI.

Finally, there are certain components of an AI regulatory strategy that are not principles, but **operational approaches and solutions,** common to the different areas though requiring context-based development. This is the case with the important role played by education and training, interoperability and expert committees.

Such considerations suggest only partial harmonisation is achievable. The regulatory approach to AI should therefore be based on **a legally binding instrument that includes both general provisions** – focusing on common principles and operational solutions – and **more specific and sectoral provisions**, covering those principles that are only relevant in a given field or cases where the same principle is contextualised differently in the different fields.

[220] See also the principles of equitable access and of beneficence in health sector, or the principles of non-interference by public authorities on media activities to influence elections and the obligation to offer equal conditions to all the political parties and candidates in electoral advertising.

V. Conclusions

This analysis has confirmed the validity of the methodological approach adopted, which focuses on the **contextualisation** of guiding principles extracted from legally binding and non-binding intentional instruments. At the same time, it also highlighted the complexity of systematising the provisions of a wide variety of instruments, which differ not only in their binding nature, but also in their specific focus and approach, as well as their structure.

The results have also confirmed that the existing framework based on human rights, democracy and the rule of law can provide an appropriate and common context for the elaboration of **a more specific binding instrument to regulate AI in line with the principles and values enshrined in the international legal instruments, capable of addressing more effectively the issues raised by AI**.

This international framework necessarily leads us to reaffirm the central role of human dignity in the context of AI, where machine-driven solutions cannot be allowed to dehumanise individuals. This may also suggest the introduction of specific limitations to AI when developed or used in a way that is not consistent with respect for human dignity,[221] human rights, democracy and the rule of law.

With a view to future AI regulation, this positive methodological and substantive outcome does not exclude the existence of some gaps. These mainly concern broad areas, such as democracy and justice, where different options and interpretations are available, depending on the political and societal vision of the future relationship between humans and machines.

Further investigation in the field of human rights and AI, as well as the ongoing debate at international and regional level, will contribute to bridging these gaps. However, given the evolving nature of AI, a **co-regulatory approach** is desirable.

A binding instrument establishing the legal framework for AI, including both general common principles and granular provisions addressing specific issues, could therefore be combined with detailed rules set out in **additional non- binding sectoral instruments**. This model would provide both a clear regulatory framework and the flexibility required to address technological development.

[221] See also UNESCO. 1997. Declaration on the Human Genome and Human Rights, Article11.

References

Andorno, R. 2005. The Oviedo Convention: A European Legal Framework at the Intersection of Human Rights and Health Law. Journal of International Biotechnology Law, January 2005. https://doi.org/10.1515/jibl.2005.2.4.133, accessed 20.02.2020.

Azencott, C.-A. 2018. Machine Learning and Genomics: Precision Medicine versus Patient Privacy. Phil. Trans. R. Soc. A 376, no. 2128 (13 September 2018): 20170350. https://doi.org/10.1098/rsta.2017.0350, accessed 14.01.2020.

Barrett, L. 2017. Reasoably Suspicious Algorithms: Predictive Policing at the United States Border. 41 N.Y.U. Rev. Law & Social Change 327.

Bennett Moses, L., Chan, J. 2018. Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability. 28 Policing and Society 806.

Bychawska-Siniarska, D. 2017. Protection the Right to Freedom of Expression under the European Convention on Human Rights. Council of Europe.

Cabitza, F., Rasoini, R., and Gensini, G.F. 2017. Unintended Consequences of Machine Learning in Medicine. JAMA 318, no. 6 (8 August 2017): 517. https://doi.org/10.1001/jama.2017.7797, accessed 18.12.2019.

Calo, R., Citron, D.K. 2020. The Automated Administrative State: A Crisis of Legitimacy. Emory Law Journal, Forthcoming. https://ssrn.com/abstract=3553590, accessed 20.04.2020.

Caruana, R. et al. 2015. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Cham, Switzerland: Springer International Publishing AG.

Clay, T. (ed). 2019. L'arbitrage en ligne. Rapport du Club des Juristes. Paris. 58 https://www.leclubdesjuristes.com/les-commissions/larbitrage-en-ligne/, accessed 30.05.2020.

Committee of Ministers. 1999. Recommendation No. R (99) 5 for the protection of privacy on the internet. Adopted by the Committee of Ministers on 23 February 1999 at the 660th meeting of the Ministers' Deputies.

Committee of Ministers. 2001. Recommendation Rec(2001)10 of the Committee of Ministers to member states on the European Code of Police Ethics. Adopted by the Committee of Ministers on 19 September 2001 at the 765th meeting of the Ministers' Deputies.

Committee of Ministers. 2003. Recommendation CM/Rec(2003)4 on common rules against corruption in the funding of political parties and electoral campaigns. Adopted by the Committee of Ministers on 8 April 2003 at the 835th meeting of the Ministers' Deputies.

Committee of Ministers. 2004. Recommendation CM/Rec(2004)15 on Electroic Governance ("E-Governance"). Adopted by the Committee of Ministers on 15 December 2004 at the 909th meeting of the Ministers' Deputies.

Committee of Ministers. 2007. Recommendation CM/Rec(2007)15 of the Committee of Ministers to member states on measures concerning media coverage of election campaigns. Adopted by the Committee of Ministers on 7 November 2007 at the 1010th meeting of the Ministers' Deputies.

Committee of Ministers. 2009. Recommendation CM/Rec(2009)1 of the Committee of Ministers to member states on electronic democracy (e-democracy). Adopted by the Committee of Ministers on 18 February 2009 at the 1049th meeting of the Ministers' Deputies.

Committee of Ministers. 2009. Recommendation CM/Rec(2009)2 of the Committee of Ministers to member states on the evaluation, auditing and monitoring of participation and participation policies at local and regional level. Adopted by the Committee of Ministers on 11 March 2009 at the 1050th meeting of the Ministers' Deputies.

Committee of Ministers. 2010. Recommendation CM/Rec(2010)13 on the protection of individuals with regard to automatic processing of personal data in the context of profiling. Adopted by the Committee of Ministers on 23 November 2010 at the 1099th meeting of the Ministers' Deputies.

Committee of Ministers. 2011. Recommendation CM/Rec(2011)7 of the Committee of Ministers to member states on a new notion of media. Adopted by the Committee of Ministers on 21 September 2011 at the 1121st meeting of the Ministers' Deputies.

Committee of Ministers. 2012. Recommendation CM/Rec(2012)3 on the protection of human rights with regard to search engines. Adopted by the Committee of Ministers on 4 April 2012 at the 1139th meeting of the Ministers' Deputies.

Committee of Ministers. 2012. Recommendation CM/Rec(2012)4 on the protection of human rights with regard to social networking services. Adopted by the Committee of Ministers on 4 April 2012 at the 1139th meeting of the Ministers' Deputies.

Committee of Ministers. 2014. Recommendation CM/Rec(2014)7 on the protection of whistleblowers. Adopted by the Committee of Ministers on 30 April 2014, at the 1198th meeting of the Ministers' Deputies.

Committee of Ministers. 2016. Recommendation CM/Rec(2016)1 on protecting and promoting the right to freedom of expression and the right to private life with regard to network neutrality. Adopted by the Committee of Ministers on 13 January 2016, at the 1244th meeting of the Ministers' Deputies.

Committee of Ministers. 2016. Recommendation CM/Rec(2016)4 of the Committee of Ministers to member States on the protection of journalism and safety of journalists and other media actors. Adopted by the Committee of Ministers on 13 April 2016 at the 1253rd meeting of the Ministers' Deputies.

Committee of Ministers. 2016. Recommendation CM/Rec(2016)5 of the Committee of Ministers to member States on Internet freedom. Adopted by the Committee of Ministers on 13 April 2016 at the 1253rd meeting of the Ministers' Deputies.

Committee of Ministers. 2017. Recommendation CM/Rec(2017)5 of the Committee of Ministers to member States on standards for e-voting. Adopted by the Committee of Ministers on 14 June 2017 at the 1289th meeting of the Ministers' Deputies.

Committee of Ministers. 2018. Recommendation CM/Rec(2018)1 of the Committee of Ministers to member States on media pluralism and transparency of media ownership. Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies.

Committee of Ministers. 2018. Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries. Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies.

Committee of Ministers. 2018. Recommendation CM/Rec(2018)4 of the Committee of Ministers to member States on the participation of citizens in local public life. Adopted by the Committee of Ministers on 21 March 2018 at the 1311th meeting of the Ministers' Deputies.

Committee of Ministers. 2019. Recommendation CM/Rec(2019)2 of the Committee of Ministers to member States on the protection of health-related data. Adopted by the Committee of Ministers on 27 March 2019 at the 1342nd meeting of the Ministers' Deputies.

Committee of Ministers. 2019. Recommendation CM/Rec(2019)3 of the Committee of Ministers to member States on supervision of local authorities' activities. Adopted by the Committee of Ministers on 4 April 2019 at the 1343rd meeting of the Ministers' Deputies.

Committee of Ministers. 2020. Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems. Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies.

Council of Bars & Law Societies of Europe. 2020. CCBE Considerations on the Legal Aspects of Artificial Intelligence. Brussels.

Council of Europe - Venice Commission, OSCE/ODIHR. 2011. Joint Guidelines on Political Party Regulation. https://www.osce.org/odihr/77812, accessed 20.12.2019.

Council of Europe, Consultative Committee of the Convention of the Protection of Individuals with Regard to Automatic Processing of Personal Data. 2019. Guidelines on Artificial Intelligence and Data Protection. T-PD(2019)01. https://rm.coe.int/2018- lignes-directrices-sur-l-intelligence-artificielle-et-la-protecti/168098e1b7, accessed 16.11.2019.

Council of Europe, Consultative Committee of the Convention of the Protection of Individuals with Regard to Automatic Processing of Personal Data. 2017. Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data. T-PD(2017)1. http://rm.coe.int/t-pd-2017-1-bigdataguidelines- en/16806f06d0, accessed 16.11.2019

Council of Europe, Consultative Committee of the Convention of the Protection of Individuals with Regard to Automatic Processing of Personal Data. 2019. Profiling and Convention 108+: Suggestions for an update. T-PD(2019)07BISrev.

Council of Europe, Directorate General of Democracy – European Committee on Democracy and Governance. 2016. The Compendium of the most relevant Council of Europe texts in the area of democracy http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806b5f2c, accessed 18.11.2019.

Council of Europe, Directorate General of democracy and Political Affairs – Directorate of Democratic Institutions. 2011. Guidelines on transparency of e-enabled elections.

Council of Europe, Directorate General of Democracy and Political Affairs – Directorate of Democratic Institutions. 2009. Project «Good Governance in the Information Society», CM(2009)9.

Council of Europe, Parliamentary Assembly. 2019. Resolution 2254 (2019)1. Media freedom as a condition for democratic elections.

Council of Europe. 2008. The 12 Principles of Good Governance https://www.coe.int/en/web/good-governance/12-principles#{%2225565951%22:[0]}, accessed 16.03.2020.

Council of Europe. 2009. Additional Protocol to the European Charter of Local Self-Government on the right to participate in the affairs of a local authority, Utrecht, 16.

Council of Europe. 2017. Guidelines for civil participation in political decision making, CM(2017)83-final. Adopted by the Committee of Ministers on 27 September 2017 at the 1295th meeting of the Ministers' Deputies.

Council of Europe. 2019. Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes. Adopted by the Committee of Ministers on 13 February 2019 at the 1337th meeting of the Ministers' Deputies.

Council of Europe. 2019. Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes (Adopted by the Committee of Ministers on 13 February 2019 at the 1337th meeting of the Ministers' Deputies).

Council of Europe. Guidelines for civil participation in political decision making. CM(2017)83-final. https://www.coe.int/en/web/youth/-/guidelines-for-civil-participation-in-political-decision-making, accessed 15.03.2020.

Council of Europe-Committee of experts on internet intermediaries (MSI-NET). 2018. Algorithms and Human Rights. Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications. https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5, accessed 28.11.2019.

Crawford, K., and Joler, V. 2018. Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources. AI Now Institute and Share Lab. http://www.anatomyof.ai, accessed 27.12.2019.

Cummings, M. L., Roff H. M., Cukier K., Parakilas J. and Bryce H. 2018. Chatham House Report. Artificial Intelligence and International Affairs Disruption Anticipated. London: Chatham House. The Royal Institute of International Affairs. https://www.chathamhouse.org artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf, accessed 21.03.2020.

Dupont, B. et al. 2018. Artificial Intelligence in the Context of Crime and Criminal Justice. Korean Institute of Criminology 2018. https://www.cyberjustice.ca/publications/lintelligence-artificielle-dans-le-contexte-de- la-criminalite-et-de-la-justice-penale/, accessed 30.05.2020.

EU Code of Practice on Disinformation, 2018 https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation, 23.03.2020.

European Commission for Democracy Through Law (Venice Commission). 2002. Code of Good Practice in Electoral Matters. Guidelines and Explanatory Report. Adopted by the Venice Commission at its 51st and 52nd sessions (Venice, 5-6 July and 18-19 October 2002).

European commission for Democracy trough law (Venice Commission). 2019. Joint Report of the Venice Commission and of the Directorate of Information society and Actions Against Crime of the Directorate General of Human Rights and Rule of Law (DGI) on Digital Technologies and Elections.

European Commission for the Efficiency of Justice (CEPEJ). 2018. European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment. Adopted by the CEPEJ during Its 31st Plenary Meeting (Strasbourg, 3-4 December 2018) https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c, accessed 04.12.2018.

European Commission, Networks, Content and Technology- Directorate-General for Communication. 2018. A Multi-Dimensional Approach to Disinformation Report of the Independent High Level Group on Fake News and Online Disinformation. https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert- group-fake-news-and-online-disinformation, accessed 22.03.2018.

European Commission. 2014. Green paper on mobile health. http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=5147, accessed 12.01.2020.

European Commission. 2020. A European strategy for data, COM(2020) 66 final. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital- age/european-data-strategy_en, accessed 20.02.2020.

European Commission. 2020. Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, COM(2020) 64final. https://ec.europa.eu/info/publications/commission-report-safety-and-liability- implications-ai-internet-things-and-robotics-0_en, accessed 20.02.2020.

European Commission. 2020. White Paper on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european- approach-excellence-and-trust_en, accessed 20.02.2020.

European Union Agency for Fundamental Rights. 2018. #BigData: Discrimination in Data-Supported Decision Making. https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making, accessed 20.05.2020.

European Union Agency for Fundamental Rights. 2018. Preventing Unlawful Profiling Today and in the Future: A Guide. https://fra.europa.eu/en/publication/2018/preventing-unlawful-profiling-today-and- future-guide, accessed 20.05.2020.

Faye Jacobsen, A. 2013. The Right to Public Participation. A Human Rights Law Update. Issue Paper. The Danish Institute for Human Rights.

Ferryman, K. and Pitcan, M. 2018. Fairness in Precision Medicine. Data & Society, February, https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf, accessed 20.12.2019.

González Fuster, G. 2014. The Emergence of Personal Data Protection as a Fundamental Right of the EU. Cham-New York: Springer International Publishing.

Maisley, N. 2017. The International Right of Rights? Article 25(a) of the ICCPR as a Human Right to Take Part in International Law-Making. 28 European Journal of International Law 89.

Manheim, K., Kaplan, L. 2019. Artificial Intelligence: Risks to Privacy and Democracy. 21 Yale J.L. & Tech. 106.

Mantelero, A. 2018. AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment. 34 Computer Law & Security Review 754.

Mantelero, A. 2019. Artificial Intelligence and Data Protection: Challenges and Possible Remedies. Report on Artificial Intelligence. Consultative Committee of the Convention for the Protection of Individuals with Regard to Automatic Processing of personal data: Strasbourg. T-PD(2018)09Rev, https://rm.coe.int/artificial-intelligence- and-data-protection-challenges-and-possible-re/168091f8a6, accessed 20.02.2020.

Mayer-Schönberger, V. 1997. Generational development of data protection in Europe? In Agre, P.E., Rotenberg, M. (eds). Technology and privacy: The new landscape. Cambridge, MA: MIT Press.

Mehr, H. 2017. Artificial Intelligence for Citizen Services and Government. Harvard Kennedy School. Ash Center for Democracy and Innovation.

Meijer, A., Wessels, M. 2019. Predictive Policing: Review of Benefits and Drawbacks. 42 International Journal of Public Administration 1031.

Nunez, C. 2017. Artificial Intelligence and Legal Ethics: Whether AI Lawyers Can Make Ethical Decisions. 20 Tul. J. Tech. & Intell. Prop. 189-204.

OECD. 2013. Recommendation of the Council concerning Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data, C(80)58/FINAL, as amended on 11 July 2013 by C(2013)79.

OSCE. 2020. Non-Paper on the Impact of Artificial Intelligence on Freedom of Expression. https://www.osce.org/representative-on-freedom-of-media/447829, accessed 11.06.2020.

Osoba, O.A., Welser, W. 2017. An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence. RAND Corporation. https://www.rand.org/pubs/research_reports/RR1744.html, accessed 20.05.2020.

Parliamentary Assembly. 2019. Resolution 2254 (2019)1. Media freedom as a condition for democratic elections.

Parliamentary Assembly. 2019. Resolution 2300 (2019)1, Improving the protection of whistle-blowers all over Europe

Perry, W.L. et al. 2013. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. RAND Corporation 2013. https://www.rand.org/pubs/research_reports/RR233.html, accessed 30.03.2020.

Privacy International. 2017. Smart Cities: Utopian Vision, Dystopian Reality. https://privacyinternational.org/report/638/smart-cities-utopian-vision-dystopian- reality, accessed 21.03.2020.

Re, R.M., Solow-Niederman, A. 2019. Developing Artificially Intelligent Justice. 22 Stan. Tech. L. Rev. 242.

Rosenbaum, D., 2006. The limits of hot spots policing. In: D. Weisburd and A. Braga, eds. Police innovation: contrasting perspectives. New York, NY: Cambridge University Press, 245–266.

Rouvroy, A. 2016. "Of Data and Men" - Fundamental rights and freedoms in a world of Big Data. Consultative Committee of the Convention for the Protection of Individuals with Regard to Automatic Processing of personal data: Strasbourg. T-PD- BUR(2015)09Rev, http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806a6020, accessed 04.11.2019.

Savaget, P., Chiarini, T., Evans, S. 2019. Empowering Political Participation through Artificial Intelligence. 46 Science and Public Policy 369.

Seatzu, F. 2015. The Experience of the European Court of Human Rights with the European Convention on Human Rights and Biomedicine. 31(81) Utrecht Journal of International and European Law 5, DOI: http://dx.doi.org/10.5334/ujiel.da, accessed 07.12.2019.

Sunstein, C.R. 2015a. The Ethics of Nudging. Yale Journal on Regulation 32: 413

Sunstein, C.R. 2015b. Why Nudge? The Politics of Libertarian Paternalism. New Haven: Yale University Press.

Sunstein, C.R., Thaler, R. 2003. Libertarian Paternalism in Not an Oxymoron. University of Chicago Law Review 70 (4): 1159.

Thaler, R., Sunstein, C.R. 2008. Nudge. New Haven, CT: Yale University Press.

The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information. 2017. Joint Declaration on "Fake News," Disinformation and Propaganda. http://www.osce.org/fom/302796?download=true, accessed 02.02.2020.

T-PD(2019)07BISrev. Consultative Committee of the Convention of the Protection of Individuals with Regard to Automatic Processing of Personal Data. 2019. Profiling and Convention 108+: Suggestions for an update

UN Office of the High Commissioner for Human Rights. 1981. CESCR General Comment No. 1: Reporting by States Parties. Adopted at the Thirteenth Session of the Committee on Economic, Social and Cultural Rights, on 27 July 1981 (Contained in Document E/1989/22).

UN Office of the High Commissioner for Human Rights. 1996. General Comment No. 25: The right to participate in public affairs, voting rights and the right of equal access to public service (Art. 25). CCPR/C/21/Rev.1/Add.7.

UNESCO. 2019. Preliminary Study on a Possible Standard-Setting Instrument on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000369455, accessed 21.11.2019.

UNESCO. Declaration on the Human Genome and Human Rights (11 November 1997).

United Nations Convention against Corruption, 2003.

van Brakel, R., De Hert, P. 2011. Policing, Surveillance and Law in a Pre-Crime Society: Understanding the Consequences of Technology Based Strategies. 3 Cahiers Politiestudies, Jaargang 163.

Verbeek, P-P. 2011. Understanding and Designing the Morality of Things. Chicago-London: The University of Chicago Press.

Završnik, A. 2019. Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings. EuropeanJournal of Criminology, 1-20, https://doi.org/10.1177/1477370819876762, accessed 20.02.2020.

Annex 1. Legal instruments

Binding instruments	Related non-binding instruments
Biomedicine	
Council of Europe Convention on Human Rights and Biomedicine ('Oviedo Convention') Additional Protocol concerning Genetic Testing for Health Purposes Additional Protocol concerning Biomedical Research	Rec(2016)8 on the processing of personal health-related data for insurance purposes, including data resulting from genetic tests and its Explanatory Memorandum Recommendation CM/Rec(2016)6 of the Committee of Ministers to member States on research on biological materials of human origin Strategic Action Plan on Human Rights and Technologies in Biomedicine 2020-2025
Antidiscrimination	
- Universal Declaration of Human Rights - International Covenant on Civil and Political Rights International Covenant on Economic, Social and Cultural Rights - International Convention on the Elimination of All Forms of Racial Discrimination - Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) - Convention on the Rights of Persons with Disabilities - European Convention on Human Rights (ECHR) and its Protocols (No.12 in particular) - European Social Charter - Convention on Cybercrime and its Additional Protocol concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems - Convention on preventing and combating violence against women and domestic violence (Istanbul Convention) - Charter of the Fundamental Rights of the European Union	ECRI's General Policy Recommendations, no. 2 (on equality bodies), 11 (on combating racial discrimination in policing) and 15 (on hate speech) in particular. PACE Recommendation 2098 (2017) on Ending cyberdiscrimination and online hate CM Recommendation (2019)1 on Preventing and Combating Sexism
Cybercrime and electronic evidence	
Convention on Cybercrime	Guidance Notes by the Cybercrime Convention Committee on DDOS attacks, Critical information infrastructure attacks, Malware, Spam, Identity theft etc.
Justice	
- Universal Declaration of Human Rights - International Covenant on Civil and Political Rights	CEPEJ. 2019. European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment
- International Convention on the Elimination of All Forms of Racial Discrimination - Convention on the Elimination of All	

Forms of Discrimination against Women - Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) - Charter of Fundamental Rights of the European Union	
Congress of Local and Regional Authorities	
The European Charter of Local Self-Government	Congress Resolution 435 (2018) and Recommendation 424 (2018) "Transparency and open government." Congress Resolution 417 (2017) and Recommendation 398 (2017) "Open data for better public services. Congress Resolution 394 (2015) E-media: game changer for local and regional politicians. Congress Resolution 290 (2009) E- democracy: opportunities and risks for local authorities.
Democracy and participation	
- Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) - Convention on the protection of individuals with regard to automatic processing of personal data ETS No. 108 of 1981 and the 2018 Protocol modernising the Convention	- Committee of Ministers Recommendation Rec(2003)4 on common rules against corruption in the funding of political parties and electoral campaigns - Code of Good Practice in Electoral Matters (Venice Commission) - Joint Guidelines on Political Party Regulation (Venice Commission and OSCE/ODIHR) - Recommendation CM/Rec(2007)15 of the Committee of Ministers to member states on measures concerning media coverage of electoral campaigns - see also Recommendation CM/Rec(2018)1 on media pluralism and transparency of media ownership, Recommendation CM/Rec(2018)2 on the roles and responsibilities of internet intermediaries, Recommendation CM/Rec(2016)1 on protecting and promoting the right to freedom of expression and the right to private life with regard to network - 1999 Committee of Ministers Recommendation No. R (99) 5 for the protection of privacy on the internet, 2010 Recommendation CM/Rec(2010)13 on the protection of individuals with regard to automatic processing of personal data in the context of profiling, Recommendation CM/Rec(2012)3 on the protection of human rights with regard to search engines,

	Recommendation CM/Rec(2012)4 on the protection of human rights with regard to social networking services
Freedom of expression	
- European Convention on Human Rights - International Covenant on Civil and Political Rights - Charter of Fundamental Rights of the European Union	UDHR CM/Rec(2018)2 on roles and responsibilities of internet intermediaries CM/Rec(2020)x on the human rights impacts of algorithmic systems Decl(13/02/2019) on the manipulative capabilities of algorithmic processes CM/Rec(2018)1 on media pluralism and transparency of media ownership CM/Rec(2020)x on promoting a favourable environment for quality journalism in the digital age
Elections	
Universal Declaration on Human Rights International Covenant on Civil and Political Rights United Nations Convention on the Elimination of All Forms of Racial Discrimination United Nations Convention on the Elimination of All Forms of Discrimination against Women United Nations Convention on the Rights of Persons with Disabilities United Nations Convention against Corruption Convention for the Protection of Human Rights and Fundamental Freedoms (ETS No. 5) Protocol to the Convention for the Protection of Human Rights and Fundamental Freedoms (ETS No. 9) European Charter of Local Self- Government (ETS No. 122) European Charter for Regional or Minority Languages (ETS No. 148) Convention on Cybercrime (ETS No. 185) Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (ETS No. 108)	Code of Good Practice in Electoral Matters, adopted by the Council for Democratic Elections of the Council of Europe and the European Commission for Democracy through Law (Venice Commission) Recommendation Rec(2003)3 of the Committee of Ministers to member states on balanced participation of women and men in political and public decision making Convention on the Standards of Democratic Elections, Electoral Rights and Freedoms in the Member States of the Commonwealth of Independent States (CDL-EL(2006)031rev) Recommendation Rec(99)5 of the Committee of Ministers to member States on the protection of privacy on the Internet Recommendation Rec(2004)15 of the Committee of Ministers to member States on electronic governance (e-governance) Recommendation CM/Rec(2007)15 of the Committee of Ministers to member states on measures concerning media coverage of election campaigns Recommendation CM/Rec(2009)1 of the Committee of Ministers to member States on electronic democracy (e-democracy)

Additional Protocol to the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, regarding supervisory authorities and transborder data flows (ETS No. 181) Charter of Fundamental Rights of the European Union Framework convention for the protection of national minorities and explanatory report	Recommendation CM/Rec(2017)5 of the Committee of Ministers to member States on standards for e-voting Document of the Copenhagen Meeting of the Conference on the Human Dimension of the OSCE Report on the misuse of administrative resources during electoral processes adopted by the Council for Democratic Elections and by the Venice Commission (CDL-AD(2013)033) Report on electoral rules and affirmative action for national minorities' participation in decision making in European countries adopted by the Council for Democratic Elections and the Venice Commission (CDL-AD(2005)009) Code of good practice on referendum adopted by the Council for Democratic Elections and the Venice Commission (CDL-AD(2007)008rev-cor) Council of Europe Disability Strategy 2017- 2023 Resolution 1897 (2012) of the PACE, Ensuring greater democracy in elections Code of Good Practice in the field of Political Parties adopted by the Venice Commission and Explanatory Report adopted by the Venice Commission (CDL- AD(2009)021)
Democracy (excluding issues relating to elections and electoral cycle)	
- Universal Declaration of Human Rights - International Covenant on Civil and Political Rights - International Convention on the Elimination of All Forms of Racial Discrimination - Convention on the Elimination of All Forms of Discrimination against Women - Charter of Fundamental Rights of the European Union - Convention 108+ - Convention for the Protection of Human Rights and Fundamental Freedoms and Protocols - European Charter of Local Self-Government - Framework Convention for the Protection of National Minorities	- Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes See also Compendium Chapter A (separation of powers / good governance) Chapter B (media pluralism & diversity ; protection pf freedom of expression on the Internet) Chapter C (enabling civil society) Chapter E (citizen's participation)

Good Governance	
- Universal Declaration of Human Rights - International Covenant on Civil and Political Rights - International Convention on the Elimination of All Forms of Racial Discrimination - Convention on the Elimination of All Forms of Discrimination against Women - Convention for the Protection of Human Rights and Fundamental Freedoms and Protocols - Charter of Fundamental Rights of the European Union - Convention 108+ - European Charter of Local Self-Government and Protocols - Council of Europe Convention on Access to Official Documents	-12 principles of good democratic governance - Recommendation of the Committee of Ministers to member States on supervision of local authorities' activities CM/Rec(2019)3 See also Compendium Chapter A (good governance) Chapter E (Integration policies – standards and mechanisms) And see https://www.coe.int/en/web/good-governance/conventions- recommendations
Gender equality including violence against women	
- Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) - Council of Europe Convention of Preventing and Combating Violence against Women (article 17§1 on the participation of the ICT sector in the prevention & fight against violence against women, Article 34 cyber stalking) -European Social Charter - UN Convention on the Elimination of All Forms of Discrimination against Women - Universal Declaration of Human Rights - International Covenant on Civil and Political Rights International Covenant on Economic, Social and Cultural Rights - International Convention on the Elimination of All Forms of Racial Discrimination - Charter of Fundamental Rights of the European Union	CM Recommendation (2019)1 on Preventing and Combating Sexism CM Recommendation (2013)1 on gender equality and media ECRI's General Policy Recommendations, no. 15 on hate speech
Culture, Creativity and Heritage	
Universal Declaration of Human Rights International Covenant on Economic, Social and Cultural Rights Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR)	
EFCNM	Numerous CoE/CM and PACE and Congress RECs and Resolutions on issues of cultural identity, diversity
European Charter for Regional and Minority Languages[158]	Numerous CoE/CM and PACE and Congress RECs and Resolutions on issues of cultural identity, diversity and dialogue and minorities

CoE Conventions in the Cultural Heritage Sector (Nicosia Convention =not yet in force; Faro Convention; La Valetta Convention; Granada Convention)	Numerous CoE/CM and PACE RECs on issues of Cultural heritage
UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions	CoE Declaration on Cultural diversity CoE CM Rec on the UNESCO Convention
Council of Europe Convention on Cinematographic Co-production (revised) EU's Audiovisual Media Services Directive (AVMSD) / Directive (EU) 2018/1808 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC	Resolution (88)15 amended setting up a European Support Fund for the Co- production and Distribution of Creative Cinematographic and Audiovisual Works ("Eurimages") Recommendation CM/Rec(2017)9 of the Committee of Minister to member States on gender equality in the audiovisual sector
Universal Declaration of Human Rights International Covenant on Economic, Social and Cultural Rights Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR)	
GRECO	
Criminal and Civil Law Conventions on Corruption; GRECO monitoring	CM recommendations on model code of conduct for public officials; lobbying whistleblower protection; transparency of political party funding, etc…
Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR)	Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR)
ESC	
European Social Charter (1961 Charter, 1988 Protocol and 1996 revised Charter)	
European Social Charter rights more specifically	Some examples: - New Strategy and Council of Europe Action Plan for Social Cohesion (approved by CM on 7 July 2010); - CM Rec(2000)3, proposing an individual universal and enforceable right to the satisfaction of basic material needs; - etc.
In addition, there are many social rights (and Charter) aspects related to subjects covered by a wide range of other areas of CoE work:	Some examples: - CM Rec(93)1 on effective access to the law and to justice to the very poor; - social rights aspects of the prison rules (health care, living conditions, employment, education, family rights,…); - etc.

Annex 2. Impacted areas

Impacted areas (applications)
Biomedicine AI-based surveillance, prevention, diagnosis and intervention in healthcare settings • Prediction-based surveillance, diagnosis, monitoring, financing (insurance) treatments (e.g. user facing apps and online services beyond healthcare settings)
Antidiscrimination • Automated Decision-making covering different areas in both public and private sectors (e.g. job applications, welfare/social benefits, access to goods and services, such as bank loans, insurance) Predictive policing (which holds high risk of racial profiling) Predictive justice Facial recognition • Behavioural prediction technologies such as emotional recognition and AI-based lie detection Personal assistance tools (e.g. Siri) Content moderation Data protection
Cybercrime and electronic evidence Automated cybercrime and cyberattacks, such as: Distributed denial of service (DDOS) attacks Criticial information infrastructure attacks Man-in-the-middle attacks Phishing and similar social engineering techniques Scanning for vulnerabilities Etc. Cybercrime investigations and computer forensics: Collection and analysis of electronic evidence (in relation to any crime). Attribution Reverse engineering Cybersecurity and prevention of cybercrime: Detection of malware, intrusions, etc. Automated patching of vulnerabilities
Justice sector Processing of judicial decisions and data: to support judicial decision-making or judicial research) On-line dispute resolution Provision of legal advice to litigants Predictive policing
Congress of Local and Regional Authorities Provision of local public services. Instruments to promote citizen participation. Wide variety of digital and electronic applications in cities and local communities. • Application of information and communication technologies (ICT) to improve the quality of life and working environments in cities. Smart city-governance. The embedding of ICT within government systems. • Local roll-out of practices that bring people and ICT together in order to foster innovation and enhance the knowledge that they offer.

Freedom of expression

- Individual communication (through automated content moderation and restriction – algorithmic sorting, classification, optimisation and recommender systems)
- Media production and distribution (robo-journalism, data-journalism, NLP, micro-targeting of reader-base, automated newsfeeds based on reader profile)
- Societal and political communication/ fragmentation/polarisation of public discourse, political redlining (micro-targeting of voterbase, opinion swaying through bots, proliferation of automated local media sites)

Elections

Pre-electoral period:
Planification of electoral calendar
Training of electoral stakeholders
Delimitation of electoral constituency
Registration of voters and candidates
Accreditation of observers (international and domestic)
Update of the list of voters
Update of legal framework
Financing of political parties
Electoral propaganda by administration and by political parties/candidates

Electoral period:
Financing of electoral campaigns
Access to media
Voting
Counting of ballots
Tabulation of results

Post-election period:
Publication of electoral results
Electoral dispute resolution

Democracy (excluding issues relating to elections and electoral cycle)

Separation of power
Civil society participation
Citizen's participation
Privacy
Citizenship
Protection of minorities
Pluralism & diversity
Legitimacy

Good Governance
Local governments
Regional
Administration
Service delivery
Budgetary allocation
Social security and social benefit systems
Police and judiciary
Smart cities
Public tender and procurement
Institutional capacities

Gender equality including violence against women (VAW)

General issue of inherited gender bias from the data systems algorithms train on (valid for many areas), which may lead to aggravated gender and social inequalities.

General issues related to AI as an employment sector:

- The lack of participation /under-representation of women exacerbates the potential gender biases and excludes them from a powerful sector
- Exploitation of “click workers” in Europe and worldwide (low salaries, no social protection, no labour rights , long term exposition to damaging content for content moderators etc.)

Specific challenges
Discriminatory job screening
Automated decision-making for public and private services
Facial & speech recognition (performing worse for women, especially some groups)
Surveillance /stalking facilitated by AI tools ex in the context of domestic violence

- Automated decision-making exacerbating the possibility for multiple discrimination based on sex/gender, race and social origin by combining secondary data like level of education, address, level of income.

Predictive justice (ex VAW)

- Predictive health based on gender-biased data (ex some diseases characterised as “female” or “male”)
- Inherited biases in machine-led content moderation (high tolerance for sexism, sexist hate speech & VAW)

Gendered virtual assistants / robots perpetuating gender stereotypes
Gendered marketing perpetuating gender stereotypes
Differential pricing based on sex/gender

Positive impacts

- Use of GPS tracking devices to ensure respect of protection orders in cases of VAW

Use of AI by law enforcement agencies to conduct risk assessment in DV cases

- Use of AI to identify and track gender bias and being able to quarantine or eliminate the spreading of (sexist) hate speech on platforms

Developments of Apps to support and inform victims of VAW

- Use of AI-based tools to analyse content and track gender bias / analyse representation (ex in movies or other media)

Culture, Creativity and Heritage

Access and participation in public / cultural life;
FoE (incl. freedom of artistic expression)
Access to impartial information?
Automated decision making, targeting, profiling
Automated decision making, targeting, profiling;
But also learning of endangered languages to preserve/ protect them

- Automated assistance in administration, health etc. for speakers from minority groups/ languages
- Geolocalisation, Predictive policing, criminal analytics (re destruction, looting, trafficking of cultural property; targeting; learning re endangered heritage can help with its protection
- Automated creation of content, targeting, profiling (re cultural creation, exchange, consumption)

Audiovisual content development & production:
Predictive audience analysis
Automated script analysis
Assisted or automated script writing
Computer Generated Images (SFX, Animation...)
Automated location scouting, scheduling and budgeting (impact yet to be assessed)
Content distribution

Recommendation algorithms
Targeted advertising
- Automated control of content (compliance with regulations) / Censorship (ref. Study "Entering the new paradigm of artificial intelligence and series" commissioned by DG2 and Eurimages)
Access and participation in public / cultural life;
FoE (incl. freedom of artistic expression)
Access to impartial information?

GRECO

Anti-corruption
Criminal liability related to the use
of automated vehicles
Article 8: Right to respect for private and family life

ESC

All areas of social rights, social security, social cohesion, etc. Including, but not limited to:
many aspects of employment (including but not limited to monitoring and surveillance, job screening and work in the platform economy, etc);
ditto different aspects of health (the right to enjoy the highest standard of health attainable);
ditto education;
equally for social protection, integration and participation;
let alone non-discrimination;
housing and protection from social exclusion;

For example:
justice (both as regards the administration of justice, and criminal justice and prisons;
trafficking in human beings (forced labour and exploitation, …);
migration and refugees;
gender equality, plus violence against women;
children and youth, plus education;
bioethics;
non-discrimination, Roma and Travellers, SOGI ;
drug policy;
participation and culture;
sport;

Annex 3. Principles

Guiding principles and legal values	Missing principles
Biomedicine	
Primacy of the human being Privacy and confidentiality Informed consent Autonomy Non-discrimination Non-maleficence/beneficence Accountability Transparency and Equitable Access Public debate	Precautionary principle Human control/oversight Explainability Liability for AI-based decision making Gender equality/equity
Antidiscrimination	
Non-discrimination and equality Diversity and inclusion Intersectionality Right to an effective remedy Right to a fair trial Right to privacy Presumption of innocence and burden of proof Transparency Impartiality Fairness Human control/oversight Access to digital skills	Explainability of AI systems Inclusiveness in design, development and deployment of AI systems
Cybercrime and electronic evidence	
Specific conduct to be criminalised. Specified data in specific criminal investigations to be secured for use as evidence. Effective powers to secure electronic evidence limited by the rule of law conditions and safeguards.	Problem of evidence in the cloud versus territorial enforcement jurisdiction for criminal justice (to be addressed in the 2nd Additional Protocol to the Budapest Convention).
Justice sector	
Non-discrimination Data quality & security Transparency Impartiality Fairness Freedom of choice/ Independence of judges (decision-making process) Human control/oversight Guarantees of the right of access to the judge Guarantees of the right to a fair trial	Precautionary principle for applications missing fundamental transparency requirements
Congress	
Transparency Human control (oversight) Impartiality Right to privacy Data security Cyber security Non-discrimination Inclusive cities Financial sustainability Monitoring safety Service efficiency Digital literacy	Democracy and participation – Deep fakes, Microtargeting and propaganda in the framework of electoral processes

Democracy and participation	
Right to free elections Freedom of expression Right of individuals to access the internet Right to private life; Data protection Equality of opportunity for parties and candidates Requirement of a neutral attitude by state authorities with regard to the election campaign, to coverage by the media, and to public funding of parties and campaigns Requirement of a minimum access to privately owned audio-visual media, with regard to the election campaign and to advertising, for all participants in elections Transparency in campaign funding Prevention of improper influence on political decisions through financial donations Responsible, accurate and fair media coverage of electoral campaigns; right of reply, modalities of disseminating opinion polls, transparency requirements on paid advertising content; media pluralism Network neutrality Protection of individuals with regard to the collection and processing of personal data on information highways Non-discrimination Data quality & security Transparency Impartiality Fairness Freedom of choice/ Independence of judges (decision-making process) Human control/oversight Guarantees of the right of access to the judge Guarantees of the right to a fair trial	Balance between sometimes conflicting rights such as e.g. - right to free elections / freedom of expression - right of access to information including on the internet / right to private life, data protection Standards which would be applicable and adequate for digital advertising/campaigns, e.g. with respect to - equality of opportunity for parties and candidates - election campaign and campaign funding, transparency and enforcement - fair media coverage, media pluralism - accountability of internet intermediaries in terms of transparency and access to data enhancing transparency of spending, specifically for political advertising - net neutrality - data protection
Freedom of expression	
Individual autonomy Equality Democratic security Transparency and accountability Independence of the media Diversity and pluralism	Precautionary principle for applications missing fundamental transparency requirements

<table>
<tr><td colspan="2">Elections</td></tr>
<tr><td>- Free and fair elections
- Freedom of choice/opinion/speech
- Universal suffrage
- Equal suffrage
- Free suffrage
- Secret suffrage
- Direct suffrage
- Frequency of elections
- Transparency of electoral process
- Inclusiveness of electoral process
- Gender balanced
- participation/representation in public decision-making</td><td>- Principle of use of AI systems in electoral processes (especially e-voting systems, etc.)
- Opportunities offered by AI to have more inclusive electoral processes (AI as tool for the Electoral Management Bodies and election commissions, AI as an assistant for the voters).</td></tr>
<tr><td colspan="2">Democracy (excluding issues relating to elections and electoral cycle)</td></tr>
<tr><td>Transparency Impartiality Fairness
Freedom of choice Freedom of expression
Freedom of assembly and association
Access to information
Human control/oversight Diversity
Equality
Non-discrimination Data quality & security
Data protection
Independence</td><td>- Role of intermediaries
- Tech & digital literacy
- Question of who owns the data
- Democratic oversight
- Open data and open government
- Risk assessment</td></tr>
<tr><td colspan="2">Good Governance</td></tr>
<tr><td>- Non-discrimination
- Data quality & security
- Impartiality
- Fairness
- Participation, Representation Fair Conduct of Elections
- Responsiveness
- Efficiency and Effectiveness
- Openness and Transparency
- Rule of Law
- Ethical Conduct
- Competence and Capacity
- Innovation and Openness to Change
- Sustainability and Long-term Orientation
- Sound Financial Management
- Human rights, Cultural Diversity and Social Cohesion
- Accountability
- Redress mechanisms
- Access to remedy
- Independence</td><td>- Democratic oversight
- Access to remedy and redress mechanisms in case of automated and algorithmic decisions making by public officials
- Role of intermediaries
- Tech literacy & competences
- Questions of who actually owns the data
- Open data and open government
- Civil and criminal liability
- Risk assessments and risk management</td></tr>
</table>

Gender equality including violence against women	
Equality and non-discrimination Integrity / Elimination of violence (against women) Equal access to justice Guarantees of the right to a fair trial and to redress	Un-biased data (Gender) inclusiveness of AI as a sector AI as an employment sector respecting labour and social rights Data quality & security Transparency & explainability Accountability Impartiality Fairness
	Human control/oversight Digital literacy and closing existing digital (gender) gaps, essential with regards to right to redress – if citizens & consumers do not understand AI, they will not be able to claim their rights Precautionary principle for applications missing fundamental transparency requirements Ethical principles such as “do no harm” are not respected because some of the spyware apps are developed and advertised for the sole purpose of “knowing what your wife is up to”.
Culture, Creativity and Heritage	
Non-discrimination Access, Freedom of Association, Right to participate in cultural life and create and learn (Covenant) Freedom of Expression Access to impartial information	Precautionary principle for applications missing fundamental transparency requirements Need to develop cultural paradigms and techniques to deal with Autonomization (only exist for Automatisation) “Avoid further centralisation of knowledge and power in the hands of those, who already have it and further dis-empower those who don’t” (M. Whitaker) Need to stress rules and rights on access to common goods, and to participate in public life (citizen-centred practices)
Non-discrimination – Impartiality (Protection of National Minorities)	
Non-discrimination – Impartiality Protection of Minorities and their cultural expressions (languages / linguistic diversity, cultural heritage)[222]	Ownership and possible bias of information fed into AI-driven learning applications
Promote/ protect European identity, diversity, co-operation Access to and participation in cultural heritage; protection of cultural heritage	Protection of Human creativity (distinctive nature of human creativity)

[222] An AI-Language Recreation Machine could fill gaps and help develop a living global language archive, a “Louvre of Languages”.

Human control/oversight over creative process, transparency Protection and Promotion of cultural diversity Creating conditions for culture to flourish and freely interact Recognise the distinctive nature of cultural activities, goods and services as vehicles of identity, values and meaning	IP and copyright management Protection of Human Creativity (distinctive nature of human creativity)
Cultural diversity Cultural cooperation in Europe and beyond Availability of works Non-Discrimination Data protection Freedom of expression and of creation	Visibility of works Transparency of decision-making (to develop and produce / to censor / to recommend a work) IP ownership, Copyright and moral rights issues
Human control/oversight, transparency	
Non-discrimination Access, Freedom of Association, Right to participate in cultural life and create and learn (Covenant) Freedom of Expression Access to impartial information	Precautionary principle for applications missing fundamental transparency requirements Need to develop cultural paradigms and techniques to deal with Autonomization (only exist for Automatisation) “Avoid further centralisation of knowledge and power in the hands of those, who already have it and further dis-empower those who don’t” (M. Whitaker) Need to stress rules and rights on access to common goods, and to participate in public life (citizen-centred practices)
GRECO	
Guiding Principles for the Fight against Corruption	Nothing specific on: AI applications to prevent corruption; need to make sure algorithm are not corrupted
	Ongoing work by the CDPC on Criminal liability related to the use of automated vehicles
Article 8: Right to respect for private and family life	Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR)

ESC	
Various broad principles emerge from the Charter and the monitoring activities under the Charter about transparency and participation in decision-making	Automated or computer-assisted or AI- enabled decisions-making would require: - mandatory human oversight in order to mitigate and/or avoid errors in the management, attribution or revocation of entitlements, assistance and related benefits which could amplify disadvantage and/or disenfranchisement; - effective arrangements to protect vulnerable persons from destitution, extreme want or homelessness, and from serious injury or irreparable harm, as a result of the implementation of computer- assisted or AI-enabled decisions in the area of social services; - a proactive approach with a view to ensure that those affected by computer- assisted or AI-enabled decisions in the area of social services, in particular persons in a situation of extreme deprivation or vulnerability, can effectively assert their rights and seek remedies.

Annex 4. Data Protection

Binding and non-binding instruments in the field of data protection

Convention 108+ Consultative Committee of the Convention for the Protection of Individuals with Regard to Automatic Processing of personal data (Convention 108). 2019. Guidelines on the data protection implications of artificial intelligence[223]	**Human control** I.6 AI applications should allow meaningful control by data subjects over the data processing and related effects on individuals and on society **Value-oriented design** II.1. AI developers, manufacturers and service providers should adopt a values-oriented approach in the design of their products and services, consistent with Convention 108+, in particular with article 10.2, and other relevant instruments of the Council of Europe. **Precautionary approach** II.2 AI developers, manufacturers and service providers should assess the possible adverse consequences of AI applications on human rights and fundamental freedoms, and, considering these consequences, adopt a precautionary approach based on appropriate risk prevention and mitigation measures. **Human rights by-design approach and bias detection** II.3 In all phases of the processing, including data collection, AI developers, manufacturers and service providers should adopt a human rights by-design approach and avoid any potential biases, including unintentional or hidden, and the risk of discrimination or other adverse impacts on the human rights and fundamental freedoms of data subjects. **Data quality and minimisation** II.4 AI developers should critically assess the quality, nature, origin and amount of personal data used, reducing unnecessary, redundant or marginal data during the development, and training phases and then monitoring the model's accuracy as it is fed with new data. The use of synthetic data may be considered as one possible solution to minimise the amount of personal data processed by AI applications. **Risk of decontextualization** II.5 The risk of adverse impacts on individuals and society due to de-contextualised data and de-contextualised algorithmic models should be adequately considered in developing and using AI applications. **Independent committees of experts** II.6 AI developers, manufacturers and service providers are encouraged to set up and consult independent committees of experts from a range of fields, as well as engage with independent academic institutions, which can contribute to designing human rights-based and ethically and socially-oriented AI applications, and to detecting potential bias. Such committees may play an especially important role in areas where transparency and stakeholder engagement can be more difficult due to competing interests and rights, such as in the fields of predictive justice, crime prevention and detection.

[223] See also T-PD(2019)01, Guidelines on Artificial Intelligence and Data Protection [GAI]; T- PD(2017)1, Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data.

	III.7 Appropriate mechanisms should be put in place to ensure the independence of the committees of experts mentioned in Section II.6. **Participation and democratic oversight on AI development** II.7 Participatory forms of risk assessment, based on the active engagement of the individuals and groups potentially affected by AI applications, should be encouraged. III. 8. Individuals, groups, and other stakeholders should be informed and actively involved in the debate on what role AI should play in shaping social dynamics, and in decision-making processes affecting them. **Human oversight** II.8 All products and services should be designed in a manner that ensures the right of individuals not to be subject to a decision significantly affecting them based solely on automated processing, without having their views taken into consideration. **Freedom of choice** II.9 In order to enhance users' trust, AI developers, manufacturers and service providers are encouraged to design their products and services in a manner that safeguards users' freedom of choice over the use of AI, by providing feasible alternatives to AI applications. **Algorithm vigilance** II.10 AI developers, manufacturers, and service providers should adopt forms of algorithm vigilance that promote the accountability of all relevant stakeholders throughout the entire life cycle of these applications, to ensure compliance with data protection and human rights law and principles. **Transparency and expandability** II.11 Data subjects should be informed if they interact with an AI application and have a right to obtain information on the reasoning underlying AI data processing operations applied to them. This should include the consequences of such reasoning. **Right to object** II.12 The right to object should be ensured in relation to processing based on technologies that influence the opinions and personal development of individuals. **Accountability and vigilance** III,2 Without prejudice to confidentiality safeguarded by law, public procurement procedures should impose on AI developers, manufacturers, and service providers specific duties of transparency, prior assessment of the impact of data processing on human rights and fundamental freedoms, and vigilance on the potential adverse effects and consequences of AI applications (hereinafter referred to as algorithm vigilance). **Freedom of human decision makers** III. 4. Overreliance on the solutions provided by AI applications and fears of challenging decisions suggested by AI applications risk altering the autonomy of human intervention in decision-making processes. The role of human intervention in decision-making processes and the freedom of human decision makers not to rely on the result of the recommendations provided using AI should therefore be preserved.

	Prior assessment III.5. AI developers, manufacturers, and service providers should consult supervisory authorities when AI applications have the potential to significantly impact the human rights and fundamental freedoms of data subjects. **Cooperation** III.6. Cooperation should be encouraged between data protection supervisory authorities and other bodies having competence related to AI, such as: consumer protection; competition; anti- discrimination; sector regulators and media regulatory authorities. **Digital literacy, education and professional training** III.9. Policy makers should invest resources in digital literacy and education to increase data subjects' awareness and understanding of AI applications and their effects. They should also encourage professional training for AI developers to raise awareness and understanding of the potential effects of AI on individuals and society. They should support research in human rights-oriented AI.
Recommendation CM/Rec(2019)2 of the Committee of Ministers of the Council of Europe to member States on the protection of health-related data	Processing of health-related data should always aim to serve the data subject or to enhance the quality and efficiency of care, and to enhance health systems where possible, while respecting individuals' fundamental rights **Interoperability** 1. […] It therefore highlights the importance of developing secure, **interoperable** information systems **Professional standards** 4.4 Data controllers and their processors who are not health professionals should only process health-related data in accordance with rules of confidentiality and security measures that ensure a level of protection equivalent to the one imposed on health **professionals**. **Consent withdrawal** 5.b Health-related data may be processed if the data subject has given their **consent**, except in cases where law provides that a ban on health-related data processing cannot be lifted solely by the data subject's consent. Where consent of the data subject to the processing of health-related data is required, in accordance with law, it should be free, specific, informed and explicit. The data subject shall be informed of their right to withdraw consent at any time and be notified that such withdrawal shall not affect the lawfulness of the processing carried out on the basis of their consent before withdrawal. It shall be **as easy to withdraw consent as it is to give it**. **Right not to know** 7.6 The data subject is entitled to know any information relating to their genetic data, subject to the provisions of principles 11.8 and 12.7. Nevertheless, the data subject may have their own reasons for not wishing to know about certain health aspects and everyone should be aware, prior to any analysis, of the possibility of not being informed of the results, including of unexpected findings. Their wish not to know may, in exceptional circumstances, have to be restricted, as foreseen by law, notably in the data subject's own interest or in light of the doctors' duty to provide care.

	Transparency 11.3. Where necessary and with a view to ensuring fair and transparent processing, the information must also include: […] - the **existence of automated decisions,** including profiling, which is only permissible where prescribed by law and subject to appropriate safeguards. **Interoperability** 14.1. Interoperability may help address important needs in the health sector and may provide technical means to facilitate the updating of information or to avoid storage of identical data in multiple databases, and contribute to data portability. 14.2. It is, however, necessary for interoperability to be implemented in full compliance with the principles provided for in this Recommendation, in particular the principles of lawfulness, necessity and proportionality, and for data protection safeguards to be put in place when interoperable systems are used. 14.3. Reference frameworks based on international norms offering a technical structure which facilitates interoperability should guarantee a high level of security while providing for such interoperability. The monitoring of the implementation of such reference frameworks can be carried out through certification schemes. **Scientific research integrity** 15.10. Where a data subject withdraws from a scientific research project, their health-related data processed in the context of that research should be destroyed or anonymised in a manner which does **not compromise the scientific validity of the research** and the data subject should be informed accordingly.
Recommendation CM/Rec(2016)8 of the Committee of Ministers to the member States on the processing of personal health-related data for insurance purposes, including data resulting from genetic tests	8. The processing for insurance purposes of health-related personal data obtained in a research context involving the insured person should not be permitted.
Recommendation CM/Rec(2010)13 of the Committee of Ministers of the Council of Europe to member States	**Risk of re-identification** 8.5. Suitable measures should be introduced to guard against any possibility that the anonymous and aggregated statistical results used in profiling may result in the re-identification of the data subjects.[224]

[224] See also Convention 108+. Explanatory Report, 19 and 20 ("Data is to be considered as anonymous only as long as it is impossible to re-identify the data subject or if such re-identification would require unreasonable time, effort or resources, taking into consideration the available technology at the time of the processing and technological developments. Data that appears […] When data is made anonymous, appropriate means should be put in place to avoid re-identification of data subjects, in particular, all technical means should be implemented in order to guarantee that the individual is not, or is no longer, identifiable. They should be regularly re-evaluated in light of the fast pace of technological development").

on the protection of individuals with regard to automatic processing of personal data in the context of profiling	
UNESCO. 2019. Preliminary Study on a Possible Standard-Setting Instrument on the Ethics of Artificial Intelligence	[Principles-based approach] • Diversity, inclusion and pluralism (including a multilingual approach should be promoted) • Autonomy • Explainability • Transparency • Awareness and literacy • Responsibility • Accountability • Democracy ("AI should be developed, implemented and used in line with democratic principles") • Good governance ("Governments should provide regular reports about their use of AI in policing, intelligence, and security") • Sustainability • Human oversight • Freedom of expression (including universal access to information, the quality of journalism, and free, independent and pluralistic media, avoiding the spreading of disinformation)
OECD. 20 19. Recommendation of the Council on Artificial Intelligence	[Principles-based approach] • Human-centred values and fairness • Transparency and explainability (**awareness** of the interactions with AI systems; understanding of AI outcome; enabling those adversely affected by an AI system to **challenge** its outcome based on plain and **easy-to-understand information** on the factors, and the **logic** that served as the basis for the prediction, recommendation or decision) • Robustness, security and safety (not pose **unreasonable safety risk**; **traceability**, including in relation to datasets, processes and decisions made during the AI system lifecycle; **risk management approach** to each phase of the AI system lifecycle on a continuous) • Accountability

40th International Conference of Data Protection and Privacy Commissioners. 2018. Declaration on Ethics and Data Protection in Artificial Intelligence [ICDPPC]	[Principles-based approach] • Continued attention and vigilance ("establishing demonstrable governance processes for all relevant actors, such as relying on trusted third parties or the setting up of independent **ethics committees**") • Transparency and intelligibility (**explainable AI**, algorithmic **transparency** and the **auditability** of systems, **awareness** of the interactions with AI systems; adequate information on the purpose and effects of AI systems, **overall human control**) • Risk assessment and privacy by default and privacy by design approach ("assessing and documenting the expected **impacts on individuals and society** at the beginning of an artificial intelligence project and for relevant developments during its **entire life cycle**") **Public engagement** **Mitigation of unlawful bias and discrimination**

TITLE II. NATIONAL PERSPECTIVES OF AI SYSTEMS REGULATION

CHAPTER I. Harnessing Innovation: Israeli Perspectives on AI Ethics and Governance

Prof. Isaac Ben-Israel[225], Prof. Eviatar Matania[226] , Leehe Friedman[227]

I. Executive Summary

This article sets forth the current state of play in Israel's policy development, with respect to the opportunities and challenges presented by artificial intelligence (AI) in relation to human rights and ethics. It is based, to a large extent, on the report of Israel's National Initiative for Secured Intelligent Systems, which has been recently submitted to the Israeli government. The present survey describes Israel's unique approach in attempting to leverage opportunities presented by AI while addressing the challenges that it poses. This article outlines how Israel's governance approach thus far seeks to balance the need to enable innovation, both in the public and private sectors, with moral and human rights imperatives which are omnipresent in AI developments.

Israeli policy-makers tend to view AI developments not just as a disruptive but as a transformative: AI technology is seen as critical to the welfare, economy and security of Israel's citizens. Taking this as the starting point, the priority for Israel has been to establish a holistic and sustainable secured AI ecosystem, driven by the private sector but in which government, private industry and academia all participate, and which supports the use of AI at all levels. Bearing this in mind, this article highlights the key challenges that have been identified by policy makers, in Israel and abroad, in the fields of human rights, democracy and the rule of law – security, privacy, autonomy, civil and

---

[225] Professor Isaac Ben-Israel is a retired Major General who joined Tel-Aviv University in 2002, where currently he is the Director of the Blavatnik Interdisciplinary Cyber Research Centre and the Yuval Ne'eman's Workshop for Science, Technology & Security. Outside the university he is the Chairman of Israel Space Agency. In 2010, he was appointed by the PM to lead a task force that led the Cyber Revolution in Israel. During the last two years Prof. Ben-Israel has been co-chairing Israel's National Initiative for Secured Intelligent Systems (AI) to recommend the PM and the government about a national plan to promote Israel as a global power in AI.

[226] Eviatar Matania is a professor at the School of Political Sciences, Government and International Affairs at Tel-Aviv University, where he heads the MA program of Security Studies and the MA program of cyber politics and government. Matania is also an Adjunct Professor at Oxford's Blavatnik School of Government, where he convenes the Cyber Module. Matania was the founding head and former Director General of the Israel National Cyber Directorate (INCD) in the Israeli Prime Minister office, where he reported directly to the Prime Minister, and was responsible for Israel's overall cyber strategy, policy and its implementation to defend the Israeli civilian sector. During the last two years Prof. Matania has been co-chairing Israel's National Initiative for Secured Intelligent Systems (AI) to recommend the PM and the government about a national plan to promote Israel as a global power in AI.

[227] Leehe Friedman is Adjunct Professor and the Director of the Honors Track in Strategy & Decision Making at the Lauder School of Government, Diplomacy and Strategy at IDC Herzliya. During the last two years Ms. Friedman has been the Coordinator of Israel's National Initiative for Secured Intelligent Systems (AI) to recommend the PM and the government about a national plan to promote Israel as a global power in AI.

political rights, safety, fairness – including fair competition – and accountability. Israel's proposed approach in response to these challenges, according to the National Initiative's Report, breaks new ground. While it is firmly anchored in established governance principles and international AI policy best practices, it nonetheless represents a novel governance approach, focusing on balanced regulation to foster innovation. To that effect, it proposes original policy tools, such as risk assessment tool that match different regulatory approaches based on the risk level associated with a particular activity, and a dynamic frequency map that helps locate challenging areas in term of applying ethical values to the a particular AI system's development.

It signals also the need for engagement with countries and international forums, to learn from and contribute to international processes involving questions of AI, ethics, law and governance.

This article is not intended as an official government paper, and does not necessarily reflect Israeli government policy. Its authors are writing in their personal capacity, though they received relevant information from various government officials.

Acknowledgments

The authors would like to thank Cedric Sabbah, Director of International Cybersecurity & IT Law at the Office of the Deputy Attorney General (International Law), within Israel's Ministry of Justice, Prof. Karin Nahon, Head of the Ethics and Regulation working group of the National Initiative for Secured Intelligent Systems, Dr. Roy Schöndorf, Deputy Attorney General (International Law), and Dror Ben Moshe, Big Data Department Manager at the Ministry of Health, for their valuable contribution, professional guidance and insightful comments.

II. Introduction

Israel perceives AI as a core emerging technology and as "*an infrastructure of infrastructures, one that is critical to the future of the State of Israel – to its security, its economy and to the welfare of its population*".[228] AI applications, due to their potential to enhance availability, reliability and efficiency of national infrastructures, services and systems, at lower costs to the state and its citizens, hold key roles in Israel's capacity to meet some of its national challenges in the 21st century.

Looking ahead, AI is likely to fundamentally transform all aspects of private and public life. In order to harness the positive potential of AI technologies, Israel strives to establish a holistic and sustainable AI ecosystem that includes the government, private industry and academia. A feedback loop involving these three sectors would benefit society as a whole by: (a) increasing the use of AI applications; (b) enhancing the work of the government and the services it provides; (c) fostering the economy and innovation of new techno-scientific developments which in turn would increase again the demand for new AI applications.

Accordingly, the Israeli approach towards AI is based on two complementary efforts:

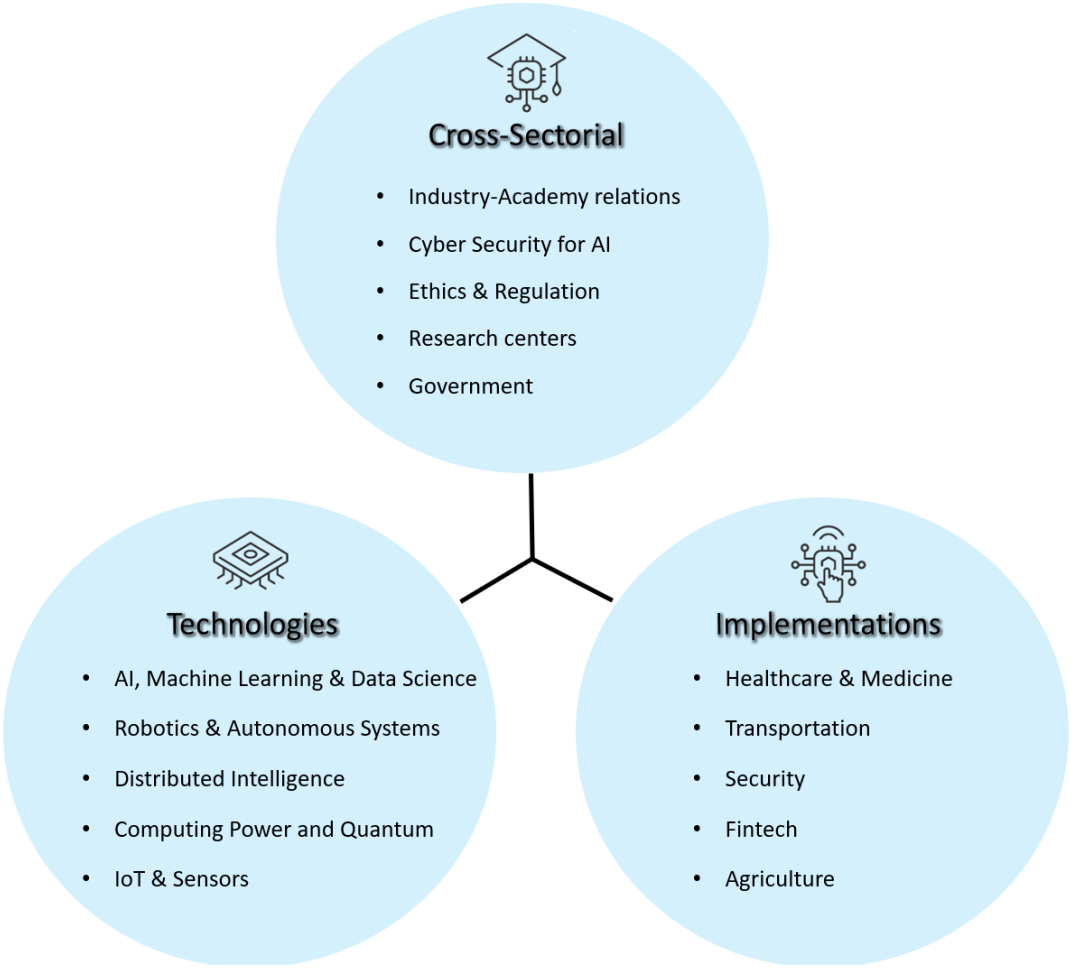
[228] Ben-Israel, I., Matania, E. & Friedman, L. (Eds.) (Sep. 2020). *The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: A National Strategy for Israel. Special Report to the Prime Minister*. (Hebrew) p.3.

i. Promoting a wide and fair use of AI applications both in the public and private sectors.

ii. Fostering a leading technological industry that would develop AI-based solutions for emerging challenges in Israel and around the globe.

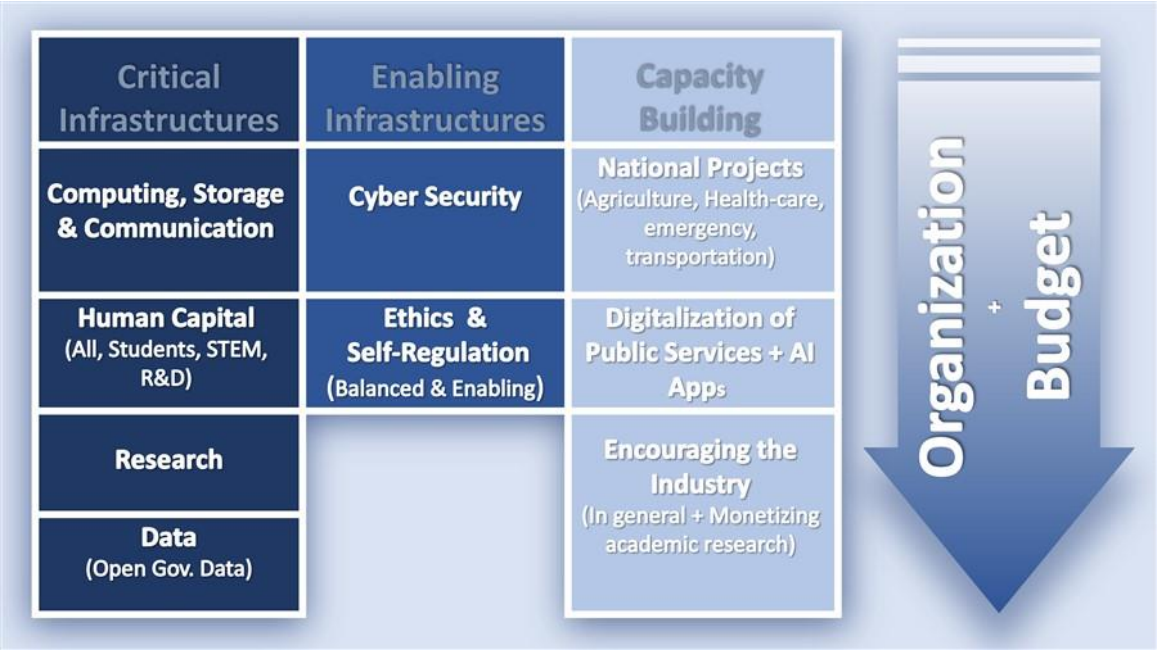
Israel's National Initiative for Secured Intelligent Systems

In order to understand Israel's current AI landscape and examine how its broad policy goals can be achieved given the characteristics of Israel society, Israel's Prime Minister launched in 2018 the National Initiative for Secured Intelligent Systems, and appointed two of the authors of this article, Prof. Isaac Ben-Israel and Prof. Eviatar Matania, to co-lead the initiative. Its mandate was to generate a national plan in the field of AI and related intelligent technologies. The work initiative used a multistakeholder approach: hundreds of Israeli experts in various domains and from the academic, industrial and governmental sectors volunteered to take part in this endeavor. The experts were divided into 15 working groups dealing with various technological, sectorial and cross-sectorial aspects of intelligent systems, according to the following model[229]:



[229] Ibid. p.15.

Each working group analyzed the Israeli AI environment according to its thematic perspectives. The working groups used as a comparative point previous and cutting-edge work that has been conducted by other jurisdictions, in order to present a national plan customized to the specific characteristics of Israel. Conclusions and recommendations were integrated into a final report which proposes a National Strategy for Israel in the field of Secured Intelligent Systems (the "**National Initiative Report**"). It defines intelligent technologies as a national priority and draws an operative national plan for the establishment of a sustainable eco-system in the field of secured intelligent systems. The national plan is based on three layers: (1) critical infrastructures; (2) enabling infrastructures; (3) capacity building; and consists of the following building blocks:[230]



The National Initiative Report has been recently submitted to the Israeli Prime Minister.

III. AI applications in Israel – A public policy opportunity

"Israel is now number three in the world for AI solutions. With only 8.5 million citizens, Israel has a market share of 11% and is equal to China. Israel has 40x more AI companies per capita than the market leader USA, and that makes Israel the clear hidden champion of Artificial Intelligence".[231]

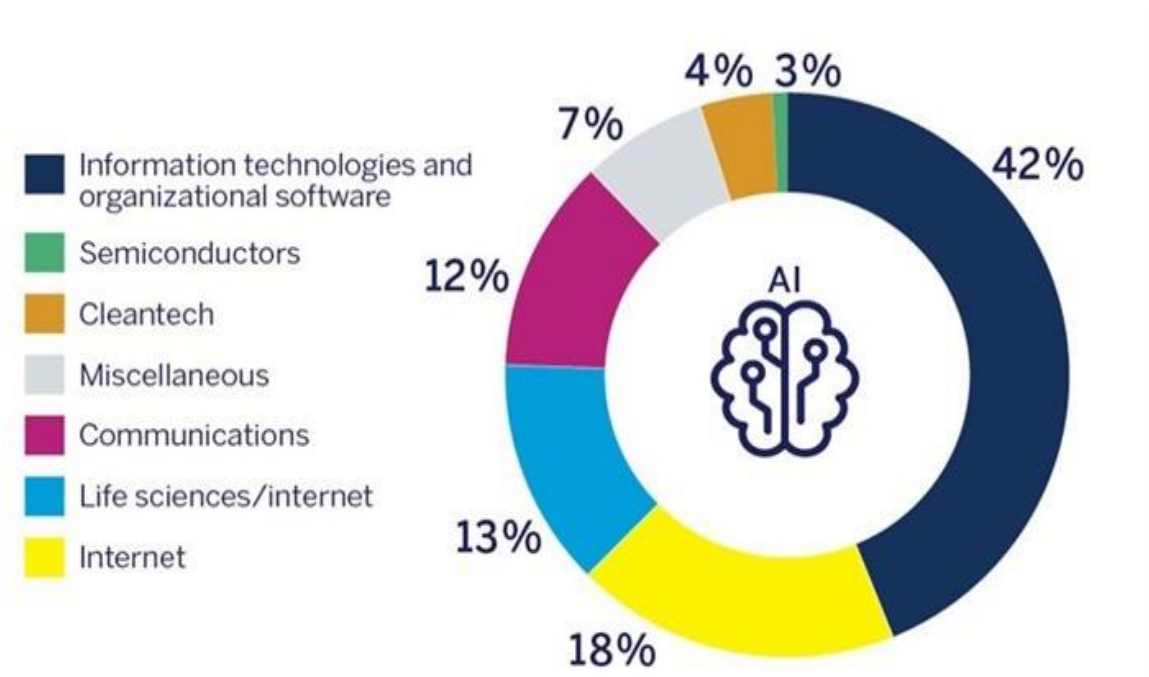
Israel has a strong high-tech and innovation ecosystem coupled with a culture that embraces and adapts to technological developments. The prevailing atmosphere in the "Startup Nation" is one that encourages both the public and the private sectors to explore and use AI applications in various fields. However, the AI applications landscape in Israel is shaped, first and foremost, by the private market.

[230] Ibid. p.23.
[231] ASGARD. *The Global Artificial Intelligence Landscape*. Retrieved from https://asgard.vc/global-ai/7

i. The private sector

According to the Israel Innovation Authority, investments in Israeli high-tech AI projects increased in 2011-2019 by a factor of 12.5, from 305 million dollars to 4 billion dollars. In 2019, 42% of the total sum invested in Israeli high-tech went towards AI technologies.[232] Around 1,400 AI companies are currently operating in Israel, developing and utilizing AI technologies in various domains such as business analysis, cyber and healthcare applications and more. Over 40% of the companies deal with information technologies and organizational software, while 30% focus on internet services and communications[233].

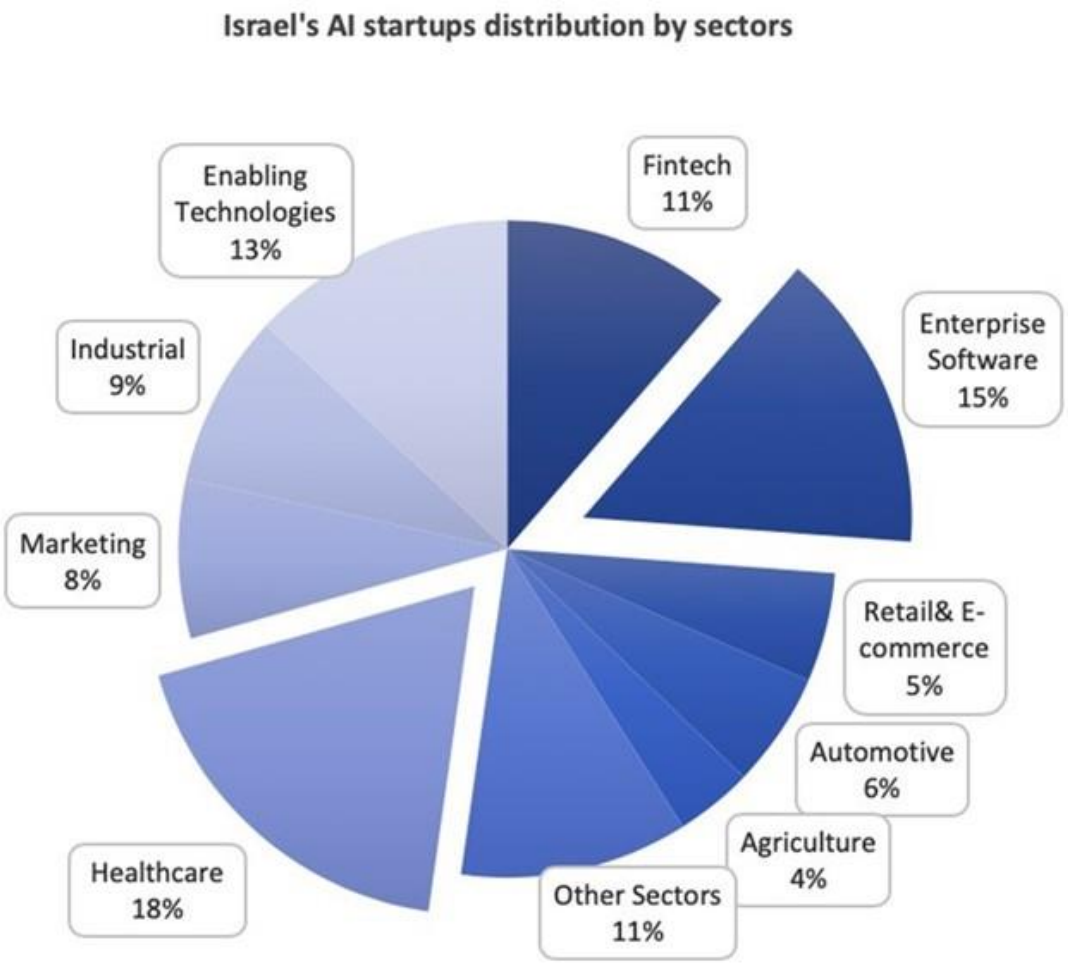
Distribution of AI companies by sector



Source: Israel Innovation Authority (2020). Bolstering Artificial Intelligence

1,024 of these companies are startups. Despite Israel's small size and limited resources, it ranks third in the world in terms of the number of AI startups, after the United States and China, and first in terms of the number of AI companies per capita.[234] In the past five years, an average of 140 new startups have emerged annually, offering applications and products which cover all sectors and areas of life. However, the leading sector is healthcare with 188 startups (18%) offering AI solutions in the fields such as diagnostics, monitoring, disease management, personalization and clinical workflow. Enterprise software closely follows with 152 startups (15%) developing and utilizing AI products and services in the fields of sales and training, HR, data and intel, customer and support, development and IT, management and teamwork, security, privacy and finance.[235]

[232] Israel Innovation Authority. (2020). *Bolstering Artificial Intelligence*. Retrieved from https://innovationisrael.org.il/en/reportchapter/bolstering-artificial-intelligence-0#footnoteref2_bscx727
[233] Ibid; Israel Innovation Authority. (2020). *Innovation in Israel - 2019 Innovation Report*. p.62 (Hebrew).
[234] ROLAND BERGER GMBH & ASGARD. (2018). Artificial Intelligence – A strategy for European startups. P. 17
[235] Cardumen Capital (June 9, 2020) Israel's Artificial Intelligence Startups, June 2020. *Medium*. retrieved from https://medium.com/@cardumencapital/israels-artificial-intelligence-startups-june-2020-81e27d9332d8



Source: Cardumen Capital. Israel's Artificial Intelligence Startups, June 2020. *Medium*

Israel also ranks high in the number of companies that develop infrastructure technologies for AI such as special-purpose chips, infrastructure algorithms, and complex systems for the acceleration of computing.[236]

For a full map of Israel's AI startup landscape in 2020 segmented by sectors and applications, see Annex I.

ii. Government initiatives and policy

In addition to the activity of the private sector, the Israeli government has important role in promoting AI applications. It can do so by initiating projects itself, or by creating an encouraging environment for the private sector to further develop and use AI technologies, for example through enabling regulation, incentives for the industry, etc. The great potential which the government attributes to AI technologies and their applications originates in two characteristics of Israeli society:

[236] Israel Innovation Authority. (2020). *Bolstering Artificial Intelligence*. Retrieved from https://innovationisrael.org.il/en/reportchapter/bolstering-artificial-intelligence-0#footnoteref2_bscx727

- **High population growth rate –** Israel's demographics are exceptional for developed economies and as such they present its government with some unique challenges. With 3.1 children per women, Israel has the highest fertility rate among the countries of the OECD, of which the average fertility rate is 1.6 children.[237] Furthermore, the average life expectancy in Israel is of 82.9 years, the fifth highest within the OECD.[238] Consequently, the annual population growth rate in Israel – 1.9%[239] – is almost four times higher than the average population growth rate of OECD member countries (0.54%).[240] The consistent fast growth of the population requires the State of Israel to adjust its public and social services and to maintain and increase accordingly its infrastructures in all fields of life (e.g. healthcare, transportation, education, energy, etc.).

- **Population density and overloads on infrastructures –** the majority of the Israeli population lives in the center of the country, and over 40% of the population is spread over less than 7% of the country's territory.[241] This leads to severe overloads on the infrastructures and services in highly populated areas. One prominent example is the growing traffic congestion in the center of the country, which have negative ramifications on productivity, the environment and the number of accidents and casualties

The high population growth rate and the population density shape Israel's approach to AI, as they significantly increase the make the Israeli demand for national infrastructures and services. The government's motivation to increase the use of AI, beyond the areas that are already covered by the private sector, lays in the technology's potential to answer the growing need to enhance availability, reliability and efficiency of public services and national infrastructures, at lower costs to the state and its citizens. For this reason, among the first sectors for which the government promotes AI solutions are healthcare and transportation, where the overload on current infrastructures is most acute.

However, governmental ministries in Israel differ in their readiness to embrace AI applications due to variance in digital maturity and in some cases even digital gaps. Israel's National Digital Initiative, also called "Digital Israel", is the government body responsible for e-government services. Digital Israel spearheads government efforts for digital transformation, to reduce socioeconomic gaps, promote economic prosperity, and create a smarter, friendlier government. Its scope of activity encompasses a broad array of e-government services, at all levels of government (including municipalities). Thus, Digital Israel works with other government ministries, assisting them in developing and deploying digitization plans. Digital Israel also leads the government initiatives plans for Smart Cities as well as the National Plan for Digital Literacy.

To a very large extent, the fulfillment of Digital Israel's mandate depends on the availability and transferability of data. Indeed, in order to maximize the full potential of digital transformation, government bodies must be able to collect large amounts of information, combine it with data from other sources, and deploy technical tools to analyze the data and draw conclusions. In many cases, the data that must be collected and shared includes personally identifiable information ("PII"). In addition, the software tools that can be used include big data analysis,

[237] OECD (2020), Fertility rates (indicator). Retrieved from https://data.oecd.org/pop/fertility-rates.htm#indicator-chart
[238] OECD (2020), Life expectancy at birth (indicator). Retrieved from https://data.oecd.org/healthstat/life-expectancy-at-birth.htm
[239] Central Bureau of Statistics. (2020). Israel in Figures Selected Data From the Statistical Abstract of Israel 2019. Retrieved from https://www.cbs.gov.il/he/publications/DocLib/isr_in_n/isr_in_n19e.pdf p.6.
[240] World Bank, Population Growth for OECD Members [SPPOPGROWOED]. Retrieved from FRED, Federal Reserve Bank of St. Louis https://fred.stlouisfed.org/series/SPPOPGROWOED
[241] Central Bureau of Statistics. (2020). *Statistical Abstract of Israel 2019 - No.70.* Retrieved from https://www.cbs.gov.il/he/publications/DocLib/2019/Shnaton70_mun.pdf (Hebrew). p.21.

much of which can be enhanced by machine learning. Thus, the projects that Digital Israel wishes to implement face a key challenge, namely, how to balance between the data needs, on the one hand, and the legal and ethical considerations on the other.

To date, there is no all-encompassing government policy to address this challenge. Such a policy is currently being finalized. Pending its adoption, Digital Israel's activities are informed by the existing legal framework, which includes constitutional human rights protections (privacy, non-discrimination, freedom of expression) as well as administrative law rules and principles applicable generally to all government bodies (transparency, accountability, fairness, due process and reasonableness).

Below is a description of the main projects involving AI that are in the process of development (in each case, in conjunction with the relevant government ministry). In the development of each of those projects, Digital Israel has worked with in-house counsel and Ministry of Justice constitutional counsel, to ensure that the development and deployment of each project complies with applicable constitutional and administrative law limitations

Digital health

The project has ambitious goals, including:[242]

- Customized treatment: promoting research, development and implementation of tools that allow the patient to receive the best and most personalized treatment;
- Promoting health and patient prevention through use of digital tools in a way that shifts the focus from patient care to preventive medicine;
- Sustainable health: promoting the development and implementation of systems that increase the operational and managerial effectiveness of the health system, in a way that frees up existing resources;
- Development and implementation of digital tools that streamline communication between the Ministry of Health and those it serves;
- Delivery of emergency treatment services through an appointment management system and an application informing the patient on the progress of the treatment. The information collected will enable better management of resources to avoid congestion in emergency rooms;
- Sharing clinical information across service provider platforms has been expanded, to connect different service providers and allows them to view treatments and diagnoses made by other health professionals in different organizations.
- Some of these goals are already being implemented. By a government resolution,243 the Ministry of Health has created a platform called "TIMNA", which grants third-parties controlled access to health data in order to promote applied research. The data includes vast quantities of health records, gathered by hospitals and clinics from around the country, providing an invaluable resource. Access to the data is subject to strict privacy and ethical restrictions. First, the institutions, researchers and start-ups seeking access to this data must provide Helsinki committee approval for their project. They are required to identify the specific types

[242] Ministry of Health digital services home page (Hebrew) https://www.health.gov.il/About/projects/DigitalHealth/Pages/default.aspx.
[243] Israel Government Resolution 3709, "National plan to advance digital health as a means to improve health and foster growth " August 23, 2018, https://www.gov.il/he/Departments/policies/des3709_2018 (Hebrew).

of data that they need, and only that data is provided. They must sign privacy commitments. All personal data is anonymized. The research takes place entirely within the digitized platform – no personal data can be extracted from the platform, further protecting confidentiality. Furthermore, before the research is published and an algorithm is used, the Ministry of Health reviews it to ensure that no personal information is used or disclosed. Thus far, the TIMNA projects making use of AI are as follows:

- The Israel Center for Disease Control applies an AI algorithm to review diagnostic forms of patients and verify their cancer diagnosis. This saves significant amounts of time, as it automates the process of reviewing over 100,000 forms a year. Audits are conducted to ensure that there are no false negatives.
- Similarly, an algorithm is being developed to assist with medical follow-ups in two areas: child development and pregnancy. In both cases, the algorithm analyzes in real time status reports, diagnoses and notes and recommendations of doctors and nurses, comparing them against standard protocols. It then alerts the hospital or clinic of potential errors, misdiagnosis, or issues that might require additional testing or follow-up. The system is geared towards assisting health professionals in catching mistakes and does not entail significant ethical risks to the patient.
- Another field of study is the use of AI to analyze of medical images (MRI, CT etc.). Thus far the results have been promising, in that the algorithms have been able to detect cases that were missed by doctors. The intent is not to supplant the doctor's decision-making but rather to streamline the process and assist him/her in analyzing the images.
- Finally, the Ministry of Health has deployed AI algorithms to assist with its efforts in slowing the spread of COVID-19. Often, the epidemiological study based on discussions with an infected individual are incomplete, due to failure to remember all locations visited, the interviewer's failure to enter all the information correctly, or other human error. The algorithms are used to form a more complete picture of the likely course of previous infections and predict future infections. This information is then used to inform government policy with respect to closure measures at a general scale. It should be noted that the information is not used to make decisions about specific individuals or communities.

Transportation

- The Ministry of Transportation is establishing a pilot project to enable testing of autonomous vehicles. The project would allow manufacturers to apply for a special license, under which they may test their product in real-world conditions, in low-risk driving environments. To that effect, the Ministry published a draft bill, which is open for public comments.[244] The draft bill does not yet contain all the rules that will apply to trials or to the requirements of an autonomous driving system (capabilities, safety, and oversight). These issues, as well as ethical issues that could arise, will be addressed at a later stage. This approach reflects a cautious and incremental innovation philosophy: in order to understand the impacts of a technology, and given the risks to human life, the trials are permitted in an environment that is not "controlled" but that presents a relatively low risk. This should enable policy-makers to make adjustments before moving forward with larger scale experimentation.

[244] The text of the draft bill can be found here: https://www.nevo.co.il/law_word/law11/200820-2.doc (Hebrew).

- Progress is also being made in the field of public transportation as innovative solutions are being developed to reduce traffic congestion[245].

Taxation

Israel's Tax Authority launched a project to assist investigators in detecting tax fraud. The project uses AI tools to predict the likelihood of tax fraud, based on certain indicators. Privacy concerns were central to how the project was designed. Indeed, the project is based on a layered approach for access to information: initially, few indicators and little information is used to flag risks of fraud; if the initial investigation suggests a higher risk of fraud, only then is personal information required in order to determine with greater certainty the identity of the potential offender.

Proposals for national projects

In addition to the aforementioned unfolding projects, the National Initiative Report recommends that the government launch, in cooperation with the industrial and the academic sectors, four more national projects in the fields of healthcare, transportation, security and agriculture. All four were conceived to answer genuine national needs deriving from developments withing Israeli society which trigger demands for improved and novel infrastructures and services.

- **Healthcare** – Reforming the national healthcare system by improving the quality and availability of medical services, and relieving the overloads on hospitals by launching a national system based on intelligent technologies for: (i) remote patient management; (ii) more efficient triage and treatment in emergency medicine; (iii) generating comparative quality indices for measuring clinical outcomes.[246]

- **Transportation** – *"Of the many ways in which intelligent systems can solve acute problems in the field of transportation, we chose to recommend, at the first stage, the installation of Smart Traffic Lights in an entire pilot metropolitan area, with the intent to address traffic congestion, which is the most severe and acute transportation problem in Israel."*[247]

- **Security** – Creating a national dual-system that will harness the potential of intelligent technologies to improve predicting abilities and decision-making processes, for better management on the national level. By collecting and analysing data for civilian applications of command and control in normal times, it will enhance the national capacity to prepare for times of emergency (natural disasters, epidemics and security threats from enemies) and to make decisions during crises.[248]

- **Agriculture** – "*In an age when food security, water management and other areas in agriculture become acute global challenges, […] We recommend promoting a national project to develop an intelligent technologies based system for early detection of pests and diseases in agricultural crops; alongside integration of intelligent systems into the agricultural sector for optimizing the use of nature resources and [other] inputs to ensure optimal food production."*[249]

---

[245] Ben Dror, M. and Azaria, M. (July 24, 2020). Israel's 'smart commuting' shows what public transport could be like after COVID-19. *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2020/07/israel-smart-commuting-after-covid-public-transport-innovation/

[246] Ben-Israel, I., Matania, E. & Friedman, L. (Eds.) (Sep. 2020). *The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: A National Strategy for Israel. Special Report to the Prime Minister*. (Hebrew) p.35.

[247] Ibid. P.37

[248] Ibid.

[249] Ibid. P.36

- Finally, acknowledging the variance in digital maturity within the government, the National Initiative included a working group dedicated to the government sector, in order to set guidelines for preparing the entire government to the age of AI. The working group assessed the required organizational, technological and regulative measures to foster implementations of AI applications within the government, in order to improve both the inter-ministerial work and the interactions between the government and the citizens.[250]

IV. Risks and challenges posed by AI in the fields of human rights, democracy and the rule of law

The topics of human rights and ethics were addressed by the National Initiative, through a dedicated working group that studied the issue in depth from a regulatory and ethical perspectives. The working group's report ("**Ethics Report**")[251] identifies issues that are novel and unique to AI and the ways in which it is expected to affect our lives. In light of these issues, it sets forth a series of ethical challenges posed by the technology regarding human rights, democracy and the rule of law.

Below is an outline of the issues identified by the Ethics Report, followed by an overview of the risks and ethical challenges that it suggests addressing.

i. What Is New and Special about AI?

a) *"AI systems tend to radicalize existing social relations. For example, if there is inequality between different social groups, AI systems can reproduce and even exacerbate it. This is true of discrimination, stereotype, rights violations, political extremism, etc. For the* sake *of convenience, we will demonstrate that claim with regard to inequality. There are several main reasons for that phenomenon:*

- *Since AI systems depend on the information provided to them, their input can reflect inequality that already exists, and if the data entered have been manipulated, the system will learn that manipulation.*
- *AI systems are becoming increasingly common in a growing number of social contexts. Therefore, their impact – and potential biases – affect larger audiences.*
- *There is an erroneous tendency to treat the products of AI systems, which analyze data quickly and on a large scale, as scientific truth. Consequently, there is the danger that such systems would not be subject to the controls applied to equivalent human decisions, when a bias is suspected.*
- *Due to the systems' complexity, it is difficult to anticipate and validate their behavior in advance. Consequently, it is often hard to distinguish between "true" diagnosis based on a valid review and monitoring process, as done with regard classical algorithms or human decision-making, and a biased diagnosis."*[252]

[250] Sharvit, S. et al. (2020). Government Working Group Report. In Ben-Israel, I., Matania, E. & Friedman, L. (Eds.). *The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: A National Strategy for Israel. Special Report to the Prime Minister*. pp. 214-227. (Hebrew).
[251] Nahon K., Ashkenazi A., Gilad Bachrach R., Ken-Dror Feldman D., Keren A. and Shwartz Altshuler T. (2020). Working Group on Artificial Intelligence Ethics & Regulation Report. In Ben-Israel, I., Matania, E. & Friedman, L. (Eds.). *The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: A National Strategy for Israel. Special Report to the Prime Minister*. pp. 172-119. (Hebrew).
[252] Ibid. p. 179.

b) *The procedural challenge: How to "engineer" values*. This issue arises in areas where AI systems are developed to replace human decision-makers who are skilled and authorized to apply normative considerations. "*When developing AI systems that replace human discretion, the responsibility for these normative considerations is transferred from professionals such as doctors and lawyers to engineers and information scientists, which does not occur as often when dealing with classical algorithms*." [253]

c) Privacy and autonomy risks of unprecedented scope and scale. see page 137 and 139 below.

d) *Complexity that erodes public trust*. Lack of clarity and public understanding of how AI systems operates and how it affects our lives often leads to distrust, which may result in reluctance to embrace the technology, even in areas where AI systems offer a clear business – and social – benefit. The report mentions in this regard, the assessment of the EU's High-Level Expert Group on Artificial Intelligence.[254]

e) *"Unfair economies of scale. […] powerful players with the big data required to develop AI systems take advantage of internet economies of scale to shape the way new players enter the market, with a negative effect on competitiveness. When it comes to completely new players, the fact that they lack the amount of data required could mean they are in effect barred from the AI market*." [255]

f) *"Changes in familiar warranty categories. The ability to collect and process data through products* [*IoTs*] *enables companies to offer new related services, but also raises new questions about the warranty for these services, and the division of responsibility between the producer and those providing the services in practice. AI-integrated products, in particular, also include the combination of a physical product and remote computability and operability. Thus, the classical division between product and service and product warranty and service warranty needs to be reexamined. Things become even more complex when such products and services are used by other business entities. For example, when a grocery chain uses a drone for deliveries. The drone is capable of flying, navigating and dealing with the environment. In addition, it provides mapping and weather forecast services. All these are acquired by a grocery chain, for the modest purpose of delivering groceries."* [256]

ii. Ethical risks and Challenges

The Ethics Report address the following ethical challenges. It should be noted that it relies extensively on the EU's High-Level Expert Group on Artificial Intelligence in elaborating upon the different ethical principles.

Security

The right to life and security is often overlooked in global discussions involving human rights. And yet, it remains the most fundamental right of all, absent which other human rights cannot be applied. At a basic level, with respect to AI, the Ethics Report underlines the need to secure AI applications and AI-enabling networks and computers. The report makes several important

[253] Ibid. p. 180.
[254] Ethical Guidelines for Trustworthy AI, The High-Level Expert Group on Artificial Intelligence, EU, 2019, https://ec.europa.eu/futurium/en/ai-alliance-consultation. (Hereafter, EU)
[255] Nahon K. et al. (2020). p.180.
[256] Ibid. p.181.

observations in that regard. It notes that information is the "energy that fuels the current wave of AI", such that security of AI applications and networks is a precondition to development and implementation of AI technology[257]. It further notes that the information that can be collected by AI to build and deploy AI tools includes vast amounts of personal and commercial information, including personal, medical, economic and other sensitive information. It notes that even information that appears "non-sensitive" can become sensitive when cross-referenced with other information.

From a policy perspective, this ties security directly with human rights. For example, protecting privacy requires, at a fundamental level, securing private information from malicious cyber operations. Similarly, beyond commercial and performance considerations, protecting data integrity is also a human rights imperative: in order to protect against bias in a particular AI application, the data as well as the algorithms upon which the application is based must not be tampered with. Freedom of expression and access to information are also highly dependent on the security of AI-related applications and networks.

The use of AI further deepens the reliance upon computers, hence creating new vulnerabilities for cyber-attacks. As part of the national initiative, a dedicated working group for cybersecurity in the age of intelligent systems identified new cyber threats presented by AI technologies. First, attacks against AI systems, which can result in damage to the decision-making mechanism, thus leading to false, misleading or biased decisions, or to threats against the system's IP. Second, malicious exploitation of AI capabilities as cyber weapon to launch sophisticated "intelligent attacks". These vulnerabilities of AI systems raise the question – whether and how an AI system can be authenticated as secured and reliable.[258] AI security is thus a basic layer over which richer interactions can take place. Its importance is indeed reflected in the title of the Israeli national AI initiative - the National Initiative for *Secured* Intelligent Systems.

Privacy

Tellingly, privacy is the first and foremost of human rights addressed by the Ethics Report. The report underlines that AI applications "are largely based on information about individuals or on deriving conclusions about them from personally identifiable information".[259] Protection of privacy is largely dependent upon a robust legislative framework. In Israel, this framework consists mainly of Israel's Basic Law: Human Dignity and Liberty (1992),[260] its Privacy Law (1981), several privacy regulations including data transfer regulations[261] and the comprehensive 2017 Protection of Privacy Regulations (Data Security).[262] The legal regime is complemented by extensive case law and a robust judiciary. Over time, a number of privacy protection principles have emerged: the need for legal cause for collecting and processing information (e.g. informed consent), usage limitations, the right to review and correct one's personal information, transparency vis-à-vis the information owner and the obligation to protect the information.

---

[257] Ibid. p.190.
[258] Zack, H. et al. (2020). Working Group on Cyber and Intelligent Systems Report. In Ben-Israel, I., Matania, E. & Friedman, L. (Eds.). *The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: A National Strategy for Israel. Special Report to the Prime Minister*. pp. 168-171. (Hebrew). p.168.
[259] Nahon K. et al. (2020). p.188.
[260] http://knesset.gov.il/laws/special/eng/BasicLawLiberty.pdf. This is a quasi-constitutional law, whose underlying principles are seen as constitutionally mandated, even in the absence of a formal written constitution.
[261] Privacy Protection (Transfer of Data to Databases Abroad) Regulations (2001). See unofficial translation here: https://www.gov.il/BlobFolder/legalinfo/legislation/en/PrivacyProtectionTransferofDataabroadRegulationsun.pdf.
[262] See for unofficial translation here: https://www.gov.il/BlobFolder/legalinfo/data_security_regulation/en/PROTECTION%20OF%20PRIVACY%20REGULATIONS.pdf

Against this backdrop, the Ethics Report notes that there remain gaps between traditional conceptions of privacy protection and the challenges raised by AI. Indeed, to the extent that AI relies on the collection and processing of PII, it can be expected that novel privacy issues will arise, which may require adapting existing privacy laws further down the road.

The Ethics Report also observes that in certain cases, there might arise a conflict between privacy and fairness. If individuals belonging to a certain group refrain from sharing their personal information with an AI application, that application will not be able to draw from data that takes this group into account, potentially leading to greater discrimination.[263] There is thus a policy imperative to enabling the collection of PII while ensuring that such data will be both secured and subject to robust privacy protections.

Autonomy

The Ethics Report defines autonomy as "the individual's ability to make intelligent decisions, including the prevention of unfair or unconscious influence on individual behavior."[264] In human rights terms, this can refer to concepts such as human dignity and the right of access to information. The Ethics Report states[265]:

"Autonomy is based not only on an individual's ability to choose among options, but also on the availability of the information allowing cogent choice and assessing its reliability. These issues cannot be taken for granted in the AI era. Moreover, the ability to conduct in-depth analysis of information about a person enabled by AI makes it possible to devise highly intrusive persuasion attempts, again with potential implications that are not fully understood as yet.

Autonomy is also related to the range of human decisions involved in interaction with technology, which technology might narrow. We must therefore always examine whether a given application affects autonomy and how. Note that within this discussion, there may be cases where autonomy is narrower to begin with (due to certain socioeconomic or normative characteristics), or where narrower autonomy is seen as more appropriate normatively, making the special steps to protect freedom of choice may not be necessarily required.

Some AI technologies, such as "deep fake", are designed to produce unreliable information that can hardly be distinguished from reliable one. These technologies have the potential of reducing the ability of individuals to understand reality and make autonomous, informed decisions, and of eroding the trust between people and between them and their government. For example, we are not far from the day when it would be possible to artificially produce a film where a leader declares war, leading to catastrophic results. The Committee believes that the State of Israel should examine ways of dealing with these technologies in a separate report.

One final area relevant to autonomy is the penetration of AI tools into the news media. Many communication channels use AI to produce individually customized news. This tool has many advantages, but also poses the danger of selective exposure: certain groups in the population are exposed to standardized information and are unaware of evidence and arguments that are inconsistent with their worldview. This would deny such a population the freedom of choice or the freedom to be exposed to a diversity of opinions, and make them vulnerable to unfair and highly effective influence campaigns by interested parties. In particular, this could enable foreign governments to intervene in elections."

[263] Nahon K. et al. (2020). p.188.
[264] Ibid. p.182.
[265] Ibid. p.189.

Civil and political rights

The Ethics Report defines civil and political rights as including the "right to elect, freedom of speech and freedom of conscience religion."[266] These go to the core of democratic values and warrant special protection. In that respect, the Ethics Report notes with concern how the automated manipulation of global discourse is manipulated, for example by over-amplifying certain views while silencing others, polarizing the discourse and giving legitimacy to views that could be offensive to certain groups, and disseminating false information on a large scale.[267]

All these can harm the democratic process itself, creating rifts within society and undermining faith in the democratic process, and produce.

Fairness

This is a broad ethical principle, that refers to the need to achieve substantial equality, to prevent of biases (in information, in the process and in the product), prevent discrimination, and avoid widening socioeconomic and educational gaps. The Ethics Report notes: *"Technology is not neutral, as it is based on human programming and various commercial interests. Moreover, the AI systems are based on information related to human behavior, which may reflect and exacerbate various types of social biases*".[268] It provides the following examples of AI systems that raise fairness questions:

- The system decides on allocating resources such as funds and medical treatments.
- The system evaluates candidates for a workplace or higher education.
- The system evaluates people for the purpose of criminal punishment or the mitigation thereof.
- The system makes decisions that threaten users' property and financial interests.

To address these risks, the Ethics Report underlines the importance of proactively studying the target population and identifying in advance groups that are liable to be misrepresentation or underrepresentation. In addition, it emphasizes the need to consult with representatives of the target users themselves to help produce fairer systems.[269]

Accountability

The Ethics Report separates accountability into three categories: *transparency, explainability and responsibility.*

Transparency is about *"Providing information about the process and related decision making"*[270] and is referred to as a *"key value in technological development and in developing AI products in particular"*.[271] It is both a value that stands alone, and an aspect of accountability as well. It is enables the monitoring and realization of other values such as fairness. Transparency is a core component of public trust.

[266] Ibid. p.182.
[267] Ibid. p.190.
[268] Ibid. p.184.
[269] Ibid.
[270] Ibid. p.182.
[271] Ibid. p.185.

Explainability is an AI system's capability of explaining its decision-making process, whether to the individual end-user, or on a collective level if the decision affects group. It also includes a system's capability of providing meaningful explanation to the operators of the system themselves. The Ethics Report adopts the EU Experts Group Report's position that explainability includes a principle of "meaningful information", that is that the level of information provided should be sufficient without being exceedingly technical or detailed.

Responsibility, involves making appropriate rules to prevent risk, based on the the context and the estimated severity of the risk, managing the risks and appointing an employee in charge of risk management. The Ethics Report notes that the diversity of stakeholders and the complexity of AI systems make this a particularly challenging task. This is compounded by the fact that AI systems also make their own "decisions".

Safety

The Ethics Report recognizes the need to address safety risks that arise from AI systems. Indeed, the more an AI system is empowered to make decisions with a direct impact on human life, the riskier is it to use. The risk arises both in ordinary operation of the AI system, as well as in extreme situations. The Ethics Report thus distinguishes between safety risks occurring as part of a malfunction, and those that occur when the system operated without malfunction but in a manner that nonetheless causes death or physical harm.

Safety risks can be mitigated by implementing a number of measures. For example, in order to prevent incorrect decisions based on faulty bias, a diverse dataset should be used. Similarly, safety considerations must be borne in mind at the design stage. The Ethics Report provides an interesting – and perhaps counter-intuitive – example. In the design of an autonomous car, it is important for the system to have trained on diverse conditions, including conditions where harm could be imminent. For this training process to occur, it would be necessary to place individuals in risk situations, which can then form part of the data set.[272] Of course, this could at least in part be done through simulations not involving physical human beings, but it highlights the tensions that exist in order between different ethical and human rights principles, and the trade-offs that are sometimes necessary in order for AI systems to be as "good" as possible.

Fair competition

Among the different ethical principles that are commonly referred to in various AI ethics documents and standards, fair competition is probably the least often quoted. By contrast, it is a standalone ethical principle in the Ethics Report. Fair competition refers to the need to enable innovators, entrepreneurs, software engineers and other stakeholders in the supply chain, to benefit from equitable access to data and opportunities to create and deploy AI systems. This principle may, at first glance, be characterized as an economics goal in disguise, but the Ethics Report develops it on distinctly human rights and ethics-based grounds. Indeed, the report notes: that fair competition is needed "*for innovation and social welfare. Thus, maintaining a free market with fair competition would allow all actors in the value chain, particularly small-to-medium enterprises and startups to benefit and profit from the activity.*"[273] The ethical imperative is societal as opposed to individual: in order for society to reap the benefits of AI transformation, greater innovation is needed, by a diversity of actors.
As noted by the report, examples of the challenges posed by unfair competition include:

[272] Ibid. p.190.
[273] Ibid.

- The system produces an advantage for competitors with big data.
- The system is based on a large database accessible to only few market players.
- In the course of its operations, the system produces a large and unique database that is inaccessible to competitors.
- Non-competition agreements and automatic coordination between companies based on AI systems.

The Ethics Report notes: "*Concentrations of economic power can also lead to concentrations of political power, allowing tech giants to dictate the rules of the game in the market. The fear is that the influence of these mega-players on the market could make it difficult for new technologies or applications to enter the market, and compromise the innovation so critical for AI*."[274] It further suggests that competition laws, standardized contracts and consumer protections be updated to meet the anticipated challenges. "*To that we must add the international challenge, resulting from the fact that some of the key players are based in the United States*."[275] In short, a major part of the challenge is to enable SMEs to have access to large databases.

V. Israel's approach to address the challenges

Israel is aware of the potential risks and challenges presented by the growing use of AI applications and is determined to address them. At the same time, as noted by the National Initiative Report, experience shows that over-regulation can stifle innovation, particularly when dealing with emerging technologies[276]. The National Initiative included two working groups dedicated respectively to ethics and regulation, and to cybersecurity for AI systems. Their main challenge was to establish a model that would **balance** the need to: (1) ensure ethical and secured development and deployment of AI applications in accordance with the values of Israel as a democracy; (2) foster technological innovation and scientific research and development which are fundamental to the Israeli economy and national security. As will be elaborated below, the approach of the National Initiative Report is novel in its manner of combining ethics, human rights and innovation.

i. Six Ethical Principles for AI

Acknowledging that human rights and ethical considerations remain paramount, the Ethics Report's lists “6 Ethical Principles for AI” that should inform public policy making:[277]

1. ***Fairness:*** *Striving for substantial equality, prevention of biases (in information, in the process and in the product), prevention of discrimination, and avoidance of widening socioeconomic and educational gaps.*
2. ***Accountability:***
    a. *Transparency: Providing information about the process and related decision making.*

[274] Ibid. pp.190-191.
[275] Ibid.
[276] Ben-Israel, I., Matania, E. & Friedman, L. (Eds.) (Sep. 2020). p.32.
[277] Nahon K. et al. (2020). p.182.

b. *Explainability: Being able to explain the system's decision-making process (on the level of individual users, as well as on a collective level if the system affects group, as well as for the system operators themselves).*
c. *Ethical and legal responsibility – to be divided among the relevant actors in the value chain, together with risk management. Determining the responsibilities for setting rules for reasonable measures to prevent the risk according to the context and the estimated severity of the risk, for managing the risks and for appointing an employee in charge of risk management.*

3. ***Protecting human rights:***
    a. *Bodily integrity: Preventing any harm to life or limb.*
    b. *Privacy: Preventing damage to privacy due to collecting, analyzing and processing information, sharing the information and making new and different uses of the information.*
    c. *Autonomy: Maintaining the individual's ability to make intelligent decisions, including the prevention of unfair or unconscious influence on individual behavior.*
    d. *Civil and political rights: Including the right to elect, freedom of speech and freedom of conscience religion.*
4. ***Cyber and information security:*** *Maintaining the systems in working order, protecting the information they use, and preventing misuse by a malicious actor.*
5. ***Safety:*** *Preventing danger to individuals and to society and mitigating any damage.*
    a. *Internal safety: In developing the AI tool.*
    b. *External safety: For the environments and clients, in using the tool.*
6. ***Maintaining a competitive market*** *and rules of conduct that facilitate competition.*

In light of these guiding ethical principles, the National Initiative Report suggests a balanced regulatory model based on applying the minimal regulatory intervention required for maintaining adequate ethical environment, on the one hand, while refraining from any unnecessary restrains on innovation and scientific progress, on the other hand. Accordingly, the National Initiative Report calls "*to encourage self-regulation through the use of the tools developed within the framework of the [national] initiative to assess risks and identify in advance ethical challenges in the stages of development and production. Ethical limitations should be integrated into the intelligent systems, forbidden conducts should be defined, and the ethical principles should be implemented during the learning and training process of those who deal with AI systems".*[278]

ii. Balanced regulation to foster innovation

One of the key features of the Ethics Report is its approach to regulation. Rather than setting out a list of activities that must be regulated, it takes a systematic approach to the question, comprising three steps: (1) mapping of different types of regulatory approaches, along with their respective advantages and drawbacks; (2) identifying the main areas and activities of AI that could benefit from some level of regulation, and the risks associated with each of them;

[278] Ben-Israel, I., Matania, E. & Friedman, L. (Eds.) (Sep. 2020). p.31.

(3) matching different regulatory approaches to the various AI activities. This provides a roadmap for the government to craft tailored, sector-specific regulations.

Details of the approach are provided below.

The Ethics Report identifies the following broad regulatory approaches:

1. Legislation or regulation
2. Judicial decision making to interpret existing legislating or fill the gaps
3. Professional standards (by government, industry, academia or civil society)
4. Self-regulation by ethical rules or professional standards usually developed by the relevant professional community.

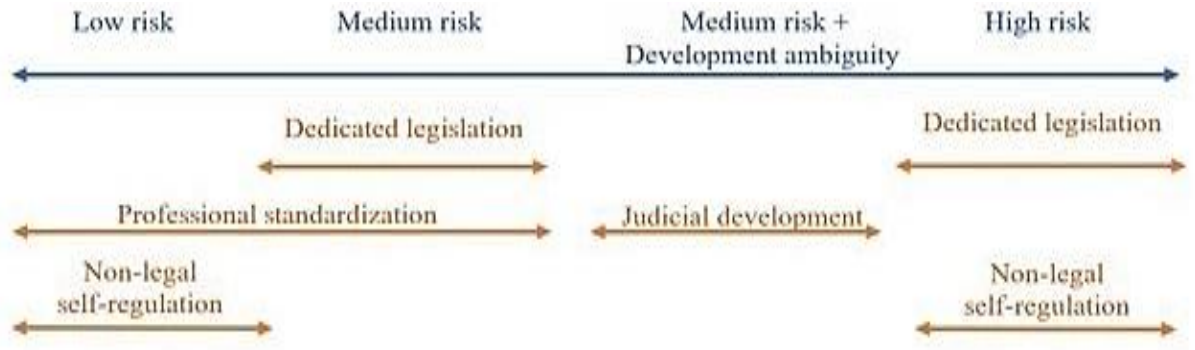
The report then highlights the advantages and drawbacks of each approach. The following table is reproduced as-is from the Ethics Report[279]:

Type of Regulation	**Characteristics**	**Strengths**	**Weaknesses**	**Committee Recommendations**
Dedicated legislation	Dedicated law or amendment enforced by a state authority or private entities	• Increased clarity about protected values • Allows concrete judicial development based on legislator guidelines • Partial flexibility	• Lack of professional expertise in a single organization • Retroactive enforcement only • Potential for increased uncertainty • Lack of involvement in present power relations that may privilege certain players	Suitable mainly for medium & high risk areas
Judicial development	No specific law	• No direct regulatory or legal friction • Flexibility • Enables judicial development	• Usually applicable to more obvious cases of harm, and may therefore fail to meet the entire range of harm risks • Lack of professional expertise in a single organization • Uncertainty • Advantage for strong players	Suitable for medium risk situations with development ambiguity
Professional standardisation	Allows future adoption by the legal system	• Flexibility • High legitimacy in the professional community • Participatory process	• Risks excluding the law and its values • Dependency on the law for binding validity, oversight & enforcement • Advantage for strong players	Suitable for medium and low risk situations + as a framework for developing & reviewing the application of ethical values

[279] Nahon K. et al. (2020). p.200.

Non-legal regulation	No legal norm (e.g. applying ethical principles)	• Flexibility • High legitimacy in the professional community	• Risks excluding the law and its basic values (equality, fairness, human rights) • Dependency on the professional community for development • Lack of reliable enforcement mechanism • Advantage for strong players	Suitable for low risk situations, where non-legal regulation is sufficient, and for high risk situations, where technological development is relatively rapid for the legal channel

The Ethics Report then proposes the following model, to match different regulatory approaches based on the risk level associated with a particular activity:



Thus, for example, high-risk activities are better addressed by legislation and self-regulation ex ante, than by post facto judicial intervention. At the other end, low risk activities do not necessarily require dedicated legislation, and can be addressed through standards and self-regulation.

This model, of course, is not meant to apply in a rigid fashion. Rather, it presents a framework that enables policymakers and regulators to gauge the appropriate means of an activity, factoring in a multitude of variables. It further notes that the question of "who regulates" is no less important: regulation by a central AI body enables the development of consistent policies; however, there is a risk of over-regulation and chilling innovation if a regulation is adopted across the board. Conversely, regulation could be left to different sector-based bodies, which would allow for greater experimentation, at the expense of uniformity of rules.

In light of the foregoing, the Ethics Working Group proposes 11 regulatory guidelines:

1. **Alignment of Israel's regulation with international legislation and standardisation, and promoting Israeli policy in global arenas** – this is essentially about participating in the international discussion around AI regulation, to be attuned to emerging international standards, while taking part in the shaping of those standards going forward.

2. **Mapping the actors to create an adapted responsibility and incentive framework** – this requires a multi-stakeholder approach, to enable policymakers to understand their respective roles in the value chain, their incentives, and their responsibility.

3. **Adjusting the accountability principle to the dynamism of the AI area** – the suggestion here is to require that organizations implementing AI technologies implement a testing environment and control perimeters prior to implementing the technology, in order to determine how to best apply the accountability principle in a given case taking account the anticipated effects of the technology.

4. **Promoting normative clarity in critical stages of the AI product value chain** - this emphasizes the importance of guidance in the early stages of AI development. An AI risk assessment tool, and perhaps in certain cases a regulatory requirement for AI impact assessment, would be useful in mitigating risks and in enabling developers to implement the various ethical principles and legal rules.

5. **Constant review of the regulatory policy by the regulator** – beyond monitoring the implementation of existing regulation and updating legal texts, this principle calls for regulatory experimentation. It requires regulators to take an agile, flexible approach, promote innovation while factoring in risks.

6. **Regulatory sandboxes** – the concept of regulatory sandboxes is well known. Controlled testing is particularly useful in an AI context "*because of the need to allow innovation on the one hand and address unpredictable risks to social interests on the other*".[280]

7. **The interface between the proposed principles and existing regulations** – given that laws and regulations already apply in many fields of activity (health, transportation, finance, education, etc.), the existing legal landscape must be borne in mind, along with the specific values, interests and potential social benefits of regulation, in determining whether new regulation is needed and what ought to be its focus and scope. At a basic level, every government body is already responsible to undertake this examination within the scope of the field it regulates.

8. **The role of the Privacy Protection Authority** - Privacy is a cross-sectoral issue, such that the Privacy Protection Authority has a predominant role to play in assessing the privacy implications of AI systems, and making regulation as needed, in coordination with other government bodies. Furthermore, it is important that the Privacy Protection Authority obtain the resources required for developing an up-to-date legal and technological framework for the area of information anonymization, as it is a fundamental to the development of AI.

9. **The role of the Competition Authority** – as mentioned above, competition and a fair market is not just about economy – it is also an ethical matter. Thus, the Competition Authority should be tasked with *"formulating regulations designed to maintain fair competition in the AI area, protect consumers and ensure the accessibility of technology; and prevent technological risks and costs from being rolled over to weaker players at the bottom of the value chain, in a way that is socially inefficient."*[281]

10. **The need for interministerial coordination** - to ensure coherent policy and regulation development, an interministerial coordination mechanism should be implemented.

11. **Authorities responsible for information resources**.

    Authorities that are responsible on substantive information resources used for AI technologies "*have a key role in examining whether the regulatory framework they*

[280] Ibid. p.203.
[281] Ibid. p.204.

apply is suitable for achieving societal benefit in this field, while maintaining a fair and free competitive market and protecting human rights. Consideration must be given in this regard not only to risks but also to innovation spaces and […] promoting societal interests.

The Committee therefore recommends that authorities responsible for areas of activity affected by the products of information processing will be required to undergo evaluation in light of the principles detailed above. Specifically, the authorities need to examine whether, when deploying AI technologies or using them in the activity areas regulated by them there is need for adjusting the applicable framework in order to promote the protection of the regulated interests."[282]

iii. Original Ethical Risk Assessment Tool

As the Israeli approach encourages self-regulation, the Ethics Report stresses the responsibility of all those involved in AI to remain up-to-date with the risks of the dynamic technology. To assist them in this demanding duty, the Ethics and Regulation working group developed an original *Decision-Maker Instrument for Assessing Ethical Challenges.* The instrument is designed to enable AI professionals to identify ethical risks throughout the development and production change and to respond properly. It consists of two parts:

A set of preliminary questions that should be addressed to AI product developers in order to assess the influence of the product:[283]

1. *What is the level of potential individual harm?*
2. *What is the extent of potential perceptual impact?*
3. *What is the degree of potential damage to the public?*
4. *Is there any impact on the allocation of public resources?*
5. *Is the development team diverse enough?*
6. *What is the expected extent of damage due to misuse of or loss of control over the product?*
7. *Is there a fast way to identify unpredicted ethical failures?*

A dynamic frequency map that helps locate challenging areas in terms of applying ethical values to the system's development. The map presents the six ethical principles juxtaposed cardinal milestones along the development process. It indicates the frequency of ethical issues along the product's development chain by highlighting areas where failures have been found in the past and providing information about their rate of incidence (See Annex II for a sample frequency map). The map is based on assessment of real-life past cases of AI systems which presented ethical challenges or conflicts, thus raising awareness to areas where AI organizations experienced trouble in the past, and areas for particular attention by decision makers. It is important to note that as the map is shaped by the test cases used to create it, each organization is expected to select a set of cases that are relevant to the product it develops. Furthermore, to remain relevant, the map needs to be frequently updated with new test cases. The Ethics Report explains in detail how an AI organization can create and update a frequency map relevant to its product. See Annex II for further information.

[282] Ibid. p.205.
[283] Ibid. pp.192-193.

iv. International activity and cooperation

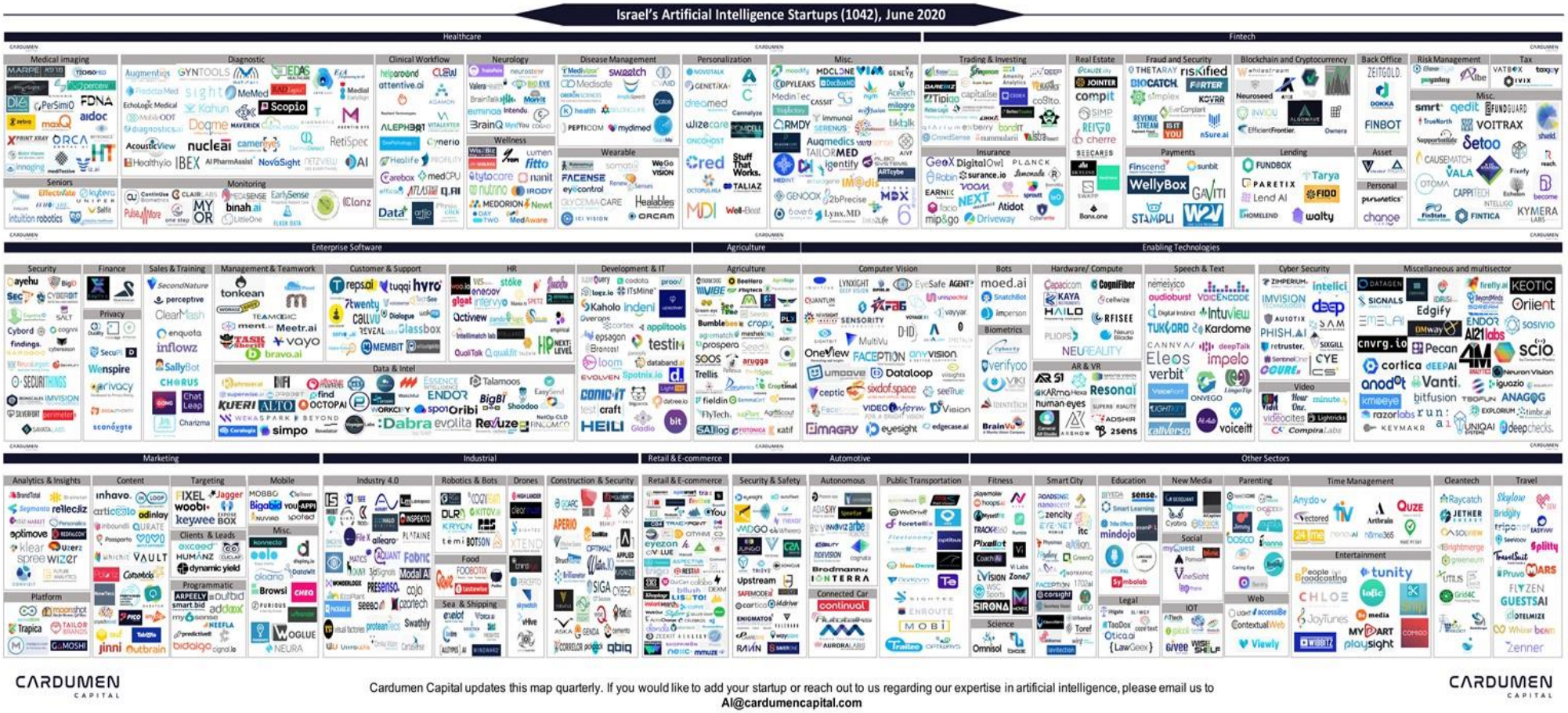
Israel has been involved in international forums dealing with AI ethics and human rights. Israeli representatives were active in the drafting of the OECD's AI Recommendations and guiding principles. In addition, Israel is a member of the "Digital Nations" ("DN"), regrouping 10 of the world's leading digital economics. In 2018, Israel hosted the annual DN meeting, in which a declaration on responsible AI was adopted.[284] In 2019, the DN also adopted a declaration on data governance.[285] While these declarations are not legally binding, they reflect the Digital Nations' commitment to abide by high standards of human rights, ethics and accountability in they use of digitization. Israel has also partnered with the World Economic Forum's C4IR project, in conducting research projects in the fields of transportation and health.[286]

[284] Shared Approach on the Responsible Use of AI, https://fdfd812d-4234-49d8-8755-ff45ad565157.filesusr.com/ugd/189d02_ef802d92ba5147d2901bde25c6e954a3.pdf, November 2018, Written by the Artificial Intelligence working group, this framework was adopted at the D9 Ministerial Summit in Israel in 2018.

[285] Data 360 Declaration, https://fdfd812d-4234-49d8-8755-ff45ad565157.filesusr.com/ugd/189d02_abce8f2b8cc140e4baeec7dcab7bee97.pdf, November 2019, Drafted by the Data 360 working group, this shared declaration was presented at the D9 Ministerial Summit in Uruguay in 2019.

[286] Israel Innovation Authority. *Establishment of the Israeli Center for the Fourth Industrial Revolution – World Economic Forum*. Retrieved from https://innovationisrael.org.il/en/contentpage/establishment-israeli-center-fourth-industrial-revolution-world-economic-forum .

Annex I. Israel's AI startup landscape segmented by sectors and applications



Source: Cardumen Capital (June 2020). https://www.cardumencapital.com/ai-israel-landscape

Annex II. Frequency Map of Ethical Challenges in the AI Development Process

The following part is reproduced as is from pages 194-197 of the Ethics Report

Frequency map

The frequency map indicates the frequency of ethical issues along the product's development chain. It pinpoints areas where failures have been found in the past and provides information about their rate of incidence. As the frequency can change with time and new events found, we recommend updating the map on a regular basis, as also demonstrated below.

In order to create the frequency map, we used ten test cases selected out of real-life past cases that represent various challenges. The map illustrates all the ethical principles listed under "Ethical Principles for AI" on p.141 above.

Table 1: Prototypical Test Cases of Ethical Challenges

1	**AI system for screening workplace candidates** Companies are contacted by multiple candidates wishing to work for them. In order to select the best candidates, several companies have developed AI-based tools trained based on past decisions by the companies. When one such system developed by Amazon was tested, it was found to discriminate against women candidates for technical job. It is assumed that in the past company executives used to discriminate this way, and the system learned to emulate this behavior[287].
2	**Using AI for political influence** Cambridge Analytica collected personal data of millions of Facebook profiles without the users' agreement or knowledge, and used them to influence the users for political purposes. There was probably use of AI technology to manipulate minds. This activity went on for several years[288].
3	**Predicting disease risk** During the 1990s, several research centers joined hands to develop a system that would estimate the degree to which pneumonia represents a life risk for specific patients. This was designed to help doctors decide which patients to hospitalize and which can be treated in the community. Shortly before the system's launch, it was found that its recommendations for asthmatics could risk their lives, because the information used to build the system was biased: asthmatics with pneumonia had received preliminary intensive care that saved their lives, and the system deduced that pneumonia was not risky for asthmatics.[289]
4	**System for assessing detainee dangerousness** When deciding whether to remand a detainee, one of the considerations is the danger he poses to others. The decision is based on multiple parameters, such as criminal history. Several US

[287] https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine
[288] https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data, https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
[289] http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf

	districts have adopted an AI system called Compas to help judges assess suspects' dangerousness. The system was tested and was found to assess white detainees as less dangerous than black ones.[290]
5	**Virtual AI-guided players accumulate tie-breaking weapons** In a game called Elite Dangerous, human players compete against AI-guided players. To make the game more interested, restrictions on the virtual players were changed in Version 2.1, to enable them to fly and fight better. The AI mechanisms found a way of taking advantage of those changes to accumulate weapons in a way that prevented human users from being able to match them.291
6	**The racist bot** Microsoft launched a bot in order to teach it to correspond freely with Twitter users. The idea was that the bot would engage in conversation and learn to improve its dialogue skills in the process. Less than 24 hours after the launch, it was found that since it emulated the users, several users chose to turn it into a racist bot by using racist comments themselves. [292]
7	**The impersonator bot** Google Duplex enables a bot to hold a conversation in a manner that made it difficult for its interlocutors to determine whether it was human. Building this tool required access to huge amounts of data available to only very few knowledge-intensive companies. [293]
8	**Autonomous car runs over pedestrian** A pedestrian that crossed the street in a dark area was killed in Arizona by an Uber autonomous vehicle. Apparently, the vehicle identified an "obstacle" and could have avoid crashing into it. Nevertheless, since the engineers had previously lowered the software's sensitivity to barriers, the vehicle did not stop and the woman was killed. The human driver in the vehicle was not alert enough to prevent the accident.[294]
9	**Face recognition bias** Amazon developed a tool for engineers enabling them to add face recognition capability to the system they were developing. The system was designed, among other things, to be used by law enforcement, border police, etc. A test revealed that the system erred much more frequently when activated on people with a dark skin than on people with a light skin.[295]
10	**Content recommendation systems show different information to different groups** Various companies use AI to offer more personally relevant information for users. It was found, however, that Google's ad system presents ads seeking information related to criminal acts when a user searches for information under a name more common in minority populations. [296]

[290] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[291] https://futurism.com/this-video-games-artificial-intelligence-turned-on-players-using-super-weapons
[292] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist
[293] https://www.androidcentral.com/google-duplex-will-let-people-know-its-not-human
[294] https://www.cbc.ca/news/business/uber-arizona-crash-1.4594939
[295] https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender
[296] https://www.bostonglobe.com/business/2013/02/06/harvard-professor-spots-web-search-bias/PtOgSh1ivTZMfyEGj00X4I/story.html

Ethical milestones along the development process

Below, we present examples for ethical issues arising during the development process and follow up on them as they unfold, in order to identify particularly sensitive development milestones. To do so, we present a typical AI development process.

1. **Product definition**
    a. Understanding the business need or problem the system is trying to solve and creating the R&D organization
    b. Data collection – identifying information sources from within and outside the organization to be used for building the system and assessing its performance
2. **Product training**
    a. Processing and filtering the raw data into a form that would enable the AI algorithms to receive the data and perform calculations with them
    b. Modelling – applying an AI algorithm to the information processing in an attempt to identify generalizable patterns
3. **Integration**
    a. Evaluating the model for accuracy
    b. Connecting the AI components with the rest of the system and distributing it for wide use
4. **Market management**
    a. Performance monitoring to make sure the system works as expected
    b. Ecosystem – together with the process within the organization, there is need to also address the ethical considerations arising out of the fact that the process takes place in the Israeli ecosystem. Integrating AI could affect the socioeconomic, regulatory and other systems, and this should be continuously monitored after launch.

Creating the frequency map

Review the list of test cases and the implications and reported events considering the list of ethical values on p.8 above. Fill in the table according to the emerging ethical challenges. The numbers within the table cells refer to the event number. Next, check the accumulated number of events. Cells with low, medium and high event frequencies are colored beige, yellow, and red, respectively. Note that this table does not indicate the degree and scope of the potential harm. A more sophisticated tool can take these factors also into account. The decision regarding what constitutes low or high frequency should be taken when selecting the number of events the organization refers to. In Table 2, we have ten events, and the frequencies have been determined accordingly.

The Committee recommends that decision makers discuss and offer solutions for emerging challenges according to the frequency map throughout their development process. Since the map depends on a list of test cases, each organization needs to choose a set of test cases relevant to the product under development, assuming that this set changes in time.

Table 2: Frequency Map of Ethical Challenges in the AI Development Process

	Business need	Data collection	Data organization	Modelling	Model evaluation	Distribution	Performance monitoring	Ecosystem
Fairness		1,3,4		3,4	1,3,4	1,4,9	1,3,4,9	1,4,9
Transparency	4			3,4				4
Explainability	4			3,4				9
Accountability	1,2,3,4				1,3,4	3,4	3,4,5,6	2,5,6,9
Privacy	2,9	1	1,2	1	1			2,9
Freedom of choice	7,10						10	6,7,10
Infosecurity			2					2,9
Human rights	4,9			4		4	4	4,9
Safety	3,4	3,4		3,4	3,4	3,4,5,8	3,4,5	3,5,8
Free market	5					5,6	5	

Legend

1	Job candidate screening
2	Political influence
3	Predicting disease risk
4	Assessing detainee dangerousness
5	AI-guided players gain tie-breaking weapons
6	Racist bot
7	Impersonator bot
8	Autonomous vehicle runs over pedestrian
9	Face recognition bias
10	Content recommendation systems present different information to different groups

Low frequency of problematic cases (single case)

Medium frequency of problematic cases (two cases)

High frequency of problematic cases (three cases or more)

CHAPTER II. AI Governance in Japan

Arisa Ema[297] & **Hideaki Shiroyama**[298] (The University of Tokyo)

I. Introduction

Various actors are involved in the process of research and development of new technologies and their utilization in society. At each stage of the research, development and utilization of a technology, each actor makes decisions and policies for its social implementation based on an assessment of its social impact. The totality of the roles of various actors in the above-mentioned social impact assessment, decision and policy-making, and implementation can be called technology governance[299]. Currently, there are various approaches toward governance of artificial intelligence (AI) technologies in Japan and abroad. There is no single right answer in regards to the systems and means for implementing technology in society, and in fact this issue intertwines a combination of different factors, including technology, culture, policy, people's values, existing legal systems, the environment, and the economy. As a result, the debates in these governance attempts are diverse.

Although it is difficult to cover all policies and activities, this report summarizes the nature of AI governance and the characteristics of the discussions in Japan. Since this report mainly deals with governance attempts and discussions up to 2018[300], it is easy to imagine that the details contained within it will change with future technological developments and changes in social conditions. However, organizing discussions at a fixed point is useful for future discussions on AI governance and for comparative research with other cutting-edge technology governance.

This report is organized as follows. Chapter II "AI Governance in Japan" first organizes the activities and reports of each actor in industry, academia, and the government. This is followed by an analysis of the white papers and reports published by the ministries and agencies for fiscal years 2016-2017. An overview and comparison of how the ministries position AI and related technologies (e.g. big data and IoT) is provided in addition to a section on how they perceive their areas of use and challenges. In chapter III, "Comparison of AI governance in Japan and abroad", we summarize the issues and perspectives that are missing in AI governance discussions in Japan. Finally, in chapter IV, we summarize how discussions on AI governance should be handled in the future.

[297] Project Assistant Professor, Institute of Future Initiative, The University of Tokyo.
[298] Professor, Graduate School of Public Policy, The University of Tokyo.
[299] Hideaki Shiroyama, Science, Technology and Politics (Minerva Shobo, 2018), chapters 7 and 9 (in Japanese).
[300] This report is based on a partial translation from a chapter “AI Governance” written by Arisa Ema and Hideaki Shiroyama, in *Artificial Intelligence, Humanity and Society* (Keisho Shobo, 2020) in Japanese (https://www.keisoshobo.co.jp/book/b498075.html). Some Japanese references are replaced with English versions and supplemented with information from 2020 by the author and colleagues.

II. AI Governance in Japan

i. The role of each actor

This chapter summarizes the discussions on AI governance in Japan by different actors.

Academic societies

In February 2017, the Ethics Committee of the Japanese Society for Artificial Intelligence (JSAI) released its ethics guidelines. These are not ethics for AI technology but are rather more directed at the ethics of AI researchers, as they provide guidelines for behavior that should be obeyed as researchers and which include sections such as Accountability and Social Responsibility and Abidance of Laws and Regulations[301].

In addition, the Information Processing Society of Japan (IPSJ) established the SC 42 Expert Panel on Artificial Intelligence in January 2018 to work with SC 42, a subcommittee established by ISO/IEC JTC 1 (joint technical committee of the International Organization for Standardization and the International Electrotechnical Commission) and the society has been disseminating Japan's opinions on the subject[302]. In addition, the Academic Promotion Council within the Japan Medical Association released a report entitled "Artificial Intelligence (AI) and Medicine" in June 2018[303].

Universities

There are many research institutions in Japanese universities with the words "AI" and "intelligence" in the names of their faculties and departments. For example, the University of Tokyo also has a Next Generation Artificial Intelligence Research Center, which hosted the AI and Society Symposium in 2017[304], and co-hosted the Beneficial AI Tokyo with CFI in the UK, to which people from IEEE, PAI, and other relevant research institutions and companies[305] were invited.

Industry organizations

The Japan Business Federation (Keidanren)

The Japan Business Federation (Keidanren) has also been initiating discussions on the impact of AI and other advanced technologies. In May 2018, the federation publicly released a proposal entitled "Creation of New Values through the Implementation of Three Business Principles for the Achievement of the SDGs - Toward the Formulation of an IPR Strategy Vision'"[306]. The proposal states that, "Japan has been lagging behind in the use of data in business," and that it is important to regain international competitiveness by combining data with AI and other technologies. The proposal flags the concept of Society 5.0 in solving social challenges, and urges business to be promoted in line with the UN's Sustainable Development Goals (SDGs). In terms of the use of data the proposal notes that, "There is a need to build

[301] The Ethics Committee stated that they developed these guidelines to first gain the trust of the research community before addressing the impact of AI on society and ethical perspectives. In the "Japanese Society for Artificial Intelligence Ethical Guidelines" (http://ai-elsi.org/archives/514), Prof. Yutaka Matsuo, then chairperson of the ethics committee, said, "Professor Shun Tsuchiya, said that: “The general public is concerned about what artificial intelligence researchers would do with the technology. Therefore, it is important to first make known that researchers are aiming to create a better society, and that they are not mad scientists. So, I would like to praise the JSAI for issuing such Ethical Guidelines."

[302] Information Processing Society, International Standardization on Artificial Intelligence Launched, https://www.ipsj.or.jp/release/20180110_itscjnews.html (in Japanese)

[303] Artificial Intelligence (AI) and Medicine", a report of the IXI meeting of the Academic Promotion Council of the Japan Medical Association. http://dl.med.or.jp/dl-med/teireikaiken/20180620_3.pdf (in Japanese)

[304] AI and Society, http://www.aiandsociety.org/

[305] Subsequently, Beneficial AI Japan (http://bai-japan.org/) has been established. Note that the Next Generation Intelligence Science Center joined the Partnership on AI (PAI) in August 2018.

[306] Creating New Value through the Three Principles of Society 5.0 Enabled Business, May 15, 2018, http://www.keidanren.or.jp/policy/2018/042_honbun.html (in Japanese)

systems that enable different entities to use data while giving due consideration to privacy and cybersecurity, and to spark innovation in a variety of areas," and adds that active involvement in rule-making, including international regulations and standards, and human resource development is a challenge.

At the first meeting of the Cabinet Office's Council for Social Principles of Human-Centric AI on May 8, 2018, a document entitled "Formulation of Principles for Utilization of AI for Implementation of Society 5.0" was provided by Keidanren[307], which submitted information on the formulation of an industrial version of the AI development guidelines and strategic concept for deployment[308].

National research institutes

Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology

This was established in May 2015, and as of October 2018, 12 research teams are working on different projects[309]. Seminars are also held regularly, including a discussion on "AI and the Law" in December 2017[310].

RIKEN Center for Advanced Intelligence Project

Established in 2016, the Riken project has three groups carrying out research, the Generic Technology Research Group, the Goal-Oriented Technology Research Group, and the Artificial Intelligence in Society Research Group. The Artificial Intelligence in Society Research Group consists of sub-teams such as the Information Legislation Team and the Artificial Intelligence Ethics and Society Team, which not only carry out technical research, but also survey research from the humanities and social sciences fields[311].

AI Science Research and Development Promotion Center, National Institute of Information and Communications Technology

The Center was established in 2017 as an open-innovation, strategic research and development promotion organization for AI technologies in the field of intelligence science [312]. The center shares a variety of text and speech data, and is building a state-of-the-art AI data testbed, to accelerate open innovation.

Japan Science and Technology Agency (JST), Research Institute of Science and Technology for Society (RISTEX)

RISTEX established the Human-Information Technology Ecosystem (HITE) focus area in 2016[313]. The purpose of this focus area is to promote the co-evolution of technology and society by organizing issues that arise when new information technologies, such as AI, are introduced into society and by building a place and platform where such issues can be directly fed back to same research and development fields.

Government organizations

Cabinet Office

In April 2016, the Strategic Council for AI Strategy was established in response to the prime minister's directive in the "Public-Private Sector Dialogue for Future Investment." The council

[307] Keidanren, AI Utilization Strategy For an AI-Ready Society, February 19, 2019, https://www.keidanren.or.jp/en/policy/2019/013_outline.pdf.
[308] Hiroaki Kitano, Mission Statement Keidanren AI Application Principles Task Force, http://www8.cao.go.jp/cstp/tyousakai/humanai/1kai/sanko1.pdf (in Japanese)
[309] Artificial Intelligence Research Center: https://www.airc.aist.go.jp/teams/
[310] Artificial Intelligence Research Center [19th AI Seminar] "AI and the Law" https://www.airc.aist.go.jp/seminar_detail/seminar_019.html (in Japanese)
[311] Innovative Intelligence Integration Research Center (AIP): https://www.riken.jp/en/research/labs/aip/index.html
[312] Center for the Promotion of Integrated Research and Development of Intelligent Science: http://www2.nict.go.jp/oihq/en/
[313] RISTEX-HITE website: https://www.jst.go.jp/ristex/hite/en/

serves as a command post. In addition to the three ministries (Ministry of Internal Affairs and Communications; Ministry of Education, Culture, Sports, Science and Technology; and Ministry of Economy, Trade, and Industry), research and development and social implementation are currently being promoted through cooperation and collaboration with related ministries and agencies such as the Cabinet Office, the Ministry of Health, Labour and Welfare, the Ministry of Land, Infrastructure, Transport and Tourism, and the Ministry of Agriculture, Forestry and Fisheries. The concept of Society 5.0 and a "super-smart society" have been proposed as societal benefits that should be realized.

In March 2017, the Artificial Intelligence Technology Strategy was formulated, and in addition an industry roadmap was created. Under the roadmap, national agencies are supposed to work on (1) research and development, (2) human resource development, (3) development of an environment for data and tools, (4) support for start-ups, and (5) promotion of an understanding of AI technology. The draft Artificial Intelligence Technology Strategy Implementation Plan was compiled[314] in August 2018.

In fiscal 2016, the Cabinet Office established the Advisory Board on Artificial Intelligence and Human Society, which organized various issues into a single report[315], and the Council for Social Principles of Human-centric AI, which was established in the Cabinet Office in May 2018, prepared a draft summary of the Social Principles of Human-centric AI at the end of 2018[316]. In March 2019, the Social Principles of Human-Centric AI was released by the Cabinet Secretariat in February 2019[317], and, when Japan hosted the G20 in June 2019 the principles were introduced.

Ministry of Internal Affairs and Communications (MIC)

The Institute for Information and Communications Policy (IICP) of the Ministry of Internal Affairs and Communications is a research organization that does research on information and communications policy, and which is carrying out research on AI. In October 2016, the IICP established the Conference toward AI Network Society with the aim of examining social, economic, ethical, and legal issues for the promotion of AI networking. This conference includes the participation of stakeholders from industry and other sectors, in addition to experts in the fields of science and technology, humanities, and social sciences. The Committee on AI R&D Principles and the Committee on Impact and Risk Assessment were established under the Conference[318]. In July 2017, the Conference released the AI R&D Guidelines[319]. These guidelines were released with the aim of contributing to international discussions and the guidelines were presented at the G7 and other conferences. The IICP Conference also released the Draft Guidelines for AI Utilization which was released as the AI Utilization Guidelines in 2019[320].

Ministry of Education, Culture, Sports, Science and Technology (MEXT)

The Minister's Meeting on Human Resource Development for Society 5.0 [321] was established in December 2017 and met until June 2018, and was focused on human resource development while parallel discussions were held by the Ministry's Task Force on Fostering the Power to Live Richly in the New Era. A report released in June 2018 flags in its very first section policy directions which need to be addressed in societies where AI technology has developed, such

[314] The Strategic Council for AI Technology is summarized on NEDO's AI Portal: http://www.nedo.go.jp/activities/ZZJP2_100064.html (in Japanese)
[315] Cabinet Office, "Report on Artificial Intelligence and Human Society" https://www8.cao.go.jp/cstp/tyousakai/ai/summary/aisociety_en.pdf
[316] Cabinet Office, Council for Social Principles of Human-centric AI, https://www8.cao.go.jp/cstp/stmain/aisocialprinciples.pdf
[317] Cabinet Secretariat, Social Principles of Human-Centric AI, 2019, https://ai.bsa.org/wp-content/uploads/2019/09/humancentricai.pdf
[318] Shiroyama, Hideaki, "Artificial Intelligence and Technology Assessment: framework, system and experimental trials", Journal of Science, Technology and Society, 16, 2018. (in Japanese)
[319] The Conference toward AI Network Society, Draft AI R&D Guidelines for International Discussions, 2017, https://www.soumu.go.jp/main_content/000507517.pdf
[320] The Conference toward AI Network Society, AI Utilization Guidelines, 2019, https://www.soumu.go.jp/main_content/000658284.pdf
[321] MEXT, Minister's Meeting on Human Resource Development for Society 5.0, https://www.mext.go.jp/b_menu/activity/detail/pdf2018/20180605_001.pdf

as the provision of diverse learning opportunities and venues to realize "fair and individually optimized learning," the acquisition of basic academic skills and the ability to use information, and to transcend the humanities/science divide.

Ministry of Economy, Trade and Industry (METI)

The Ministry of Economy, Trade and Industry established a Study Group for Ideal Approaches to Competition Policies for the Fourth Industrial Revolution in January 2016 to examine, and consider future ways of dealing with, the current situation and challenges related to cross-cutting systems, such as competition policy and intellectual property policy, while keeping in mind the possibility of rapid changes in industrial structures that will come about because of the fourth industrial revolution, including the development of AI. The study was carried out over July and a report was compiled[322].

The New Industrial Structure Committee of the Industrial Structure Council of the Ministry of Economy, Trade and Industry[323] has been meeting since September 2015 and has formulated a vision which could be shared by the public and private sectors. In addition, the committee discussed the measures that would be required of the public and private sectors in order to accurately respond to changes caused by IoT, big data, and AI. In May 2017, a final report on the New Industrial Structure, which has been renamed as the "Future Vision Toward 2030s,[324]" was released, and four strategic fields (Mobility, Supply-chain, Healthcare and Living) were defined for achieving Society 5.0 as an ideal social vision to be aimed at. Furthermore, the report presents a roadmap and breakthrough projects.

In addition, the Information Economy Division, Commerce and Information Policy Bureau started a study group on AI and data contracting guidelines in December 2017, and the "Contract Guidelines on Utilization of AI and Data" was released in June 2018[325]. The guideline consists of a data chapter and an AI chapter, which, as a reference for private enterprises, summarize the main issues in contracts, sample contract clauses, and factors to be considered when drafting clauses, when concluding contracts related to the use of data, or contracts related to the development and use of software that uses AI technology.

Ministry of Health, Labour and Welfare (MHLW)

In February 2015, the Ministry of Health, Labour and Welfare began meetings on formulating a vision for health care policy for the next 20 years (2035) at the Japan Vision; Health Care 2035 round-table conference[326]. After eight rounds of discussions, their proposals were released in June 2015, and in them they set forth three basic values and criteria: fairness and equity, solidarity based on autonomy, and prosperity and coexistence for Japan and the world[327].

Subsequently, the Council on AI Utilization Promotion in Insurance and Medicine starting meeting in January 2017[328], and they compiled a report in June 2017 which selected six priority areas for the promotion of AI development. Subsequently, the Consortium for AI Development Promotion in Insurance and Medicine was held in July 2018[329].

---

[322] METI, A Study Group for Ideal Approaches to Competition Policies for the Fourth Industrial Revolution Compiles a Report: https://www.meti.go.jp/english/press/2017/0628_001.html

[323] METI, Industrial Structure Council, https://www.meti.go.jp/english/policy/economy/industrial_council/index.html

[324] METI, a final report on the new industrial structure vision was compiled: https://www.meti.go.jp/english/press/2017/0530_003.html

[325] Ministry of Economy, Trade and Industry, English Translation Version of the Contract Guidelines on Utilization of AI and Data Released, 2019, https://www.meti.go.jp/english/press/2019/0404_001.html

[326] the Japan Vision; Health Care 2035: https://www.mhlw.go.jp/seisakunitsuite/bunya/hokabunya/shakaihoshou/hokeniryou2035/assets/file/healthcare2035_proposal_150703_slide_en.pdf /

[327] Health Care Insurance 2035, Summary of Recommendations: https://www.mhlw.go.jp/seisakunitsuite/bunya/hokabunya/shakaihoshou/hokeniryou2035//future/

[328] MHLW, Council on AI utilization promotion in insurance and medicine: https://www.mhlw.go.jp/stf/shingi/other-kousei_408914.html (in Japanese)

[329] MHLW, Consortium for AI development promotion in insurance and medicine, https://www.mhlw.go.jp/stf/shingi2/0000148680_00002.html (in Japanese)

In addition, Future of Work: 2035 meetings were held from January 2016 regarding work style changes due to technological innovation, mainly in AI technology, and a report on "Future of Work: 2035" was released in August 2016[330]. In addition, in March 2017, Mitsubishi UFJ Research and Consulting published a study report, commissioned by the Ministry of Health, Labour and Welfare, on the impact of IoT, big data, and AI on employment and labour[331].

Ministry of Land, Infrastructure, Transport and Tourism (MLIT)

The Ministry of Land, Infrastructure, Transport and Tourism has been mainly holding discussions on automated driving, and in September 2018, the MLIT formulated the Guidelines for Safety in the Automated Operation of Agricultural Machines, which clarifies the safety requirements that level 3 and 4 automated vehicles must meet[332]. In April 2018, they also announced a study into the “AI Center Concept,” which will evaluate and certify provided data and AI needed for the development of bridges and tunnels, etc., in order to use AI for infrastructure inspections[333].

Ministry of Agriculture, Forestry and Fisheries (MAFF)

In 2013, the Ministry of Agriculture, Forestry and Fisheries established the Study Group on the Promotion of Smart Agriculture with the cooperation of robot and IT companies, agricultural machinery manufacturers, and related ministries and agencies. This study group set out to study measures and guidelines for promoting the use of AI and robots in agriculture (smart agriculture)[334]. In 2017, the research group formulated the Guidelines for Safety in the Automated Operations of Agricultural Machinery, which defines the requirements for manufacturers and users to ensure the safety of technology for the unmanned automatic operation of agricultural machinery (robotic agricultural machinery). In 2018, a summary table concerning sources of hazards as well as hazardous conditions for the automation of tea plantation management robots was added to the guidelines[335].

ii. Discussions on AI in ministries and agencies

Purpose

Various actors in Japan are discussing experiments in the governance of AI. Therefore, this section is focused on a comparative study of the discussions, as they stood in 2017, that appeared in white papers and the reports of ministries developing policies on AI. This is in order to understand the current situation from a multifaceted perspective that is not specific to any one ministry. Specifically, we will examine: (1) the overlap of references to AI-related technologies in each ministry and agency; (2) what areas of application and benefits, challenges and measures are mentioned in each ministry and agency with regard to AI; and (3) what kind of relationship there is between social visions, such as Society 5.0 and the fourth industrial revolution, and AI-related technologies.

[330] MHLW, Future of Work: 2035 - For everyone to Shine-: https://www.mhlw.go.jp/file/06-Seisakujouhou-12600000-Seisakutoukatsukan/0000152705.pdf
[331] Mitsubishi UFJ Research and Consulting, Report of the Study Group on the Impact of IoT, Big Data, AI. on Employment and Labour, March 2017, https://www.mhlw.go.jp/file/04-Houdouhappyou-11602000- Shokugyouanteikyoku-Koyouseisakuka/0000166533.pdf (in Japanese)
[332] MILT, Development of Safety Technology Guidelines for Self-Driving Vehicles, September 12, 2018, http://www.mlit.go.jp/report/press/jidosha07_hh_000281.html
[333] Nikkan Kogyo Shimbun, AI Centralized Management of Infrastructure Inspections, Ministry of Land, Infrastructure and Transport considers 'AI Center Concept', April 26, 2018, https://www.nikkan.co.jp/articles/view/00471269
[334] MAFF, Study Group on Promotion of Smart Agriculture, https://www.maff.go.jp/e/policies/tech_res/smaagri/attach/pdf/robot-1.pdf
[335] MAFF, Guidelines for Ensuring Safety for Automatic Operation of Agricultural Machinery, March 2018, http://www.maff.go.jp/j/kanbo/kihyo03/gityo/g_smart_nougyo/attach/pdf/index-6.pdf (in Japanese)

Survey targets

Japanese ministries and agencies have published various white papers and reports[336]. Among them, the following 19 documents, which seemed likely to contain discussions on AI, were included in the study. Only the versions available as of December 2017 were included in this study.

1-3	Growth Strategy 2017 (Headquarters for Japan's Economic Revitalization, Cabinet Secretariat)
1-4	Report on Priority Measures for Strengthening Industrial Competitiveness 2016 (Headquarters for Japan's Economic Revitalization, Cabinet Secretariat)
3-1	Annual Report on the Japanese Economy and Public Finance (Cabinet Office)
5-1	The White paper on Police (National Police Agency)
7-1	Financial Services Agency Annual Report (FSA)
9-1	White Paper on Information and Communications (Ministry of Internal Affairs and Communications)
10-6	Highlights of the Draft FY2018 Budget (Ministry of Finance)
12-1	White Paper on Science and Technology (Ministry of Education, Culture, Sports, Science and Technology)
12-2	White Paper on Education, Culture, Sports, Science and Technology (MEXT)
13-1	Annual health, Labour and Welfare Report (Ministry of Health, Labour and Welfare)
13-2	White Paper on Labour Economy (Ministry of Health, Labour and Welfare)
14-1	Annual Report on Food, Agriculture and Rural Areas (Ministry of Agriculture, Forestry and Fisheries)
15-1	White Paper on International Economy and Trade (Ministry of Economy, Trade and Industry)
16-1	Annual report on Transport Policy (Ministry of Land, Infrastructure, Transport and Tourism)
16-2	White Paper on Land, Infrastructure, Transport and Tourism (Ministry of Land, Infrastructure, Transport and Tourism)
17-1	Annual Report on Current Status of Meteorological Service (Japan Meteorological Agency)
18-1	Annual Report on the Environment (Ministry of the Environment)
19-1	Defense of Japan (Ministry of Defense)
20-1	White Paper on Consumer Affairs (Consumer Affairs Agency)

Research methods

Word selection

The following 13 words (in Japanese) were searched and extracted as possible key words in the current debate on information technology.

- Artificial Intelligence, AI (人工知能/AI)
- Big Data (ビッグデータ)
- IoT (IoT)
- Robot (ロボット)
- Drones, Unmanned Aerial Vehicles, and Small Unmanned Aerial Vehicles (ドローン/無人航空機/小型無人機)
- Automatic driving and automated driving (自動運転/自動走行)
- VR, AR, and MR (VR/AR/MR)
- Society 5.0 (Society 5.0)
- Fourth Industrial Revolution (第4次産業革命)

[336] The Prime Minister's Office website shows a list of white papers by year (https://www.kantei.go.jp/jp/hakusyo/index.html)

- Connected Industries (Connected Industries)
- Artificial General Intelligence (AGI) (汎用人工知能)
- Singularity (シンギュラリティ)
- Quantum (量子)

Word count

We searched for the words in the 19 surveyed documents and counted their occurrence.

Structuring

The descriptions around the words were extracted and were labeled under one of the following four points according to their content.

(a). Positioning: a section which describes the definition and position of the selected word and an explanation of the situation to be addressed
(b). Areas of application and benefits: a section where the selected word describes an area of application or benefit
(c). Challenges and measures: a section where the selected word describes a challenge, including policies and measures to address that challenge
(d). Challenges: where the selected word describes only challenges

iii. Results

Number of occurrences per white paper and report

Table 1 shows the structure of each word for technology in the 19 surveyed documents, extracted and structured by technology label. The same word occurring under the same technology label is counted as one[337]. Words that were mentioned in more than 10 places are shown in bold in Table 1.

Table 1 shows that there was no mention of MR, singularity, or Artificial General Intelligence (AGI) in any of the materials. References to VR/AR, Connected Industries and Quantum were found, but they appeared less frequently.

On the other hand, when comparing each white paper and report, 1-3 "Growth Strategy 2017", 3-1 "Annual Report on the Japanese Economy and Public Finance" and 9-1 "White Paper Information and Communications" refer to many of the keywords raised in this survey.

Some white papers and reports make heavy use of certain keywords. For example, "Society 5.0" is used a lot in 12-1 "White Paper on Science and Technology", "robot" in 14-1 "Annual Report on Food, Agriculture and Rural Areas", "Fourth Industrial Revolution" in 15-1 "White Paper on International Economy and Trade", and "automatic driving/automated driving", "drones" and "big data" in 16-2 "White Paper on Land, Infrastructure, Transport and Tourism". In addition, although the words used in this survey were not found in 7-1 "Financial Services Agency Annual Report", cybersecurity and fintech are mentioned in Chapter 22 "Other Issues", Section 1 "International Moves to Address New Challenges".

[337] For example, on page 9 of the "Future Investment Strategy 2017", it says, "We will promote the development of a cloud environment for AI development and the establishment of a certification mechanism, and the development of rules for the manner of evaluation to ensure the quality and safety of medical devices that use AI. Based on the above, the government aims to evaluate the quality of medical care through accurate support for physicians' medical care using AI in the next and subsequent revisions of medical reimbursement fees." Here, the word "AI" is used three times. This is then labeled as (c) challenges and measure, and further labeled with a sub-label of "evaluation" to make it one count. Note that some footnotes, figures, and columns that were considered trivial in structuring were excluded and therefore differ from the actual number of word searches.

Inter-agency comparison for each keyword

The searched words were categorized into the following four groups, (a) Positioning, (b) Areas of application and benefits, (c) Challenges and measures, and (d) Challenges, for three specific technological areas, (i) Artificial Intelligence, (ii) Other technology keywords, and (iii) Society 5.0/Fourth Industry Revolution/Connected Industries. Table 2 (References to Artificial Intelligence/AI) lists the main sub-item labels used in (b) areas of application and benefits, (c) challenges and measures, and (d) challenges. Note that count "1" does not refer to the number of occurrences, but rather to the number of times they are mentioned.

Artificial Intelligence/AI

As shown in Table 2, "Artificial Intelligence/AI" is heavily used in 1-3 "Growth Strategy 2017", 3-1 "Annual Report on the Japanese Economy and Public Finance" and 9-1 "White Paper Information and Communications". Conversely, “The White Paper on Police” (5-1), “Financial Services Agency Annual Report” (7-1), “Highlights of the Draft FY2018 Budget” (10-6), “the White Paper on Education, Culture, Sports, Science and Technology” (12-2), and “Annual Report on Transport Policy” (16-1) and “the White Paper on Land, Infrastructure, Transport and Tourism” (16-2) barely mention the term AI. These white papers do use other key words, apart from 7-1 "Financial Services Agency Annual Report" which did not use any of the key words. For example, 16-2, "White Paper on Land, Infrastructure, Transport and Tourism," uses words such as automatic driving, drones, and big data.

Positioning

AI has been discussed as a subject that should be used for the realization of Society 5.0 and the fourth industrial revolution. What is also characteristic is that it is not just AI, but also big data and other technologies that are treated in the same way.

Through the use of AI and IoT and the creation of innovation (omitted), we need to work towards the future vision of an ultra-smart society (Society 5.0). (1-3 "Growth Strategy 2017" pp. 90-91)

New technologies in the fourth industrial revolution, such as IoT and AI, can be seen as an extension of the digital economy (3-1 "Annual Report on the Japanese Economy and Public Finance," pp. 149)

The utilization of big data is the key to realizing data-driven economic growth and social transformation. The Internet of Things (IoT) is the means for collecting big data, and Artificial Intelligence (AI) is the means for analyzing and utilizing big data. (9-1 "White Paper on Information and Communications," pp. 52)

The swell of the fourth industrial revolution across industries, driven by technological innovations such as IoT, big data, artificial intelligence, and robotics, is bringing signs of change to every corner of society in the form of innovative products and services that transcend national borders. (15-1 "White Paper on International Economy and Trade," pp. 209)

The fourth industrial revolution, which leverages technological breakthroughs in the Internet of Things (IoT), big data, artificial intelligence, robots and sensors, is expected to create new businesses that will solve social problems and awaken the latent needs of various stakeholders. (17-1 "Annual Report on Current Status of Meteorological Service," p. 21)

In order for the Japanese economy to grow, it is important to improve added value and promote efforts to resolve supply constraints. To this end, it is necessary to respond to innovations such as the IoT and AI in the fourth industrial revolution, to increase the added value produced per worker, in other words, labour productivity, and to respond to changes in the environment surrounding working styles, such as the participation of women and the elderly in the workforce. (13-2 "White Paper on the Labour Economy," pp. 71)

Areas of application and benefits

Various fields of application for AI as shown in Table 2 are exemplified, including medical, health, and finance. Medical care, for example, is dealt with in 13-1 "Annual Health, Labour and Welfare Report". On the other hand, 1-3 "Growth Strategy 2017", 3-1 "Annual Report on the Japanese Economy and Public Finance" and 9-1 "White Paper on Information and Communications" provide many examples without stopping at any specific sector.

In addition, the reports flag as benefits changes that have been revealed through use scenarios in “Business/commercialization,” as well as that of “Labour/Working styles.”

The realization of Society 5.0 will increase the ways of working that are not limited by time and space. In addition to the possibilities of people developing their own abilities, and implementing their own work styles, through collaboration with AI, robots and other machines, we could also see "smarter ways of working" through advanced telework that utilizes virtual reality, augmented reality and other forms of ICT. (3-1 Annual Report on the Japanese Economy and Public Finance, pp. 191)

Although there are concerns that AI and robots will take jobs away, the results of the analysis in Chapter 3 show that, despite the possible impact on some occupations and skilled workers, if many new goods and services are created to replace existing goods, the productivity gains from these product innovations can have the effect of increasing employment. (3-1 Annual Report on the Japanese Economy and Public Finance, pp. 207)

Part II of this white paper also points out that the use of technological advances through the realization of innovation activities, including AI, can contribute to the realization of a work-life balance by creating flexible work styles such as non-employment-based work, furthermore it has been pointed out that the realization of a work-life balance can lead to the effective use of human resources and create a virtuous circle of innovation activity. (13-2 White Paper on the labour Economy pp.172)

The fourth industrial revolution is bringing about a major change to the nature of employment. The fourth industrial revolution has (1) made it possible for all kinds of businesses and information in the real world to be freely exchanged through data and networks, and (2) for large amounts of data to be analyzed and used in ways that create new value. In addition, (3) artificial intelligence has made it possible for machines to learn and make sophisticated decisions that surpass those of humans, and (4) automation of diverse and complex tasks has become possible. These technological innovations have made it possible to realize a society that was previously thought to be unfeasible, and have given rise to the possibility of dramatic changes in the structure of industry and employment. (15-1 White Paper on International Economy and Trade pp. 316)

Challenges and measures

Challenges to be addressed and measures to be taken to reap the benefits listed in (b) include items related to the development of technology and research, human resources, and intellectual property. Various examples of the creation of mechanisms for international/overseas expansion and collaboration were also mentioned.

As in (b), 1-3 "Growth Strategy 2017" and 9-1 "White Paper on Information and Communications" refer to challenges and measures, but in addition, 12-1 "White Paper on Science and Technology" is unique in that it covers various measures.

In particular, human resource development is under the jurisdiction of the Ministry of Education, Culture, Sports, Science and Technology, and is described as follows.

In order to develop human resources with a basic knowledge of mathematics and data science, which is necessary to invigorate Japan's industrial activities in the future, we will strengthen education for mathematics and data science, without distinction for humanities or the sciences, at universities. The purpose of this is to develop human resources with mathematical ability, that can analyze and utilize data, solve various social challenges, develop new ideas, and

create new technologies. In addition, we will enrich practical education such as project-based learning by forming a practical education network through industry-academia cooperation and promote the development of systematic education programs for the re-education of working adults in order to strengthen the information technology human resource development functionality of universities. Furthermore, based on the growing expectations for engineering education, which plays a central role in human resource development for the fourth industrial revolution and the realization of society 5.0, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) has convened the "Review Committee on the Engineering Education in Universities" since January 2017. The committee is discussing the ideal educational system and curricula of undergraduate and graduate schools in engineering education and the ideal approach for industry-university cooperative education, which is necessary to reform the engineering education system so that it can flexibly respond to changes in the industry-academia structure (12-1 "White Paper on Science and Technology", p. 180).

Although 1-3 "Growth Strategy 2017" and 9-1 "White Paper on Information and Communications" and 13-1 "Annual Health, Labour and Welfare Report" refer to "laws", "institutions" and "regulations" to introduce initiatives for social implementation, only 12-1 "White Paper on Science and Technology" mentions "ethical, social and legal issues".

At the newly established RIKEN Center for Advanced Intelligence Project, under the Ministry of Education, Culture, Sports, Science and Technology, there is collaboration with related ministries, companies, universities, and research institutes to construct innovative basic artificial intelligence technologies for the next 10 years. In addition, they are seeking to further develop fields in which Japan has strengths, such as iPS cells and manufacturing, and promote the development of the healthcare and disaster prevention sectors, and the development of new technologies. At the same time as conducting applied research to solve social issues in Japan, such as infrastructure, we also conduct research on the ethical, legal, and social issues that arise with the spread of artificial intelligence technology (12-1 "White Paper on Science and Technology," p. 182).

In addition, 13-1, Annual Health, Labour and Welfare Report, refers mainly to councils and institutions.

In January 2017, the Ministry of Health, Labour and Welfare established the Task Force on Promotion of Data Health headed by the Minister of Health, Labour and Welfare to study the implementation of ICT infrastructure which organically links the health, medical and nursing care sectors, with a view to its full-scale operation in FY2020. (13-1 “Annual Health, Labour and Welfare Report," p. 153).

Artificial intelligence (AI) is expected to enable the creation of new diagnostic and treatment methods, the development of an environment in which people can receive the most advanced medical care anywhere in Japan, and a reduction of the burden on medical and nursing care workers. Therefore, the Ministry of Health, Labour and Welfare (MHLW) held the Health Reform Promotion Headquarters to discuss necessary measures for the utilization of AI in the health and medical fields (13-1 “Annual Health, Labour and Welfare Report," p. 163).

14-1 The "Annual Report on Food, Agriculture and Rural Area" also mentions that the Artificial Intelligence Future Agriculture Creation Project is being implemented to accelerate the research and development of AI technology in the agricultural sector, and to put it to practical use as soon as possible.

Challenges

While (c) shows the challenges and measures, (d) is a list of items that only present the challenges. As can be seen from Table 2, (c) is more common in 1-3 Growth Strategy 2017, while (d) is more common in 3-1 Annual Report on the Japanese Economy and Public Finance. This can be attributed to the different nature of the reports.

For example, the "1-3 Growth Strategy 2017" for "Human Resources" is described in (c) as follows, indicating measures to address the issue:

We will consider specific revisions for reforming the engineering education system to develop the necessary human resources, based on information technology to promote industrial structural reform such as AI, IoT, and Big Data to create new industries, and by fundamentally reviewing the vertical segmented structure of departments and faculties, for example by making educational periods, such as the 6-year integrated system for bachelor's and master's degrees, more flexible and enabling students to acquire sub-specialties in addition to their main discipline, by the end of this fiscal year. The objective is to implement revisions in phases from the next fiscal year, with full implementation from FY 2019. (1-3 "Future Investment Strategy 2017," pp. 92-93)

On the other hand, the "3-1 White Paper on Economic and Fiscal Policy" focuses on only (d) challenges, as described below. Although it states the need for measures, it does not go so far as to indicate what specifically.

In the United States and other developed countries, the phenomenon of relatively high wages for highly educated people, for whom demand is increasing, has been reported due to changes in the economic structures, and other factors such as a shift to information technology. In Japan, the percentage of students going on to higher education institutions such as universities, junior colleges, and special training colleges (vocational schools) is increasing, and so far the wage gap has not widened as there is excess demand. However, as technological innovation continues to progress, we need to secure highly educated human resources who have mastered the most advanced technologies and who can add value. In doing so, the cost-effectiveness of education will be enhanced by putting in place a system that prioritizes the development of needed human resources. This is in light of the economic and social structural changes within Japan, such as the increase in demand for the healthcare industry at home and abroad, the rise of inbound tourism demand, and the establishment and deepening of AI and IoT in the economy and society. Investment in human resources is important regardless of the age of employees. It is important to strengthen the growth potential of the economy as a whole by promoting the learning of working people (recurrent education) and encouraging the migration of labour to high value-added industries, while improving the employability of workers and raising wages. ("3-1 White Paper on Economic and Fiscal Policy" p. 83)

With respect to "work/employment," which was listed as a benefit in (b), the 13-2 White Paper on Labour Economy also shows some recognition of the challenges.

Other technology keywords

Six key words related to AI are discussed in this survey: big data, IoT, robotics, drones, automated driving/automated driving, and VR/AR/MR. Table 1 shows that all technologies are mainly mentioned in 1-3 Growth Strategy 2017, 3-1 Annual Report on the Japanese Economy and Public Finance, 9-1 White paper on Information and Communications, and 16-2 White paper on land, infrastructure, transport and tourism.

Therefore, the white papers and reports are categorized based on similar sub-labels of what themes are being discussed for each technology, not by ministry. Specifically, Tables 3 to 5 list (b) areas of application and benefits, (c) challenges and measures, and (d) challenges, respectively.

(a) Comparison of areas of use and benefits

Table 3 shows the list of areas of application and benefits. Table 3 shows that "artificial intelligence/AI," "big data," "IoT," and "robotics" are used in a comparatively wide variety of fields. Among these six words, “IoT” appeared most frequently alongside “AI”, indicating a wide range of applications, including "construction site", "tourism" and "food".

With the exception of VR/AR, which appeared less frequently, "logistics" and "addressing labour shortages" were mentioned in almost all of the keywords. Furthermore, when “drones” were excluded, it became clear that "nursing care" and "transportation" were also common focus areas.

(b) Challenges and Measures

Table 4 shows a comparative list of (c) challenges and measures. For all the technologies, promotion of technology development and research and development is mentioned, as well as the need for standards, institutions, and regulations; with the exception of VR/AR, it can be seen that "social implementation and demonstration" is taking place, as well as "practical application / entrepreneurs" and "collaboration / networking".

(c) Challenges

Table 5 shows the list of (d) challenges. “Artificial Intelligence/AI" and "IoT" share many of the same challenges. "Research and development," "regulations and standards," "human resources," "enterprise / start-ups," and "global competition" are also listed as challenges.

It is interesting to note that mention can be found in the reports on handling the psychological and emotional side of users and the general public on concepts such as a sense of security, which are common to "robots," "automated driving" and "drones."

It has been pointed out that the resistance of some people to nursing care robots is one of the reasons why nursing care facilities have not been able to introduce them, and it is important to gain the public's understanding of new nursing care systems using robots in the future. If nursing care robots become more common then this could potentially change people’s attitudes (see column). In the above "New Robot Strategy", the government has set a target of increasing the percentage of people who "want to use" and "want elderlies to use" nursing care robots to 80% each. (13-1 Annual Health, Labour and Welfare Report, pp. 166)

As shown in public awareness surveys and the aforementioned efforts to implement automated driving in society, in order for new technologies and services to be accepted by society, it is necessary to secure technical safety, as well as to provide users and society with a sense of security through rules and social experiments. For example, the use of small unmanned drones and other small-scale unmanned vehicles in the logistics business, and improvements in performance and system enhancement, is working to improve the environment for commercialization while relieving the public's anxiety about safety. (16-2 White paper on land, infrastructure, transport and tourism, pp. 89)

(d) No mention of Artificial General intelligence / Singularity / Quantum

Artificial General intelligence and singularity were not mentioned in any of the white papers and reports. Quantum was mentioned in 19-1 "Trends in Military Science and Technology" in the "Defense of Japan".

In recent years, new ICT developments have been made. For example, in August 2016, China launched a satellite called Mozi which is the first in the world to test quantum-encoded communications. In addition to quantum-encoded communications, new technologies such as artificial intelligence (AI) and big data analysis may be applied to the military sector by various countries in the future (19-1 "Defense of Japan," p. 227).

The Defense of Japan basically talks about the usefulness of information and communications technology (ICT) in defense, and technologies such as AI and big data analysis are also positioned as "new technologies," like quantum communications.

Society 5.0 / The Fourth Industrial Revolution / Connected Industries

The definition of Society 5.0 is: "A new society that is the fifth society in human history, following (1) a hunting society, (2) an agrarian society, (3) an industrial society, and (4) an information society. New values and services will be created one after another, bringing affluence to the people who are the main actors in society. 1-3 Growth Strategy 2017 (pp. 1-6) is being used as the Japanese government’s future vision moving forward. The difference between this strategy and the fourth industrial revolution is also indicated in the same “Growth Strategy 2017” report.

Industry 4.0 in Germany and Industrial Internet in the U.S. are mainly attempts to use IoT to optimize production and inventory management in manufacturing industries beyond the framework of individual factories and companies. In this regard, Japan must realize Connected Industries that go beyond manufacturing to connect goods to goods, things to things, people to machines and systems, people to technology, companies to companies in different industries, people to people across generations, and manufacturers to consumers. The Society 5.0 that Japan is aiming for is an attempt to solve a variety of social issues by incorporating cutting-edge technology into all industries and social life, and providing needed goods and services to the people who need them, when they need them and in the quantity they need. (1-3 Growth Strategy 2017 pp.1-6)

A defining characteristic of Society 5.0 is that it is one which is "yet to be seen". For this reason, the description of (b) areas of application and benefits of Society 5.0 is a process discussion for "when Society 5.0 becomes a reality".

The realization of Society 5.0 will increase the ways of working that are not limited by time and space. In addition to the possibilities of people developing their own abilities and implementing their own work styles through collaboration with AI, robots, and other machines, we could also see "smarter ways of working" through advanced telework that utilizes virtual reality, augmented reality, and other forms of ICT. (3-1 Annual Report on the Japanese Economy and Public Finance, pp. 191)

For this reason, (c) challenges and measures and (d) challenges are presented in terms of how to achieve Society 5.0, and there is little concrete discussion of (b) areas of application and benefits, and this creates an abstract discussion. As a result, inevitably, most Society 5.0 discussions are about (c) challenges and measures and (d) challenges. Specifically, we can flag the following labels of "research and development / core technology," "innovation," "human resources," "intellectual property," "overseas expansion," "institutions," "regulations," "collaboration," "implementation," "standards," "evaluation," and "demonstration," as listed in (1) "Artificial Intelligence/AI" subsection labels.

The Fourth Industrial Revolution is often used in parallel with Society 5.0, but as Table 1 shows, the Fourth Industrial Revolution is used more frequently in white papers and reports than Society 5.0. In particular, it is used more in 9-1 "White Paper on Information and Communications" and 15-1 "White Paper on International Economy and Trade". This is because many reports describe Society 5.0 as coming after the Fourth Industrial Revolution, and so descriptions of the Fourth Industrial Revolution are more prevalent[338].

Unlike Society 5.0, the Fourth Industrial Revolution is seen as an event that has already happened.

Thus, firms that are actually making use of the new technologies in the fourth industrial revolution are feeling more successful in product innovation and increased sales capabilities than in the efficiency aspects of production. (3-1 "Annual Report on the Japanese Economy and Public Finance," pp. 171)

Table 6 shows the percentage of mentions of Society 5.0 and the Fourth Industrial Revolution by sub-item label: while Society 5.0 has only 6% of mentions in (b) areas of application and benefits, the Fourth Industrial Revolution has 15% of mentions in (b) areas of application and benefits.

[338] On the other hand, 12-1, "White Paper on Science and Technology", makes extensive use of the term Society 5.0 only. In addition, 10-6, "Highlights of the Draft FY2018 Budget," also uses the term Society 5.0 instead of the term Fourth Industrial Revolution. The term Society 5.0 is also often used in parallel with the term super-smart society.

III. Comparison of AI governance in Japan and abroad

i. Comparison of the roles of different actors

This chapter examines what kind of reports are published by actors in Japan in comparison with the other countries.

The role of academic societies and universities

In Japan, the Japanese Society for Artificial Intelligence (JSAI) started discussions on ethics at a relatively early stage, in 2014, and published its Ethical Guidelines. It also collaborated with IEEE and TFS[339], however, activities are mainly directed at domestic members and the society does not actively engage with the outside world. For example, in December 2019, three academic societies including JSAI that were engaged in machine learning released the "Statement on Machine Learning and Fairness," and declared that (1) machine learning is nothing more than a tool to assist human decision making, and (2) machine learning researchers are committed to improving fairness in society by studying the possible uses of machine learning[340].

In addition, in comparison to foreign countries, university initiatives, research reports and proposals conducted mainly by university centers and faculties, and especially discussions, including those from a humanities and social sciences perspective, are not published in English very much[341]. Research is currently left to individual researchers.

The role of industry associations

Industry collaboration is essential to creating best practices, and many international industry associations working on best practices, such as the Partnership on AI and the Information Technology Industry Council (ITI), are based in the U.S.. While some companies that particularly need to take action are from the U.S., many of them are creating their own governance rules. A 2018 Accenture survey showed that of the companies that have adopted AI (72% of companies surveyed), 70% have provided ethics training for technicians and 63% have established an ethics committee to evaluate the use of AI[342], and this shows the voluntary efforts that are required of companies[343].

At present AI guidelines are mostly discussed at the initiative of Big Tech companies in the U.S. and China, and there are not so many voices from start-up companies, which have less resources to participate in these kinds of discussions. The industry structure counts also in terms of business types. There are many Business to Business (B2B) companies in Japan, compared to Business to Consumer (B2C) companies. The whole structure is like B2B2B2B2B2…B2C and the supply chain is very long in Japan. Therefore, Japanese companies need to consider accountability, reliability and responsibility along with this very long industry supply chain.

International organizations such as the ITI include many large Japanese companies, so it is expected that discussions within an international framework will be stepped up in the future. In

[339] It should be noted, however, that it was over the "cover issue" that started the debate, and that awareness of the issue is slightly different from other countries.
[340] This statement was released in response to a hate speech by an AI researcher at the University of Tokyo. https://arisaema0.wordpress.com/2020/01/08/ai-governance-in-2019/
[341] "Perspectives on Artificial Intelligence/Robotics and Work/Employment," a "Science and Technology Research Project" of the Research and Legislative Reference Bureau (RLRB), the National Diet Library, includes research and technology trends in the first part, the second part covers the impact on employment in eight fields, including medical care, nursing care, education, and agriculture, and the third part covers overseas trends in AI and employment, and human resource development, utilization, and management. This study was commissioned by the university and substantial writing was done by the university faculty. An English translation of the study is available on the AIR website (http://sig-air.org/publications/perspectives-on-ai).
[342] Accenture finds that companies are stepping up efforts to use artificial intelligence ethically and responsibly - SAS, Accenture, Intel and Forbes Insights Latest Survey, October 23, 2018, https://www. accenture.com/jp-en/company-news-releases-20181023
[343] The IEEE's "Ethically Aligned Design, Second Edition" also proposes the need for an ethics committee or employees that can cover everything in terms of value, such as a Chief Value Officer (CVO).

addition, the Japan Deep Learning Association, founded in 2017, is an organization dedicated to promoting the business applications of deep learning, and also offers certification exams as part of its human resource development. The G-test, which is aimed at human resources for business use (generalists), includes questions on industrial applications, law, ethics, and current social problems. In addition, a study group on AI governance and its assessment started being held in July 2020. The study group is considering a framework, and an ecosystem network, involving insurance companies, auditing companies as well as whistleblowing systems and third-party committees for accident investigation. The study group is also considering AI risk mitigation ecosystems, which includes not only big tech companies but also start-up companies.

It is hoped that present discussions will provide input for the ethics and social perspectives of Japan's venture companies, especially in terms of self-governance in industries.

The role of governmental organizations

In many countries, human resource and research development are important national strategies for AI, and therefore, the challenge is how to create a center of collaboration between industry, academia, and the private sector, and how to drive discussions by demonstrating leadership.

It should be noted that although Japan is promoting research which is not only limited to science and technologies but also includes social sciences and the humanities, investment in research institutions and human resource development from the government, through the budget of the Ministry of Education, Culture, Sports, Science and Technology is lower than in other countries[344]. Instead, the Institute for Information and Communications Policy (IICP), which has been stepping up conferences since 2016, brought together a diverse range of stakeholders to form a network for discussions. The network consists of dozens of people from different fields and industries.

However, in the case of Japan, there is a strong tendency to directly link the assessment of the social impact of technology with strategy formulation and policy decision-making on technology, and this means that distance and independence between the two is not maintained. The Conference toward AI Network Society, established by the IICP of the Ministry of Internal Affairs and Communications (MIC), set up two separate subcommittees, the Committee on AI R&D Principles and the Committee on the Committee on Impact and Risk Assessment, but the emphasis in both committees was placed on the preparation of principles and guidelines.

It has been also been suggested that discussions aimed at a non-Japanese audience and discussions aimed at a Japanese audience are not adequately connected. In terms of international presence, Japan has been active at OECD meetings on AI, starting with the G7 meeting in Takamatsu in 2016, and with the Ministry of Internal Affairs and Communications co-hosting subsequent OECD meetings as well. Principles on AI was one of the focal points at the G20 Digital Economy Ministers in Japan in 2019.

On the other hand, discussions on AI principles in the Cabinet Office and the Ministry of Internal Affairs and Communications, which have formed the basis for discussions aimed at non-Japanese audiences, were not necessarily linked to discussions within government ministries and agencies that took place in II-1(5), and the analysis of the White Paper does not address many ethical or social issues.

[344] The government's overall budget for AI in the overall FY2018 budget totals 77 billion yen, reportedly less than 20% of the U.S. (500 billion yen) and China (450 billion yen), Japan Times, Japan's budget for AI to be less than a fifth of that planned by U.S. and China, February 25, 2018, https://www.japantimes.co.jp/news/2018/02/25/business/tech/japanese-government-spending-ai-less-20-u-s-china/

ii. Trends and issues in focused AI technologies and fields

Intensity of Discussions on specific AI technologies in each ministry

In Section II of this report, not only AI technology but also peripheral technologies and concepts were surveyed at the same time. Table 1 shows that "Artificial Intelligence/AI" is mentioned in many white papers and reports alongside "IoT" and "Big Data". Among them, these keywords are particularly mentioned in the Cabinet Secretariat's 1-3 Growth Strategy 2017 and 3-1 Annual Report on the Japanese Economy and Public Finance, and the 9-1 White Paper on Information and Communications by the Ministry of Internal Affairs and Communications. The Institute for Information and Communications Policy (IICP) of the Ministry of Internal Affairs and Communications has been holding the Conference toward AI Network Society, and the findings and discussions from this conference have also been siphoned off into the MIC's White Paper on Information and Communications.

In the Strategic Council for Artificial Intelligence Technology, in addition to the core three ministries (Ministry of Internal Affairs and Communications, Ministry of Education, Culture, Sports, Science and Technology, and Ministry of Economy, Trade, and Industry), research and development and social implementation are currently being promoted in cooperation and collaboration with related ministries and agencies such as the Cabinet Office, the Ministry of Health, Labour and Welfare, the Ministry of Land, Infrastructure, Transport and Tourism, and the Ministry of Agriculture, Forestry and Fisheries. The comparison of the reports and white papers in Table 2 shows that the White Paper on Information and Communications discusses the fields of utilization, benefits, challenges and measures in a broad manner. In contrast, the White Paper on Science and Technology by the Ministry of Education, Culture, Sports, Science and Technology focuses on issues and measures related to technological research and development, human resources and collaboration. Meanwhile, the Ministry of Economy, Trade and Industry's White Paper on International Economy and Trade makes little mention of AI, and only touches on issues and measures related to employment, working styles, and innovation.

In turn, Table 1 shows that the White Paper on Science and Technology is characterized by its frequent references to "Society 5.0" and the White Paper on International Economy and Trade by its frequent references to the "Fourth Industrial Revolution." “Society 5.0" and "the Fourth Industrial Revolution" are key words in describing the vision of the society we should be aiming for, and AI technology is just one technology that is contributing to the realization of that society. The White Paper on Land, Infrastructure, Transport and Tourism makes a heavy use of, and therefore has a particular emphasis on, "automatic driving / automated driving," "drones" and "big data" rather than "artificial intelligence/AI”. The Annual Health, Labour and Welfare Report and the White Paper on Food, Agriculture and Rural Area do mention "artificial intelligence/AI", but more often than not there are references to "robots".

AI technology is sometimes utilized in the software of some hardware, such as self-driving cars, drones, and robots, so it is not as if they are unrelated, but there are some variations in the types of technologies which different ministries focus on.

Peripheral technologies and fields of AI

Table 3 shows that the areas of application and benefits of AI peripheral technologies are medical/health/nursing care, transportation, logistics, disaster prevention, and agriculture. These are fields in Society 5.0 that require the integration of cyber and physical space, and they are also said to be Japan's strengths.

On the other hand, financial services (fintech), housing (smart homes and assistants) and tourism are less likely to be taken up extensively than in other countries in terms of data accumulation and infrastructure development. In addition, technological developments related to national defense and security are discussed in many industry-academic-government discussions abroad, while in Japan they are only discussed by the Ministry of Defense. Relatedly, Lethal Autonomous Weapons Systems (LAWS), which is being discussed at the United Nations, is rarely discussed outside of the Ministry of Foreign Affairs in Japan, although

it was mentioned in the Diet in 2018. These may be due in part to the fact that the Ministry of Defense, the Ministry of Foreign Affairs, and the Financial Services Agency are not very involved in the relevant bodies of the Strategic Council for Artificial Intelligence Technology.

And while words like IoT and big data are used along with AI, there is little mention of VR and quantum, which were said to be emerging technologies. There is also no mention of words like singularity and artificial general intelligence.

Insufficient understanding of the current situation

There is a need for data-driven discussions to advance AI governance. Therefore, governments, universities and research institutes in many countries are evaluating the social impact of technology, and are creating indicators for this. It depends on what kind of indicators are to be created, but at least in Japan, the problem is that there is a lack of data to understand and evaluate the current situation, although objectives have been set.

However, as shown in Section II of this report, while almost all ministries and agencies discuss human resource development, such as "Develop human resources with a basic knowledge of mathematics and data science," there is no data for understanding the current situation, although this issue is mentioned as an important target. Although there are estimates that IT human resources are likely to be in short supply in the future[345], and the Entrepreneurial Activity Index TEA (2014) does set out the ratio of entrepreneurial human resources[346], there is not a collected list of data for understanding the current situation.

The Cabinet Office's "Draft AI Strategy (Overall Overview)" includes a graph on "Global AI Investment, R&D, and Human Resources," but the data for the "Number of University Graduates Trained in Data Analysis" is from 2008, more than 10 years ago[347]. While it is necessary to examine who creates which indicators, at a basic informational level the creation of indicators is needed to assess the impact of technology and identify potential needs. This is certainly an issue for the future. The development of such basic information will also serve as a basis for making the technology assessment function somewhat independent of strategy development and policy making.

iii. How to create a forum for discussion and its challenges

The place and role of collaboration by each actor

Notably, industry, academia and the private sector in the US are all proactive in implementing discussions, and in fact industry and academia are leading the way in shaping debates rather than being led by the government. In addition, since the IEEE and ISO, which are international actors, are Western-centric, it is necessary to consider how Japan participates in international platforms, and whether Japan is able to form its own platform.

In Japan, the government has been promoting the creation of principles under its own initiative. However, principles only have an abstract existence, and incorporating them into practice requires collaboration and cooperation between all fields and industries. The nature of technology governance differs by sector[348]. In Japan, ministries and agencies have started discussions with an eye for implementation, and for example, the Japan Medical Association has summarized their own discussions. However, there is a need to form a forum where industries can cooperate to formulate the structures and best practices for technology governance, and to create and disseminate standards. At present, consortia are formed by field, but it is necessary to consider the future of governance, and whether it will be like that found in Europe or the U.S., which involve various stakeholders, or in other forms.

---

[345] 15-1 White Paper on International Economy and Trade pp. 235-6
[346] 9-1 White Paper on Information and Communications pp.111
[347] The original document from the Cabinet Office (https://www.kantei.go.jp/jp/singi/tougou-innovation/dai2/siryo1.pdf) is the Ministry of Internal Affairs and Communications' 2014 White Paper on Information and Communications (http://www.soumu.go.jp/) (johotsusintokei/whitepaper/en/h26/html/nc134020.html) (in Japanese)
[348] Shiroyama Hideaki, Science, Technology and Politics, Chapter 7

In addition, governance through cooperation among actors is not only necessary for each field, but it is also important for the realization of what kind of society will be built by using technology. In Japan, Society 5.0 has been mentioned in general terms, but a vision for this society has yet to materialize. It is also necessary to establish a forum for cooperation among actors as a mechanism to flexibly consider what kind of society should be aimed for.

Necessity of discussions involving the general public

Non-profits and non-governmental organizations are essential in shaping spaces for discussions that involve a diverse range of stakeholders. However, there are not so many nonprofits organizations in Japan compared to other countries. The Japanese Society of Artificial Intelligence (JSAI) has contributed to the Future Society's online discussions. While designing an online dialogue can be difficult in some respects, it is an interesting initiative as a way to encompass diverse stakeholders. In Japan, these kinds of activities, which link public discussions to policy recommendations, is still treated as a public comment process.

In other fields, citizen-participatory dialogue events have been held in Japan in the past to make policy recommendations on subjects such as genetically modified plants, nanotechnology, and biodiversity[349]. With regard to robotics, one event was held as an "interactive public comment" event[350]. It is also necessary to consider the ideal form of governance that involves discussions among many stakeholders.

IV. Conclusions

This report outlined what social issues are being addressed by various stakeholders with regard to AI governance, and examined the characteristics and the challenges in Japan.

The leaders on governance, and connected challenges, also change depending on the country and sector. In fields where institutions and regulations play a major role, such as healthcare and transportation, national governance tends to lead the way. So it is necessary to develop trials using a regulatory sandbox. In addition, as the services provided by GAFA are developing globally and technology is advancing rapidly it is necessary to collaborate with various entities from the bottom up, not from the top down. Therefore, in order to advance discussions on AI governance, it is important to establish a mechanism and a forum for continuously thinking about how far the same principles and rules can be applied globally and locally (in each country and region), what kind of stakeholders need to be involved, and who should take the initiative.

Acknowledgements

This report is based on a translation from a chapter of "AI Governance" written by Arisa Ema and Hideaki Shiroyama in Artificial Intelligence, Humanity and Society (Keisho Shobo, 2020)[351]. We thank Ms. Chihiro Saito and Mr. Hideaki Kojima, Mr. Yoichi Iida, Mr. Koichi Takagi, and Dr. Takashi Egawa for supporting our work to fit into the CAHAI report. We thank Ms. Haruka Matsumoto, a graduate student at the University of Tokyo (at that time), for her assistance in carrying out the survey of this report. This research is part of the results of the FY 2017 Science and Technology Research Project of the Research and Legislative Reference Bureau (RLRB): "Perspectives on Artificial Intelligence/Robotics and Work/Employment," and the Grant-in-Aid for Scientific Research (A): International Governance of Emerging IT and Biotech- Role of Information Sharing and Private Actors (Grant Number 18H03620)."

[349] Tadashi Kobayashi, "Who's Thinking About Science and Technology?" (2004), Nagoya University Press, Nobuko Kori, et al.: 'From Participating Observations of WWViews Debate Process on Biodiversity' Science Communication, 13: 31-46, 2013.
[350] The robots are coming! Symposium on "Robot x Future x Dream Vision" - A bridge between "needs for robots" and "implementation in society",
http://interactive.pesti.jp/robot/wp/wp-content/uploads/F01_Introduction.pdf
[351] Artificial Intelligence, Humanity and Society (Keisho Shobo, 2020), https://www.keisoshobo.co.jp/book/b498075.html (in Japanese)

Annexes

Table 1

	Artificial Intelligence/AI	Automatic driving and automated driving	Drones, Unmanned Aerial Vehicles, and Small Unmanned Aerial Vehicles	Robot	Big Data	IoT	Society 5.0	Fourth Industrial Revolution	Connected Industries	VR, AR	Quantum	MR	Singularity	AGI
1-3	**45**	**37**	**10**	**34**	**28**	**47**	7	**15**	2	4	0	0	0	0
**1-4	Unidentified	Unidentified	0	Unidentified	Unidentified	**14**	1	Unidentified	Unidentified	1	0	0	0	0
3-1	**27**	2	0	**16**	9	**21**	**10**	**18**	2	0	0	0	0	0
4-1	0	2	0	0	0	0	0	0	0	0	0	0	0	0
*7-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9-1	**70**	8	0	**14**	**37**	**54**	8	**32**	1	5	0	0	0	0
10-6	0	0	0	0	1	1	1	0	0	0	0	0	0	0
*12-1	7	2	0	3	1	6	**32**	2	0	0	0	0	0	0
*12-2	0	0	0	0	1	0	0	1	0	0	0	0	0	0
13-1	7	0	0	9	7	5	1	4	0	0	0	0	0	0
13-2	9	0	0	1	2	1	0	5	0	0	0	0	0	0
14-1	5	3	2	**13**	4	6	0	0	0	0	0	0	0	0
15-1	8	1	0	9	5	9	1	**28**	4	0	0	0	0	0
16-1	0	5	0	0	1	0	0	0	0	0	0	0	0	0
*16-2	0	**13**	**10**	6	**12**	6	0	1	0	1	0	0	0	0
17-1	1	0	0	1	3	4	0	1	0	0	0	0	0	0
18-1	1	0	0	0	3	1	1	1	0	0	0	0	0	0
19-1	1	0	1	1	1	0	0	0	0	0	1	0	0	0
20-1	2	0	0	0	0	0	0	0	6	0	0	0	0	0

* Means that a word search was not run on the entire white paper, only a portion of the white paper was targeted.

** Means that because a normal word search was not possible due to glitches in a PDF, only the number that could be counted by eye were counted.

Table 2

Major Item	Subdivision	1-3	**1-4	3-1	4-1	*7-1	9-1	10-6	*12-1	*12-2	13-1	13-2	14-1	15-1	16-1	*16-2	17-1	18-1	19-1	20-1
(a) Positioning.		1	0	1	0	0	1	0	1	0	1	1	1	1	0	0	1	0	0	1
(b) Areas of application and benefits	Medical care	1		1			1				1									
	Health/Healthcare	1									1									
	Financing			1			1													
	Universe / Space	1					1													
	Nursing care / elderly support	1		1							1			1						
	Traffic / automobile / automated driving			1			1													
	Sports	1																		
	Manufacturing	1					1													
	Logistics / land transportation	1		1			1													
	Environment / urban development																	1		
	Weather						1													
	Infrastructure / compute environment	1					1													
	Disaster	1																		
	Administration	1																		
	Biotechnology	1																		
	Research and development	1		1			1													
	Business / business development	1												1						
	Labor / labor force / labor shortage / workplace / employment			1			1					1	1	1						
	Innovation / creation			1			1							1						
	Impact						1													
	Data			1			1				1			1						

(c) Challenges and measures	Technology development / research and development	1					1		1				1							
	Innovation / creation	1					1													
	Human resources	1					1		1				1							
	Intellectual property	1							1											
	Infrastructure	1					1													
	Ethical, legal and social issues								1											
	Law	1					1													
	Connectivity / networking	1					1		1											
	International / international expansion	1					1													
	Support / promotion	1					1													
	Practical use / commercialization	1																		
	Evaluation	1																		
	System	1					1				1									
	Standards / guidelines / regulations	1					1													
	Council	1					1				1									
	Symposium						1													
	Social implementation / demonstration	1					1		1											
	Data	1					1													
(d) Challenges	Research and development			1																
	Innovation / economic growth / transformation / creation			1			1							1						
	Human resources			1			1					1	1							
	Environmental improvement										1									
	Investment			1			1													

	Disparity			1																
	International						1												1	
	Networking			1			1													
	System	1		1			1													
	Regulations / guidelines			1			1													
	Organizational Change / organizational structure / corporate change			1			1													
	Employment / labor / labor force / labor market / labor shortage / working			1			1				1									
	productivity			1			1													
	Decentralizatio n (of power)			1																
	Entrepreneurs hip / start-up			1			1													
	Military affairs																		1	

Table 3

	Artificial Intelligence /AI	Big Data	IoT	Robot	Automatic driving and automated driving	Drone	VR/ AR
Medical care	1	1	1	1			
Health / healthcare	1	1	1	1			
Finance / FinTech	1	1					
Universe / space	1	1	1		1		
Nursing care / elderly support	1	1	1	1	1		
Traffic / automobile / automated driving	1	1	1	1	1		1
Sports	1	1	1				
Manufacturing	1		1	1			
Logistics / land transportation / ship / aircraft / cargo	1	1	1	1	1	1	

Environment / city Development / community	1	1	1		1	1	
Weather / climate change	1	1	1	1			
Infrastructure / compute environment	1	1		1		1	
Disaster / disaster prevention	1	1	1	1		1	
Administration	1	1	1				
Biotechnology	1						
Research and development	1				1		
Business / business development	1	1	1				
Labor / labor force / labor shortage / workplace / employment	1	1	1	1	1	1	
Innovation / creation	1	1	1	1	1	1	1
Impact	1						
Data	1	1	1				
Human resources		1		1			
Agriculture, forestry and fisheries / livestock / dairy		1	1	1	1	1	
Energy			1				
Sightseeing			1				1
Construction			1	1			
Residential / Housing			1				1
Foodstuffs			1				
Asset management				1			
International expansion					1		
Insurance					1		
Surveying						1	

Table 4

	Artificial Intelligence/AI	Big Data	IoT	Robot	Automatic driving and automated driving	Drone
Technology development / research and development	1	1	1	1	1	1
Innovation / creation	1	1				
Human resources	1	1	1	1	1	
Intellectual property	1	1		1		
Infrastructure	1					
Ethical, legal and social issues	1					
Law	1	1	1		1	
Connectivity / networking	1	1	1	1	1	1
International / international expansion	1	1	1	1	1	
Support / promotion	1	1	1	1	1	
Practical use / commercialization	1	1	1	1	1	1
Evaluation	1		1	1	1	
Standards / guidelines / regulations / Institutions / standards	1	1	1	1	1	1
Council	1	1	1		1	
Symposium / forum / events	1		1			1
Social implementation / demonstration	1	1	1	1	1	1
Data	1	1		1	1	
Advanced case study		1	1	1		
Manufacture			1			
Security			1		1	
Experiment / test bed			1	1	1	

Table 5

	Artificial Intelligence/AI	Big Data	IoT	Robot	Automatic driving and automated driving	Drone	VR/AR
Research and development	1	1	1		1		
Innovation / economic growth / transformation / creation	1		1			1	
Human resources	1	1	1	1			

Environmental improvement	1					1	
Investment	1		1				
Disparity	1						
International	1		1		1		
Network	1		1		1		
System	1						
Standards / guidelines / regulations / systems	1		1	1	1	1	
Organizational change / organizational structure / Corporate change	1			1			
Employment / labour / labour force / labour market / labour shortage / Working	1		1	1			
Productivity	1						
Decentralization (of power)	1						
Entrepreneurship/ start-ups	1	1	1	1			
Military affairs	1						
Global competition / competition	1	1	1	1			
Cashless		1					
Data		1	1		1		
Medical care		1					
Disaster		1					
Blockchain		1	1				
Hygiene		1					
Security			1	1			
Reassurance / acceptance / anxiety / public understanding				1	1	1	
Industrialization					1		
Experiment					1	1	
Implementation / demonstration					1		
Logistics					1		
Privacy / personal information					1		

Table 6

	Society 5.0		the Fourth Industrial Revolution	
	Number of labels	%	Number of labels	%
(a) Positioning	11	5%.	26	7%.
(b) Areas of application and benefits	14	6%.	54	15%.
(c) Challenges and measures	180	76%.	159	43%.
(d) Challenges	32	14%.	128	35%.
Total	237	100%.	367	100%.

CHAPTER III. AI-Applications in Mexico. A view from the inside

Jorge Cerdio[352]

I. Introduction

The purpose of this paper is to present some of the most pervasive and extended uses of Artificial Intelligence Applications (AI-Applications) in Mexico as well as the regulatory framework applicable to AI-Applications. We aim at representing with a high degree of accuracy the context under which each AI system operates. The context, plus a brief description of the system, will hopefully provide the reader with enough information to produce comparisons to other jurisdictions and to shed some light on the complexities surrounding the potential regulation of AI-Applications in Mexico and elsewhere. The main problem in presenting the Mexican case is the lack of a systematic source of information regarding AI-Applications. There are three main sources of information which are not equal in terms of the quality of the data contained nor in the relevance for the purpose of these paper. The first source of information is the academic works of different national research centers, scholars in the field of AI that the work across different local jurisdictions and institutions. The difficulty here lays on selecting the adequate literature to point to already mature AI-Applications and to discard early research projects or yet unproven solutions. The second source of information is the ecosystem of innovators and startups that are scattered around different cities across the Mexican territory; individuals and small corporations that emerge with business value propositions around operative solutions based on AI-Applications. Finally, the third source of information is the official source, that is, the source of the Mexican Federal and local governments that publish reports and announce public policy in different media (including official websites). Public information that amounts to public action in favor of Artificial Intelligence initiatives. The state of public data in Mexico is far from being that of full open data country making the harvesting of relevant public information a time intensive task for any researcher. But beyond the different levels of quality of information from these sources and therefore in discriminating examples of AI-Applications, we will aim at presenting an emerging puzzle, with pieces that fit together according to two main drives for the use of AI-Applications in Mexico. That is, instead of presenting groups of applications by family of technologies, we will be guided by the idea that most AI-Applications are the product of the need to solve a specific social and technical problem. This makes much more sense in a country like Mexico where innovating and pushing forward innovative ideas is twice as hard compared to other countries where there are more resources and ecosystems networked together to embrace new solutions to complex problems. At the same time, Mexico is nonetheless the 11th biggest economy in the world, and is the 2nd trade partner (by volume of exports) to the US market. Even in the face of profound inequalities in the Mexican society, economic forces keep incentivizing and rewarding the creation of AI-Applications from private and social sectors. We have reached the stage in Mexico where we can maturely ask ourselves how we want to regulate AI-Applications, and in which direction. You will see that collective intelligence is there in the Mexican arena to construct a national agenda for AI; an agenda which requires the

[352] Full Professor at the Law Department of the Instituto Tecnológico Autónomo de México (ITAM).

structuring and the leadership of public institutions. At the same time, from an economic point of view one might be tempted to say that unregulated markets for AI-Applications are best left alone. However, it is within these emerging ecosystems of innovators in Mexico that we register the generalize opinion that we need to regulate AI-Applications, bearing always in mind the principles of the Rule of Law, Human Rights and Democracy. The first part of this paper will deal with the presentation of salient examples of AI-Applications Mexico (I), after which we will look at the scarce regulatory framework, including public policy (II). We will reserve a brief last section for some conclusions.

II. Observing Public and Private AI-Applications in Mexico

In Mexico, the use of Artificial Intelligence (AI)[353] and the set of technologies associated to it are gaining traction in certain sectors of the Mexican economy and public programs. The momentum generated by the application of artificial intelligence in certain sectors of the Mexican economy is occurring against the backdrop of an under-developed country in terms of research and development expenditure. Mexico spends 0.3% of its GDP annually in innovation and technology[354]. Combined with the fact that 20% of Mexico's population lives with less than $5.50 US dollars a day at international prices, the explanation of why we have pervasive initiatives and AI-Applications is due to the specific policy led by the country's government combined with the need to seek innovation in the private sector. In these regards, there is not a normal, widespread adoption of artificial intelligence technologies in Mexico *tout-court*. What we observe are specific sectors rapidly adopting AI technologies for specific purposes and, at the same time, promising applications that will be widespread in the next years. There is indeed a new generation of innovators that are applying already mature AI technologies producing one uncoordinated wave of products and services that are nonetheless contributing to the visibility of AI in Mexico[355]. Instead of grouping AI technologies by the type of architecture behind them, it is more useful to describe the most widespread and salient initiatives using the criteria of the source of the innovation. On one hand, we have seen that the federal and local governments in Mexico are backing up on adopting AI technologies for specific governmental tasks in a way that they are having profound everyday effects on the lives of Mexican citizens (A). On the other hand, private actors, whether they are well-funded (such as the banking system) or are micro innovators, are now the driving force for a myriad of Brotherton services that have an AI-based technology; products and services that are now quotidian to Mexicans (B).

[353] Throughout this document we will assume the AI definition from the AI HLEG is an independent expert group that was set up by the European Commission in June 2018. Accordingly, instead of AI, we use the concept of AI-System or AI-Application interchangeably. See A definition of AI: Main capabilities and scientific disciplines High-Level Expert Group on Artificial Intelligence, European Commission, April 8, 2019.

[354] https://data.oecd.org/mexico.htm#profile-innovationandtechnology (accessed on October 31, 2020).

[355] It is, of course, not our intention to list every AI related Project being developed or deployed on this chapter. There are exciting innovations in the AI Mexican scene. For a broader view on AI initiatives see the "AI Use Cases" in the "TOWARDS AN AI STRATEGY IN MEXICO" below; see "Anexo II" of the Report on AI and Economic Growth https://news.microsoft.com/uploads/prod/sites/41/2018/11/IA-y-Crecimiento-MEXICO.pdf (accessed on October 31, 2020). See also de Report on Service Robots prepared by the Mexican Academy of Computation https://www.researchgate.net/profile/Yasmin_Hernandez2/publication/340634786_Robotica_de_Servicio/links/5e96334a4585150839de623a/Robotica-de-Servicio.pdf (accessed on October 31, 2020). Finally there is a vast registry of documented AI projects in the *Komputer Sapiens* a journal edited by the Mexican Society for Artificial Intelligence at http://smia.mx/komputersapiens/ (accessed on October 31, 2020).

i. Public Interest Driven AI-Applications in Mexico

AI-Applications have been recognized as a useful tool to pursue public good and to attain policy objectives in many sectors of Mexican governments, both on the federal and the local level. Some AI-Applications that we find in Mexico's public sector are comparable with similar efforts seen in other jurisdictions, such as the case of the Internal Revenue Service Agencies, the governmental agencies tasked to prevent money laundry and terrorist financing activities, or the Public Health Services. There is a global trend to use AI-Applications to execute more efficiently tasks which involve processing large collections of data from different sources, which also increment in volume in very small amounts of time and in various formats (Big Data). We also see AI-Applications trying to tackle social problems which may arise from the global COVID-19 pandemic. In this regard, we can applaud the Mexican government's effort in coping with the contention of the pandemic without the aid of any other country. There are other AI-Applications which respond to the specificities of the Mexican country: the vast historical flow of nationals migrating to the US, the nascent Eolic-Electricity Industry, and the recurrent destructive seismic activity in large areas of the Mexican territory.

AI to prevent Tax Evasion

For the past seven years, Mexico's Internal Revenue Service (SAT, by its acronym in Spanish) has been trailing an AI application to detect tax evasion[356]. This system is tries to tackle a form of organized crime. Individuals involved with shell companies use simulated legitimate transactions of services and goods. These shell companies produce millions of taxpayer's receipts to evade the payment of both income and value-added taxes to the Mexican government. In June 2020, the Ministry of Finance made a public statement announcing the filing of criminal charges to 8,212 taxpayers (both individuals and companies)[357]. The fraudulent operations amounted to approximately 55,083,957,551,086 Mexican pesos (an approximate equivalent of 2,223,125,505,255 in Euros). The SAT's AI application analyzed more than 22 million taxpayer receipts and operations. This highly complex application of AI was documented by some of the country's leading scholars in the area years before in a research paper[358]: a White Paper Report financed by the British Embassy would later reveal that the SAT's AI application worked by "identifying pattern disruptions in data analyzed using R Studio, Python Language, and DBs in-memory Redis."[359] In recent years, Mexican tax regulations have moved to an electronic-only regime for tax receipts. Every tax receipt serves as a node to identify multiple data of the operation. With nearly 12 billion tax receipts issued just in 2019, the SAT hopes to increase the internal revenue of the GDP by 3 points[360].

AI to attack Money Laundry

It is only consistent with the global trend that the fight against money laundry has seen the emergence of coordinated efforts across jurisdictions. Money laundry is a particularly sensitive topic in Mexico, given the fight against drug cartels. But perhaps in a more national scale, we

[356] https://www.gob.mx/innovamx/articulos/inteligencia-artificial-131287 (accessed on October 31, 2020).
[357] https://www.gob.mx/shcp/prensa/comunicado-no-054-shcp-mediante-sat-y-pff-anuncia-acciones-en-contra-de-defraudacion-fiscal-a-traves-de-empresas-factureras (accessed on October 31, 2020).
[358] http://omawww.sat.gob.mx/gobmxtransparencia/Paginas/documentos/estudio_opiniones/Evasion_en_IVA_Analisis_de_Redes.pdf (accessed on October 31, 2020).
[359] TOWARDS AN AI STRATEGY IN MEXICO: Harnessing the AI Revolution, White Paper Report, British Embassy in Mexico through the Prosperity Fund, Oxford Insights and C Minds, June 2018, p. 23. Electronic consultation at https://www.oxfordinsights.com/mexico (accessed on October 31, 2020).
[360] https://www.eluniversal.com.mx/cartera/sat-va-por-mas-recaudacion-con-inteligencia-artificial (accessed on October 31, 2020).

nowadays also see the use of AI-Applications in the fight against corruption, (that is, corruption at state level, which is organized from within the government's highest spheres, which aims to dilute public budget by means of complex transactions). Within the Ministry of Finance, there is a unit specialized on Financial Intelligence (UIF, by its acronym in Spanish). The UIF's main purpose is, among other things, implementing the international compromises by the Mexican government as a permanent member of The Financial Action Task Force (on Money Laundering) (FATF), also known by its French name, *Groupe d'action financière* (GAFI)[361]. There is not a more dynamic example of ever-shifting schemes which may be tackled with predictive AI techniques than that of money laundry. When the sophisticated schemes for money laundry are combined by operations devised from officials to simulate legal public procurement operations, the complexity in detecting and prosecuting these organized crime activities would require tremendous amounts of human resources without the aid of AI systems. The variety of lines of function overseen by the UIF now includes the fight against petrol stealing, human trafficking, shell companies' schemes, drug trafficking and political corruption[362]. What is salient in the case of this specialized unit is not only the extensive use on Machine Learning Techniques and Big Data, but that the result of their inquiries is taken as hard evidence by the general attorney's office[363]. That is, Mexico is a case where the result of applying AI into criminal investigations result in the admission as evidence of the conclusions produced by an AI system[364]. So far, this is a notorious precedent in Latin America.

AI to expand Public Health Services

By the end of 2019, the results of a National Survey of Health and Nutrition revealed that approximately 75.2% of Mexican adults were either overweight or obese; while 10.3% of them also suffered from diabetes[365]. This data reveals a serious public health concern affecting not only to adults but children as well. The results of the survey showed that children of women who were obese while pregnant have 1.4 more chances to be overweight at the age of 3 than children of women that had an adequate weight. Women that have obesity before pregnancy will have children with 4.5 more chances to be overweight by the age of 7 than children with mothers who had an adequate weight before pregnancy[366]. This information is key to prevent and to reverse the effects of obesity; it is known that public policy focus on information has a positive effect of the health choices people make through time; and this has been particularly true in the case of Mexico[367]. The cost of distributing relevant, real-time information to the entire population is high– and AI-Applications can help solve the problem of distributing this kind of information, which will help the population make better health decisions. *MiSalud* is an AI

[361] https://www.uif.gob.mx/ (accessed on October 31, 2020).
[362] https://www.gob.mx/cms/uploads/attachment/file/520516/Comunicado_UIF_011.pdf (accessed on October 31, 2020).
[363] In a separate precedent, a judge ruled that the UIF will be admitted as claimant along with the Generals Attorney's Office to prosecute a political corruption case. See the case file number FED/FECC/UNAI-CDMX/0000002/2020 at https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-cdmx/table/?q=FED%2FFECC%2FUNAI-CDMX%2F0000002%2F2020 (accessed on October 31, 2020).
[364] There are comparing points between the Mexican and the Australian case in this regard. Australian Transaction Reports and Analysis Centre (Austrac) as reported by the Association of Certified Financial Crime Specialist at https://www.acfcs.org/austrac-using-machine-learning-to-better-uncover-interconnected-criminal-groups-improve-aml-alerts/ (accessed on October 31, 2020).
[365] https://ensanut.insp.mx/encuestas/ensanut2018/doctos/informes/ensanut_2018_presentacion_resultados.pdf (accessed on October 31, 2020).
[366] Yu Z, Han S, Zhu J, Sun X, Ji C, Guo X. Pre-pregnancy body mass index in relation to infant birth weight and offspring overweight/obesity: a systematic review and meta-analysis. PloS one. 2013;8:e61627. DOI. https://doi.org/10.1371/journal.pone.0061627 and Weng SF, Redsell SA, Swift JA, Yang M, Glazebrook CP. Systematic review and meta-analyses of risk factors for childhood overweight identifiable during infancy. Arch Dis Child. 2012;97:1019-26. DOI. https://doi.org/10.1136/archdischild-2012-302263
[367] See the study on Obesity in Mexico by the National Institute of Public Health available in https://www.insp.mx/resources/images/stories/2019/Docs/190213_LaObesidadenMexico.pdf, p. 28 (accessed on October 31, 2020).

application that delivers interactive messages using Facebook and Twitter messaging systems. The application not only delivers health advice, but also interacts with users to produce reminders of medical appointments and controls, and it delivers real-time information on pandemics or health hazards, and has the ability to produce nudges in favor of better health choices. The pilot program started in 2017, with a focus on maternal health. *MiSalud* sent out SMS messages to 5,000 women, with advice to help improve their health and that of their babies. There was an increase both on the frequency of medical consultations and in the weight of the newborns, and a decrease on the rate of overweight mothers and babies. The program has now extended to diabetes, obesity, child vaccination and risk assessment of drug addiction[368].

There are several opportunities for the use of AI in Mexico and elsewhere. In the face of the global pandemic of COVID-19, different states in society have accelerated the implementation of several technologies in order to contain and cope with the spread of the viral disease.

AI response systems for COVID-19

In the COVID-19 scenario, early detection of clusters of potentially infected people is vital for the containment of the disease. The megalopolis of Mexico City is inhabited by nearly 9 million people. During the early stages of the COVID-19 pandemic, Mexico City's local government implemented two aid systems using AI techniques. The first system is an SMS diagnostic interactive system. It works in the following way: by sending an SMS message with the word "COVID-19", the user receives a series of questions. These questions assess the risks and the chances of the individual being infected with the virus. When the risk of infection is high, the system triggers an alarm which health services use to locate the individual, or to provide them with further information to attend a Health Center for treatment. When contagion is certain, this same system serves to deliver a rapid diagnostic test kit for the use of the cluster of individuals who have interacted with the diagnosed carrier of the virus; as well as to receive additional economic and social welfare relief for the individual and their family. The second system works along with the first one, and it provides a real-time update on the availability of hospital spaces and the crowding of hospitals according to the GPS information of the user. By estimating the closest and least crowded hospital or health facility, this system maximizes the response time for attending individuals who are potentially infected with the virus and minimizes the risk of further contagion. So far, the first system –the rapid diagnostic system– serves around 20 million users; and it was developed by the Digital Agency for Public Innovation of Mexico City (DAPI). The second system is available in the form of an app both for Android and iOS systems.[369] The information harnessed by both systems is processed in a Big Data center to create a daily strategy for the teams of epidemiologists that work on critical clusters across the city, as well as to reassess the behavior of the epidemic throughout the 1495 km² of the megalopolis. Parallel to the DAPI systems, we find COVIDBot, an AI agent for WhatsApp trained to answer questions over interactive multi-media content produced by health organizations. The AI-bot is the philanthropical creation of the Mexican company Intevolution. The main aim of the COVIDBot is to fight against fake news related to the pandemic which circulate on the WhatsApp application. COVIDBot is a free service that feeds in real-time from the registries and data of the World Health Organization, the Center for Disease Control in the

[368] https://www.gob.mx/misalud/ (accessed on October 31, 2020).
[369] https://adip.cdmx.gob.mx/proyectos (accessed on October 31, 2020).

US and the Mexican Ministry of Health. One key feature is that COVIDBot can locate the nearest certified laboratory to get testing for COVID-19[370].

Putting aside the context of the global pandemic, the use of AI in public health services could be applied to a broader policy in Mexico: one which aims to target the needs of underprivileged individuals. Attending disenfranchised groups in a country with a high index of inequality represents a huge challenge for any government in power, and an unusual greenfield for the harnessing of AI-Applications to aid in caring for those who are most in need.

AI for Assisting Mexican Immigrants in the US

Historically, Mexican people have been migrating from Mexico to the US seeking better opportunities, whether it is a seasonal activity related to agriculture or a permanent immigration aimed at a longer stay. By 2018, there were approximately 11.17 million Mexican-born immigrants in the US[371]. Mexican immigrants in the United States face a number of challenges and needs. Attending the needs of the Mexican population living in the US has also been a priority for the Mexican State. A Mexican Consulate can do extraordinarily little for the scattered (and often hidden) groups of nationals that need access to vital information regarding their rights and legal aid. Registering a newborn as Mexican, bringing back a relative defunct in foreign soil, and fulfilling judicial agreements on Alimony are some of the situations that require the aid of the Mexican Foreign Service. The most requested application is a passport reposition, since it is the main mean for identification amongst Mexicans in the US. If an individual is detained, they have the right to have the legal aid of the Mexican Consulate, and to exercise the right to a passport identification is critical. The use of bots to interact with Mexican immigrants is at the moment a promising response to tend the needs of millions of Mexicans in the US. The AI application named "*Asistente Virtual SRE-UNAM*" [SRE-UNAM Virtual Assistant] recognizes and identifies the type of administrative request an immigrant may have. This AI Application generates an automated appointment according to the answers provided during the interaction with the user. The virtual assistant is a joint development between the Ministry of Foreign Affairs and the National Autonomous University of Mexico (UNAM, by its acronym in Spanish)[372].

Public Interest driven innovation may sometimes arise from a particular context which provides an opportunity to tackle a complex problem, and such is the case of human migration. In other cases, the complexity goes in hand with the necessity to cope with the needs to create new ways to promote a better environment. That is the case of AI applied to Eolic Electricity generation in Mexico.

AI to harvest electricity from the wind

By the end of 2019, wind farms were generating 15,000 Megawatts of energy priced at 826 million USD. There are many challenges that come with implementing Eolic Electricity; among

[370] https://www.infobae.com/tecno/2020/04/01/asi-es-el-chatbot-mexicano-que-lucha-contra-la-desinformacion-sobre-coronavirus/
(accessed on October 31, 2020).
[371] Duffin Erin, Number of Mexican Immigrants in the United States 1850-2018 at https://www.statista.com/statistics/673350/mexican-immigrants-in-the-united-states/ (accessed on October 31, 2020).
[372] https://www.youtube.com/watch?reload=9&v=1dv-gVKBRBA&ab_channel=UNAMGlobal (accessed on October 31, 2020). A technical report on the application can be found here: https://cscwcivictechnologies.files.wordpress.com/2020/10/asistente-virtual-saul-esparza.pdf (accessed on October 31, 2020). Behind the team of innovator that developed the application is professor Saip Savage, she has been named one of the Innovators Under 35 Latin America 2018 from MIT Technology Review: https://www.innovatorsunder35.com/the-list/saiph-savage/ (accessed on October 31, 2020).

many, the prediction of the wind force to estimate the energy that will be delivered to the public grid electrical system. The variability challenge is usually tackled with simulation models and predictive simulations that are intensive in computing resources (because of the need for many iterations). The use of AI techniques for forecasting how much energy will be available to inject into the grid from a wind farm has greatly minimized this problem. The AI technique for wind power forecast uses weather information collected during several years using Dynamic Bayesian Networks (DBN). The results of this AI application were satisfactorily compared with forecasting results from previous time series techniques, indicating that DBN is a promising tool for wind power forecasting[373]. The trend we observe here is the use of a machine-learning approach that can be used to replace the rigorous simulation model with a surrogate model (e.g. using vector regression algorithms)[374]. The surrogate model can be obtained in a short period of time, and with far less computational resources[375]. The AI research and applications are consistent with the fast growth of wind farms across the Mexican territory. A key issue to promote competitiveness of Eolic Energy harvesting is to mitigate demand changes and variability of the wind turbines.

The economic growth associated with natural conditions is a key factor for the industry of green energies. Other natural conditions which represent a watchful challenge because of their disruptive power are phenomena which generate destruction and chaos, such as volcano eruptions or earthquakes, which are yet unpredictable. Mexico has a long-standing history of earthquakes. Much of its territory is affected by seismic movements. In the past 365 days, there have been 2,215 earthquakes with a magnitude of 1.5 or greater[376]. AI and IoT combined with machine learning techniques in cloud computing are part of the toolkits Mexicans have to manage earthquake events and casualties.

AI to prevent human harm from Earthquakes

Mexico has had one of the few Earthquake Early-Warning systems (EEW) in the world, called SASMEX, since 1991; but a heavy development phase in this same area began earlier, specifically, after the 1985 earthquake[377]. SASMEX works with sensor stations along the south pacific shore to detect seismic activity. When a seismic event is detected, a long-wave radio alert system is activated. In Mexico City, every public school has a receiver and a sound system to voice out an alert. Between public schools and other sites, there are around 90,000 receivers in Mexico City and the metropolitan area. In average, the alarm systems provide 100 seconds of warning before the seismic wave arrives. 100 seconds, in addition to intensive civic training and rehearsals to conduct safe evacuations from buildings, are the key to prevent deaths should a massive earthquake emerge. SASMEX also has an application for Android and iOS, as well as a Twitter account to widespread notices from the sensor stations. The

[373] Ibargüengoytia-González P H, Borunda-Pacheco M, Reyes-Ballesteros A, García-López Uriel Alejandro, Wind Power Forecasting using Artificial Intelligence Tools, Ingeniería, Investigacıón y Tecnología, 2018, 19-4: 1-11. DOI. http://dx.doi.org/10.22201/fi.25940732e.2018.19n4.033

[374] Santamaría-Bonfil G, Reyes-Ballesteros A, Gershenson C, Wind speed forecasting for wind farms: A method based on support vector regression, Renewable Energy, 85, 2016: 790-809, ISSN 0960-1481, DOI. https://doi.org/10.1016/j.renene.2015.07.004.

[375] Rosado-Tamariz E, Zuniga-Garcia M A, Santamaria-Bonfil G, Batres R, A machine-learning approach to speed-up simulation towards the design of optimum operating profiles of power plants. In Proceedings of the 8th International Conference on Informatics, Environment, Energy and Applications (IEEA '19). Association for Computing Machinery, New York, NY, USA: 194–199. DOI. https://doi.org/10.1145/3323716.3323735. As a NB, the DeepMind project by Alphabet has also been doing research to tackle stability and variances issues on wind farms, so Mexico is in well company to race towards a solution for a global energetic problem. See https://deepmind.com/blog/article/machine-learning-can-boost-value-wind-energy (accessed on October 31, 2020).

[376] https://earthquaketrack.com/p/mexico/recent?mag_filter=7 (accessed on October 31, 2020).

[377] http://www.cires.org.mx/sasmex_es.php (accessed on October 31, 2020).

demand for accurate and widespread early alert systems has produced AI-Applications in the Mexican market. There are two main companies in this field: Skyalert[378] and Grillo[379]. They both use IoT and cloud computing to generate seismic alerts in applications on mobile devices; and, in the case of Skyalert, to voice alarm systems. Skyalert is more oriented towards the corporate market by offering training and awareness as well as on site alarm systems. This company also provides a 120 second alert before the destructive wave arrives, with a personalized GPS focus alarm system that filters seismic events for the user. Grillo has a cero false positive record, and has consistently won benchmarking results against both Skyalert and SASMEX. Both companies use cloud computing to process data from their proprietary sensors. Grillo works with AWS cloud services, which have delivery times of 100ms from the sensor to the cloud; while Skyalert processes its data with Microsoft's Azure platform. Grillo has a line of sensors towards structural damage detection on buildings to prevent further human losses after an earthquake impacts the structure. All in all, between Grillo and Skyalert, these companies have 97% of market share for early seismic alert services. A very distinctive initiative by Grillo is a partnership with IBM and the Linux Foundation. Grillo has opened its data and code under the OpenEEW project of boosted global collaboration around Grillo's technology, in hopes of fomenting similar initiatives around the globe[380].

The pulsion behind public sector AI-Applications in Mexico is to attain the common good, the reach and protection of public interest for the benefit of the citizens at large. Many AI projects, both at the research and at the application level, strive to tackle a social problem even if the projects are a private initiative. Sometimes the line between the private and the public interest is not clear enough to classify a specific AI research or application. There are clear cut cases, however, where the AI application emerges as a response to a business opportunity, a market niche or from the purposeful innovation of corporations.

ii. Private Interest Driven AI-Applications in Mexico

In the absence of a general open registry of AI-Applications in Mexico, we can only account for some of the most noticeable private enterprises based on AI-Applications. Almost in every main industry, there are forms of AI-Applications already in use, from the automobile industry that employs robots and mechatronic tools to assemble cars; to the army of bots that political parties have used in the past to disinform or to harness public division. At the same time, the close proximity of the Mexican economy to the US market has boosted the supply of services based on the US (or with corporate matrix in the US) in Mexican territory. In this regard, Mexicans interact with the now global supply of services around digital commodities: Netflix, Uber, DiDi, Facebook, Twitter, Instagram, Amazon, and Google, for an instance. In this broad context, there are some areas of local innovation that started from now consolidated startup companies or from larger corporations who invested in using AI as part of their Research and Development programs to keep their competitive edge. On the side of startups, we have a broad range of innovators that now provide AI-Applications to the Mexican market of digital services. The range of innovators compromise FinTech, CRM-ChatBots, Social Media Analytics, Real Estate and Medical Robots as a service. Since the examples we are reviewing are only a few participants in the broader US-Mexico's market, it will be inevitable to see that

[378] https://skyalert.mx/ (accessed on October 31, 2020).
[379] https://blog.grillo.io/the-grillo-journey-e60b2dcee224 (accessed on October 31, 2020).
[380] https://grillo.io/linux-foundation-hosting-openeew/ and https://grillo.io/open-source/ (accessed on October 31, 2020).

in most cases the initiatives are the local response (or version) of other existing solutions for English speaking consumers and services in the US.

AI-Applications in FinTech

According to the Mexican Financial Regulatory Agency (CNBV, by its acronym in Spanish), there were 500 FinTech companies operating in Mexico by 2019[381]; but another independent firm, however, accounts for 640[382]. According to Statista data, Mexican FinTech companies will carry operations by 2022 which will be valued in approximately 69,000 million USD, with an annual growth rate of 17.3%[383]. With more than 158 startups, Mexico has the largest FinTech market in Latin America, even larger than those of Brazil and Colombia[384]. At the same time, only 4.5% of FinTech companies in Mexico ceased to operate from June 2019 to June 2020[385]. The main operations in FinTech are intensive in Machine Learning for risk assessment, the use of Big Data and AI. Almost the entire business cycle for FinTech companies revolve around the use of AI related technologies: from payments and remittances, personal financial management, to crowdfunding and loans. There is a vast universe of FinTech and related "techs" such as InsurTech as well in the Mexican scenario. The emphasis on the AI application each FinTech company uses is closely related to what domain the use of AI will bring the most competitiveness.

For Conekta[386], a FinTech that offers its service as a certified aggregator of payments (be it either cash, point-to-point, or e-commerce), security is everything. This company developed Conekta Shield, which uses machine learning and Bayesian networks to model consumer and users' behaviors to create dynamic patterns to minimize fraud and transactional risks for Conekta's clients[387]. Like Conekta, Dapp has a similar business philosophy, which is to automate payments across different ecosystems. Unlike Conekta, Dapp provides banks with the technology to generate QR. The users will scan the QR from their mobile phone and pay with the Wallet App of their choosing. Once the Wallet transfers the funds for the transaction, the commerce receives the payment[388].

Other FinTechs focus on using Big Data and Machine Learning to profile the client and "nudge" them throughout the re-payment process.

This is the case of Kubo Financiero[389], the first online peer to peer lending community in Mexico, and the first in Latin America authorized by a financial authority. Kubo Financiero provides online loans from a mobile phone, ranging from 400 to 4,100 USD; but also offers an attractive interest rate to micro-investors, which starts at only 22 US Dollars. The process of profiling includes an application form, uploading a few documents, and a video conference call. Credit applicants with a good credit history obtain better interest rate and terms, while investors

[381] https://www.eleconomista.com.mx/sectorfinanciero/Tenemos-mapeadas-mas-de-500-fintech-en-Mexico-CNBV-20190526-0072.html (accessed on October 31, 2020).
[382] https://elceo.com/tecnologia/cnbv-alista-autorizaciones-para-80-empresas-fintech-entre-octubre-y-noviembre/ (accessed on October 31, 2020).
[383] https://www.statista.com/study/45600/statista-report-fintech/ (accessed on October 31, 2020).
[384] https://panamericanworld.com/en/magazine/startups/meet-the-five-most-innovative-mexican-fintechs/ (accessed on October 31, 2020).
[385] https://www.finnovista.com/radar/el-numero-de-startups-fintech-en-mexico-crecio-un-14-en-un-ano-hasta-las-441/ (accessed on October 31, 2020).
[386] https://conekta.com (accessed on October 31, 2020).
[387] http://www.ebizlatam.com/seguridad-e-inclusion-soluciones-conekta/ (accessed on October 31, 2020).
[388] https://dapp.mx (accessed on October 31, 2020).
[389] https://www.kubofinanciero.com/Kubo/Portal/productos/productos.xhtml

who lend to borrowers get better return on their investments[390]. Kubo employs supervise models from data science to enrich their data collection processes from clients; as well as Machine Learning to calibrate their Credit Scoring mechanism. On a similar market segment, Kueski[391] employs AI-Applications to determine a risk assessment profile for lending money using information from non-traditional data points, including social media presence[392]. Loan applications are accepted and paid in minutes. Borrowers can progressively ask for higher amounts. Kueski's specialty are micro-loans, which range from 100 to 200 US Dollars for up to 30 days. Risk assessment using AI can be applied to enhance a prediction of the success of a small business. In this field, AI may be used to predict the likelihood of whether an idea or an entrepreneurship will succeed in becoming a high-impact company in the Mexican economy. Konfío[393] uses proprietary algorithms and data analysis to expand affordable credit operations and to accelerate the financing process of Small and Medium Enterprises (SME). SMEs in Mexico represent 95% of the market. The success of Konfío is based on the intensive use of data points across the balance sheet of their potential clients to predict their chances of success and then adapting an optimal lending rate to produce that success in the SME. Using algorithms in improvement cycles, Konfío has been able to lower the severity rate to loan value 10 times from 2016 to 2017[394].

While risk profiling is especially sensitive for money-lending platforms, it is the crux for companies using digital currencies such as Bitcoin[395]. In Mexico, Bitso is a company that believes that the key to global financial inclusion will be achieved through the use of Bitcoin and other Crypto Currencies (the company currently use 10 different types of Crypto Currencies). They have heavily invested on AI technologies for automated client due diligence to prevent money laundry and terrorism finance. At the same time, they have simplified the front-end of their application to invest, use, and transact with Crypto Currencies that rely on BlockChain technology.

There is finally a market segment that also relies in intensive Machine Learning and Big Data applications to interact with clients to offer services and assure competitive transactions. The FinTech companies referred to as "Neo-Banks" are by themselves a thriving example of the use of AI related applications. It would be out of the scope of this survey to detail every participant on the Mexican market[396]. It will suffice to acknowledge that all Neo-Banks employ some form of AI related technology: be it autonomous to interact with clients, data-mining for assessing risk, cybercrime prevention, or personalized-micro targeted services from structure models of Big Data.

AI-Applications for Real Estate

Assessing risks using non-traditional points to create a profile has been a fairly common AI technique employ in the FinTech sector. A less common use of the predictive risk modelling techniques is in the domain of leasing households. Mexico's market for renting urban property for household domestic use is a dominant market. Some surveys reveal that approximately

[390] https://publications.iadb.org/publications/english/document/FINTECH--Innovations-You-May-Not-Know-were-from-Latin-America-and-the-Caribbean.pdf (accessed on October 31, 2020).
[391] https://kueski.com (accessed on October 31, 2020).
[392] https://kueski.com/blog/tecnologia/machine-learning-aprendizaje-supervisado/ (accessed on October 31, 2020).
[393] https://konfio.mx/ (accessed on October 31, 2020).
[394] https://expansion.mx/tecnologia/2017/08/15/konfio-un-ejemplo-de-como-las-fintech-dejan-atras-a-los-bancos (accessed on October 31, 2020).
[395] https://bitso.com (accessed on October 31, 2020).
[396] There is a fairly comprehensive list of Neo-Banks in the Mexican market here: https://www.legalparadox.com/post/neo-banks-who-will-control-the-mexican-market (accessed on October 31, 2020).

42% of the Mexican population lives on a rental house[397]. Homie[398] uses a predictive model with hundreds of variables to estimate the likelihood for a tenant to default on the rent payments (tenant delinquency rate). The Machine Learning model does not only analyze the applicant's income and credit history, but also the degree of compliance to do financial commitments. Homie selects the best applicant for owners. After selecting the best applicant, Homie takes care of rent collection and even has an insurance policy where they will pay the owner in case of default, and they will recover the property in four months[399]. The customer response service of Homie is backed up by sharp predictive algorithms, which are one of the selling points which make owners entrust their assets to Homie. In this regard, the first respondent between clients and services is key to maintain a robust and engaged client base.

AI-Applications for virtual assistants

AI-Applications in the form of ChatBots with different levels of complexity, are now a global trend to tend clients' questions, comments, and claims; as well as to attract new customers. According to an Oracle report on emerging technologies, leading companies will invest on chatbot-based interactions, including Intelligent Voice Investments. 51% of respondents stated that the benefits of investing on intelligence voice is "Faster time to customer issue resolution", while 50% responded that its main benefit is "Increased operational efficiencies"[400]. Juniper Research forecast that by 2022, companies will save 8 Billion US Dollars from the use of ChatBots for customer services-related activities[401]. Similarly to Oracle's report, Gartner estimates that by 2022, 70% of white-collar workers will interact with conversational platforms on a daily basis[402].

In Mexico, there are more than a few AI-based companies that offer ChatBot services with varying degrees of autonomy and range of interaction. According to Nanalyze, YaloChat[403] has joint seven other global companies that make easy to adopt the ChatBot technology– that is, the range of companies that can "give us a ChatBot quickly and with minimal fuss"[404]. YaloChat works with companies to understand their client's relationships, and then, to design a tailor-made Bot to improve service quality. YaloChat sends and manages client notifications, including product demonstrations. The AI behind YaloChat automatizes frequently asked questions with an estimate of prediction of 90% of potential client-service conversations. Finally, YaloChat further details the clients' profiles to deliver distinctive interactive experience to increase service-product engagement. YaloChat has produce AI oriented Bots for many big companies operating in Mexico, including Amazon, Pepsi, Volkswagen, Aeromexico and Walmart. Following YaloChat's footsteps is Gus Chat[405], another Mexican company committed to the design and development of AI-Bots for the automatization of client service, with a

[397] https://www.eleconomista.com.mx/finanzaspersonales/En-el-2019-70-de-las-nuevas-familias-compraria-una-casa-20181216-0053.html (accessed on October 31, 2020).
[398] https://homie.mx/h/ (accessed on October 31, 2020).
[399] https://inmobiliare.com/homie-la-plataforma-que-facilita-la-renta-de-departamento (accessed on October 31, 2020).
[400] https://www.oracle.com/a/ocom/docs/dc/em/lpd100807811-impact-of-emerging-technology-on-cx-excellence.pdf?elqTrackId=d368a11a304041c8a7bf8e7a2f2a71e2&elqaid=82669&elqat=2 (accessed on October 31, 2020).
[401] https://www.juniperresearch.com/new-trending/analystxpress/july-2017/chatbot-conversations-to-deliver-8bn-cost-saving (accessed on October 31, 2020).
[402] https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers/ (accessed on October 31, 2020).
[403] https://www.yalochat.com (accessed on October 31, 2020).
[404] https://www.nanalyze.com/2017/07/7-chatbot-platforms-chatbots-easy/ (accessed on October 31, 2020).
[405] N.B. Perhaps the name of the company is a gest to the GUS system that pioneer the architecture for dialogue systems. See Page 11, Chapter 24 of Speech and Language Processing. Daniel Jurafsky & James H. Martin at https://web.stanford.edu/~jurafsky/slp3/24.pdf (accessed on October 31, 2020).

specialty in E-commerce, FinTech and Insurance[406]. Gus Chat designs a different algorithm for each different task assigned to a Bot: generation of sales leads, transactional Bots, client-response service Bots, B2B Bots and marketing and advertising Bots. This wide variety of algorithms makes Gus Chat an innovative company on the field of Natural Language processing in the region.

Using Natural Language Algorithms to support clients and services in the form of interactive Bots is one of the many uses of that AI technology. Character recognition and voice to text recognition are also two key technologies which are enhancing the business process in Mexico as well.

Nowports[407] is platform that connects clients, providers, customs services, ports and carriers. The process employs blockchain to make every transaction safe and transparent to every participant involved. Nowports processes every shipment request automatically via e-mail, regardless of the format or documentation. The AI system processes the request, and then places a budget with a fixed price without variation. The AI agent connects all the other end points to complete the transaction; and once approved by the client, it places the order to ship the product. The client can trace in real time the in-route of the merchandize up to the delivery point. With their AI technology, Nowports can process logistic requests up to 70% faster than other companies in the competing market share.

When using AI-Bots, the combination of cognitive computing, cloud computer and a general training AI architecture make a powerful combination. Nearshore Delivery Solutions (NDS)[408] is a Mexican company that works with IBM to train and create cognitive computing Bots. NDS could create a virtual assistant for a large Mexican bank in only 12 weeks. NDS trained IBM Watson Assistant with the analysis of e-mail, phone call transcripts and chat messages exchanged over a year between clients and the bank support area. With the knowledge obtained from this in-depth, extensive analysis, the IBM Watson Assistant was able to answer nearly 1,000 questions. The Bot also detected when a question needed a human intervention to reroute the call to a human expert. In addition, NDS used Watson Tone Analyzer to detect anxiety or anger in a client for a rapid human intervention as to prevent further client frustration. 75% of all customers calls are now handled by the cognitive Bot which resolves the issues in two minutes in average; 80% faster than a typical call center assistant[409].

The use of a general AI architecture like Watson provides a robust framework to innovate with cognitive computing. In particular, tone analysis and emotional responses are crucial information to produce accurate responses to customers and clients.

There are two Mexican companies that develop and apply AI technologies to manage messages. Metric is a Mexican company that started creating automated content for social media. They then moved to program Bots to detect emotions and attitudes, and thus, to respond with specific content to influence these emotions. As of now, their business is to model and predict social media behavior. Their products generate estimative forecasts of campaign impact, stakeholder positions and trends before they happen[410]. They can assess risk (and opportunity) scenarios for the value of trademarks including the detection of Bots, trolls and fake news. In a similar manner, but oriented towards advertisers and ad agencies, the Mexican

[406] https://gus.chat (accessed on October 31, 2020).
[407] https://nowports.com (accessed on October 31, 2020).
[408] https://nearshoremx.com/cognitive_computing (accessed on October 31, 2020).
[409] https://www.ibm.com/downloads/cas/JOEXYB08 (accessed on October 31, 2020).
[410] https://buy.metricser.com/revelio/ (accessed on October 31, 2020).

Adext[411] harnesses the power of AI for social media campaigns. Adext has created machine learning models that run simulations to determine the best timing, target audience and digital space (placement) for social media ads. The simulation determines the optimal marketing impact given the budgetary constraints from thousands of potential scenarios to produce incredible granularity for audiences. Adext claims to produce at least 25% boost in campaigns right from the beginning. This company also offers a demonstration to see in real-time the number of conversions that a campaign has with the use of their models[412].

Bots (like their physical supported counterpart– robots) operate using models and information to complete specific tasks. The notion of a digital space makes Bots and their presence less evident than that of humanoid-robots. However, the field of AI in Mexico has seen advances in the designing and developing of Robots as Service (most of them, manufactured elsewhere)[413].

AI-Applications for Health services

Detecting and providing rapid isolation to infectious clusters is one of the many parts of dealing with the COVID-19 pandemic. Another challenge is caring for the patients infected with the virus during hospitalization. There is now a massive amount of information available which describes the risks and contagion of healthcare professionals that tend infected patients. There is Mexican company, GESEDIG, which offers three models of Robots as Service that can aid the health care staff in caring for Covid-19 patients[414]. The robot humanoids can interview patients when they arrive to the health center. They can do rounds in hospital entry areas and waiting rooms showing information videos and providing with health care preventive measures. The robots can also guide patients to different sick bay areas using motion sensors, as well as delivering food and administering medicine. The robot humanoids are providing a mean to prevent further spread of the virus among health care professionals. The robot's AI has been designed and programmed by Mexican engineers. The AI core includes a Natural Language processing module, face recognition, bar-code readers, Customer Relationship Management processes and Click to Call capabilities[415].

We have seen the many uses of AI-Applications in Mexico across different domains. From the public interest side to the private interest sector, AI-Applications can be seen impacting industries, services and governmental tasks. The life cycle of an AI-Application involves financial, technological and human resources. In an emerging economy such as the Mexican economy, the variables that will impact negatively the flourishment of AI Research and Applications are vast, but not infinite. In a similar manner, there is an unknown degree of interaction between individual mindset, collective awareness, institutional infrastructure, policy making and legal framework to spring and incentivize the development of AI. An account of

[411] https://www.adext.ai (accessed on October 31, 2020).
[412] https://www.latamdigitalmarketing.com/blog/software-campanas-digitales/ (accessed on October 31, 2020).
[413] There are many cases of already deploy robots as service, see for example https://integritas.mx/robots-de-servicio/ , http://robotsmexico.mx as well as the research paper of Sucar for an analysis of the maturity of the field in Mexico in https://ccc.inaoep.mx/~emorales/Papers/2009/eduardo.pdf (all sites accessed on October 31, 2020).
[414] N.B. There has been chirurgical procedures using robots or robotic instruments in Mexico since 1993 according to Miller FHS, Cirugía robótica en México, Los sistemas inteligentes, perspectivas actuales y a futuro en el ámbito mundial, Revista Mexicana de Cirugía Endoscópica, 2003, 4(1): 45-50. With the advances of augmented reality attached to robots in the operating room there certainly is a new field of applications for AI in Medicine. In Mexico however is not widely extended. The first AI-assisted augmented reality surgery took place in La Conchita Hospital in the State of Nuevo Leon using the *BedsideXR* platform on February 2020, see https://www.christusmuguerza.com.mx/sala-de-prensa/es-hospital-conchita-sede-de-la-primera-cirugia-con-realidad-aumentada-en-mexico/ (accessed on October 31, 2020). We have left aside this field of AI-applications in Mexico because we see them as in their early stages compared to Robots as Service for Health Care.
[415] https://gesedig.com (accessed on October 31, 2020).

public policy, regulatory framework and private coordination initiatives can tell a partial account for the AI momentum in Mexico.

III. Accounting for AI (de)regulation in Mexico

In Mexico, there is not a legal, legislative framework which regulates AI whatsoever. The absence of a legislative body for AI is an absence both at the Federal and at the Local State level (i.e. the local legislative body). There is, however, a legislation and a set of rules of different hierarchies and sources in specific domains where AI-Applications appear. It is useful to acknowledge from a focused perspective the indirect regulation of AI systems in Mexico. The lack of a legislative instrument for regulating AI systems is instead filled with public policies which have been implemented at least since 2019, though there is public policy implemented in digital transformation and digital inclusion that goes back at least from one decade. Mexico is a federal republic, which means that there are two policymakers when comes to AI. We find that the federal government has seen efforts to foster innovation and development of AI technologies. At the same time, some local governments have also instated different governmental actions in favor of the adoption of AI technologies. In addition to the two tiers of regulation and of policymaking, federal and local, we find that there are state agencies with constitutional powers that also formulate regulations pertaining AI systems[416]. In Mexico there are public and private universities, as well as public and private research centers, and all of them, in one way or the other, have an interest in the public policy push for by federal and local governments. We have to take a particular look at a federal government agency in charge of technology and development to appreciate the specific mechanisms of how the AI public policy has been carried out. If we look at the cases that we have presented in part one of these chapter, we can see that most initiatives date back from before a formal federal public policy was in force in Mexico. At the same time, the private interest driven AI-Applications that we have analyze are not the direct result of any specific policy for the developing AI initiatives in Mexico. Most of AI-Applications in the private sector are the result of a combination of market forces and a special individual grit for innovation. Perhaps the absence of a strong legal framework and some indirect favorable regulatory environment produced the AI momentum in Mexico that we have seen for the past years. At the same time, organized collectives with members ranging from diverse fields (such as industry, academia, entrepreneurship, and activism) are proposing a national agenda for AI development in Mexico. The democratic, spontaneous exercise of the collective in favor of AI has no precedent in the country. We will present and analyze the legal framework that indirectly relates to AI-Applications and the public policy both at the federal and local level that has been proposed and implemented in Mexico (A); before moving to analyze the collective movement in favor of AI that is pushing forward an agenda of themes for the Mexican state to regulate (B).

i. Legal framework and public policy around and about AI

To present the regulatory framework that indirectly touches AI-Applications we will distinguish between national regulations and local regulations. In Mexico certain legal areas can only and exclusively be regulated by the Federal Congress, hence, excluding the power of local legislators. A legal area that is the exclusive power of the federal Congress is called a national

[416] An example is the Mexican Federal Institute for Telecommunications that has an express agenda for the use of AI-Applications on the Mexican telecommunications sector. See http://www.ift.org.mx/sites/default/files/contenidogeneral/transparencia/1vision19-23.pdf (accessed on October 31, 2020).

regulation. There are other legal areas where both local and federal legislators can produce norms applicable at the same time. Along the same lines, we will distinguish between federal government policy and local government policy pertaining to AI.

Personal data protection

Personal data protection rights are handled on the constitutional level. Mexico's Constitution establishes a national agency, with full autonomy, to oversee the rights relating to data protection[417]. But, at the same time, it replicates the design of an autonomous overseer of data protection rights in every local State of the Mexican Federation. The national agency to oversee the protection of rights relating to personal data (called INAI by its acronym in Spanish[418]) has enough power to audit, impose sanctions and even to recall operations if there is a breach of the principles and standards regarding personal data[419]. Interestingly enough, Mexico is a signing party to the 108 Convention as well as the Additional Protocol to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, regarding supervisory authorities and transborder data flows. To complement the national and international legal instruments INAI is part of Ibero-American Network for Data Protection[420] (The NDP). The NDP has issued Specific Guidelines for Compliance with the Principles and Rights that Govern the Protection of Personal Data in Artificial Intelligence Projects[421] as well as General Recommendations for the Processing of Personal Data in Artificial Intelligence[422]. Both instruments are considered soft-law in terms of their institutional force under the Mexican legal system, and yet, they provide a very detailed guiding framework of principles and standards for the INAI.

Anyone can bring about a claim before the INAI against any governmental agency, or private entity or individual who might have breached the constitutional and legal duties towards personal data protection. There has not been any single case brought before the national authority on personal data rights that involves data collecting, processing, or being used in an AI application in Mexico.

Financial Regulation

Unlike data protection, financial regulation is exclusive to federal regulation, nationwide, in Mexico. Should any FinTech company or startups in Mexico take money from the general public at large, or in case they provide any sort of banking services, these companies and startups are obliged to comply with their Federal Statute on FinTech[423] along with all the detail regulations issued by the Federal Authority in Financial Regulation[424] and those applicable from the Mexican Central Bank[425]. From the viewpoint of a normative system, AI-Applications are not the main theme of the normative system regulating FinTech. There is not one single

[417] Article 4 of the General Statute for Transparency and Access to Public Information at http://www.diputados.gob.mx/LeyesBiblio/pdf/LGTAIP_130820.pdf (accessed on October 31, 2020).
[418] https://home.inai.org.mx (accessed on October 31, 2020).
[419] Article 6 of the Mexican Constitution at http://www.diputados.gob.mx/LeyesBiblio/pdf_mov/Constitucion_Politica.pdf (accessed on October 31, 2020).
[420] https://www.redipd.org/es/la-red/historia-de-la-red-iberoamericana-de-proteccion-de-datos-ripd (accessed on October 31, 2020).
[421] http://inicio.inai.org.mx/nuevo/SPECIFICGUIDELINESARTIFICIALINTELLIGENCEPROJECTS2019.pdf (accessed on October 31, 2020).
[422] http://inicio.inai.org.mx/nuevo/GeneralRecommendationsfortheProcessingofPersonalDatainArti%EF%AC%81cialIntelligence.pdf (accessed on October 31, 2020).
[423] http://www.diputados.gob.mx/LeyesBiblio/pdf/LRITF_090318.pdf (accessed on October 31, 2020).
[424] https://www.cnbv.gob.mx/SECTORES-SUPERVISADOS/Fintech/Paginas/NORMATIVIDAD-FINTECH.aspx (accessed on October 31, 2020).
[425] https://www.banxico.org.mx/marco-normativo/normativa-agrupada-por-sujeto.html (accessed on October 31, 2020).

reference which indicates any standard for developing or using AI systems when procuring financial services through technology. Evidence-based financial risk assessment, different profiling methods, and information that any entity within the financial services market may possess is obliged to provide to the regulator are all the same on the regulator's rationale. There is a rationale to seek assurance when taking money and an adequate risk management when lending money, regardless of the technical and technological means for deciding the client base. It is widely recognized among the FinTech community in Mexico that the federal regulation fosters business and innovation. In this regard, the FinTech statute indirectly embraces a light approach to technological innovation in the financial services market. With a legislative policy that is not too intrusive of innovation, it is expected that AI systems and applications in the FinTech market will also flourish.

While the financial services regulation is not directed specifically towards AI-Applications there has been active public policy both by their federal and local governments to promote and support AI-Applications in Mexico.

The 21st of March 2018, there was the presentation of a study commissioned by the UK Embassy in Mexico, with support from the Office of the Mexican President, to Oxford Insights. The study was a draft of a national AI strategy for the Government of Mexico[426]. At the end of the presentation the Mexican government revealed the 2018 Strategy for Artificial Intelligence in Mexico (MX-AI2018 Strategy). The strategy consisted in five public policy actions[427]: 1) To develop an adequate governing framework to spike a multi-stakeholder open dialogue. The government would create a commission of ministries for the development of an Electronic Government. 2) To identify the needs in the industry and the best practices in government related to AI. 3) To champion an international effort on AI with specific emphasis on the Mexican role before the OECD and the G-7. 4) To open the public consultation regarding the recommendations from Oxford Insights' study. 5) To work with experts and citizens in in the sub-committee for Artificial Intelligence and Deep Learning created at the Ministry of Civil Service[428]. At the time Mexico, became one of the first ten countries worldwide to initiate concrete public policy actions fostered towards the development, adoption and use of AI.

On 2019, Mexico embraced the OECD Principles on Artificial Intelligence "to promote AI that is innovative and trustworthy, and that respects human rights and democratic values."[429] by embracing the OECD Principles on AI Mexico was compromised to follow the five complementary values-based principles for the responsible stewardship of trustworthy AI as well as the five recommendations to governments by the OECD. By that time, the federal government did create a commission of ministries for the development of an electronic government to coordinate different efforts in the deployment of the MX-AI2018 Strategy.

One salient result from the national public policy was an initiative organized by the Nacional Council for Science and Technology (CONACYT by its acronym in Spanish). CONACYT is in charge of overseeing and funding public research through National Public Research Centers (NPRC) across the country. Each center has an independent and autonomous research agenda in part because they are created in public universities and then accredited by CONACYT as NPRCs. Under CONACYT stewardship, eight NPRCs joined forces to conform

[426] The study was eventually published in June 2018. See https://www.oxfordinsights.com/mexico (accessed on October 31, 2020).
[427] https://datos.gob.mx/blog/estrategia-de-inteligencia-artificial-mx-2018 (accessed on October 31, 2020).
[428] https://www.gob.mx/cidge/articulos/crea-sfp-subcomision-de-inteligencia-artificial-y-deep-learning-de-la-cidge-161421?tab= (accessed on October 31, 2020).
[429] http://www.oecd.org/going-digital/ai/principles/ (accessed on October 31, 2020).

the Artificial Intelligence Initiative. This initiative allows for joint research in interdisciplinary fields between researchers graduate students from different backgrounds and disciplines. The initiative focused its efforts on modeling natural and social phenomena into domains of medicine, public security, human mobility and transportation, natural disasters prevention and training of human resources in AI[430]. The initiative has since produced a vast number of researches, exhibited in different publications on papers in prestigious journals across multiple disciplines. The initiative also has produced research groups between scholars and graduate students from different NPRC; and as of now 7 in total but with a permanent revision of themes to keep up with frontier research[431]. The alliance has proven the richest human resources for multidisciplinary work on AI-Applications in Mexico.

In parallel to public policy from the federal government we see that local state authorities have also looked at AI-Applications proactively. As a broader context, one should always keep in mind the vast territory that Mexico has, as well as the diversity between different local states. In that regard, some states have more development in terms of human resources, industry, and income per capita than others. It is not strange, then, that some of the initiatives on public policy either by the local government are in conjunction with the private sector there come from specific regions in the country.

The state of Puebla[432] and Queretaro[433] have started local programs to transform their capitals into "Smart Cities" with the use of AI-Applications. The governments of Yucatan[434] and Guanajuato[435] have also launched aggressive public policy plans to bring on board technological industry, they have created alliances with relevant technological companies or research centers, and they have supported their local public universities to play a major role in their public policy. In Yucatan's capital city, Merida, there is an ongoing project to create an AI Research Center. Veracruz, a southern state, has a research center on AI operating since 1994[436]. Other local states with research centers on AI are Chihuahua[437], Nuevo León[438] and Jalisco.

Jalisco is perhaps the most salient among the aforementioned states Mexico. Jalisco has been named the Mexican Silicon Valley[439] because of the number of startups, technological big players that have settled their research and development labs in the state, and for the high density of technological skill professionals[440]. Last year Jalisco's local government entered a collaboration agreement with Tec de Monterrey (one of the most prestigious private technological universities in Mexico) and the Inter-American Development Bank to create an Artificial Intelligence Hub the first in Mexico. The Hub involves a 900,000 USD investment from Tec de Monterrey, the participation of Intel, IBM y Amdocs, Sparkcognition, Tata Consultancy Services and Wizeline, the University of Berkeley California, Beijing Institute of Technology

[430] https://www.consorcioia.mx/nosotros (accessed on October 31, 2020).
[431] https://www.consorcioia.mx/grupos-investigacion (accessed on October 31, 2020).
[432] https://www.eluniversal.com.mx/estados/proyectan-puebla-como-ciudad-de-inteligencia-artificial (accessed on October 31, 2020).
[433] https://amqueretaro.com/queretaro/2020/10/22/convertiran-a-queretaro-en-ciudad-inteligente/ (accessed on October 31, 2020).
[434] https://yucatanahora.mx/yucatan-busca-desarrollar-proyectos-de-inteligencia-artificial/ (accessed on October 31, 2020).
[435] https://www.eleconomista.com.mx/estados/Guanajuato-epicentro-de-la-cuarta-Revolucion-Industrial-20201029-0104.html (accessed on October 31, 2020).
[436] https://www.uv.mx/ciia/ (accessed on October 31, 2020).
[437] https://www.facebook.com/InteligenciaArtificial.Center/ (accessed on October 31, 2020).
[438] https://www.eleconomista.com.mx/estados/Centro-de-Inteligencia-Artificial-en-NL-alista-inicio-de-operaciones-20180820-0020.html (accessed on October 31, 2020).
[439] https://english.elpais.com/elpais/2017/03/13/inenglish/1489403756_441981.html (accessed on October 31, 2020).
[440] https://inmobiliare.com/por-que-guadalajara-es-el-silicon-valley-mexicano/ (accessed on October 31, 2020).

and the Institute for Research on Informatics and Automatization from France. One of the novelties of these consortium is that there will be focus on tackling socially oriented complex problems of public policy in Jalisco on Health, Education and Public Security. The Hub will work on producing AI-Applications for the treatment and detection of the diabetic retinopathy, to prevent school dropout, and to avoid juvenile delinquency[441].

The private sector has been quite active in terms of generating AI-Applications and fostering initiatives across the country. Perhaps the most salient exercise in favor of configuring a comprehensive roadmap for a national agenda on artificial intelligence is the collective IA2030.

ii. A collective framework on and about regulating AI

IA2030 is a collective of organizations and key actors of organizations from civil society, academia, the private sector, industry at large, independent consultants, government officials and public research centers[442]. The collective has been steered by a startup called C Minds. There are two products as a result of a very intensive coordination effort: a national survey[443] and a national agenda for artificial intelligence in Mexico[444].

The survey was developed by nearly 50 independent organizations and professionals on a *pro bono* engagement. The survey was available online for one month in the AI field from August 15 to September 18, 2018. Although the survey results cannot be generalized (that is– it lacks any statistical validity), it serves well as a social thermometer around interested parties in the AI field. Out of the 1,585 respondents, nearly 90% of them where from Mexico City, the State of Mexico and Jalisco. All three locations concentrate large amounts of technological efforts and human resources related to AI in Mexico. 43% of respondents worked on a governmental position, 31% on the private sector and 18% on academia. Among the main findings the survey shows that 80% of respondents believe that AI will have a positive effect in their lives. However, 53% believes that unemployment will raise with AI and 45% are concerned with their privacy and personal data being compromised by AI-Applications. Almost 45% said there were ethical and inequality issues from the extensive use of AI without an adequate regulation. On the public policy front, there were a confluence of opinions on the active role of the Mexican State to incentivize research, development and adoption of AI-Technologies, to improve public services and generate more human resources specialized towards AI-Applications.

The National Agenda on AI is the result of a collaborative effort of 400 people from different sectors and backgrounds divided in six working groups according to each participant expertise: Data, Digital Infrastructure and Cybersecurity; Ethics, Governance, Government and Public Services; Research and Development; Education, Capabilities and Skills; and Outreaching Mexicans Immigrants. The resulting exercise is the presentation of specific problems within each topic of the working group, the proposal of lines of action and key point indicators to assess the lines of action proposed. While the specific content would need a more detailed analysis than the one provided in this chapter, there are some common themes emerging from the working groups that are worth mentioning. On one hand, the idea of Strategic Objectives in each topic. Then, the use of overarching principles. And finally, intersectional analysis

[441] https://tec.mx/es/noticias/guadalajara/investigacion/asi-sera-el-innovador-hub-de-inteligencia-artificial-del-tec (accessed on October 31, 2020).
[442] https://www.ia2030.mx/ (accessed on October 31, 2020).
[443] https://36dc704c-0d61-4da0-87fa-917581cbce16.filesusr.com/ugd/7be025_9e91bfffeea647a0a663630ea716aa8f.pdf (accessed on October 31, 2020).
[444] https://36dc704c-0d61-4da0-87fa-917581cbce16.filesusr.com/ugd/7be025_6f45f669e2fa4910b32671a001074987.pdf (accessed on October 31, 2020).

between topics throughout the document. On the other hand, the report presents key point indicators for every line of action suggested with the notice of who is responsible for implementing the course of action (i.e. the legislative, a regulatory agency, or civil society). From this viewpoint the report is not only a roadmap for the regulation of AI in Mexico, but a GPS of how possible regulatory frameworks intersect one another both in content, in scope and in normative terms.

IV. Conclusions

Looking at a moving image is not the same as looking at a movie. The AI-Applications in Mexico is a moving image, a composition of actors and initiatives, of businesses and technologies at a fast phase. Even if the arrangement of projects in this chapter may seem to have a sequential order, in reality there is a complex interaction between the AI ecosystem in Mexico that is hard to systematize. The complexity stems from a fragmentary view product of episodic initiatives that flourish with a specific grounding on an institutional program. Most of the use cases presented here are more the product of individual mindsets and talent than the result of an incubator-like environment for AI. Yet, the movement of AI initiatives captured here, and the ones left on the research for this paper hints that the Mexican market for AI-Applications is dynamic and promising. In a country with profound inequalities and with a still pending agenda on the Rule of Law and Human Rights there is always room for investing on institutions, and more if they promote welfare and inclusion via technological advancements. There are enough parties interested both in the private and the public sector –and in academia– to spark a fruitful institutionalization of AI in Mexico. Much of the inertia of a vastly unregulated market, however, could become the main trait of the Mexican case.

References

- A definition of AI: Main capabilities and scientific disciplines High-Level Expert Group on Artificial Intelligence, European Commission, April 8, 2019.
- Article 4 of the General Statute for Transparency and Access to Public Information at http://www.diputados.gob.mx/LeyesBiblio/pdf/LGTAIP_130820.pdf (accessed on October 31, 2020).
- Article 6 of the Mexican Constitution at http://www.diputados.gob.mx/LeyesBiblio/pdf_mov/Constitucion_Politica.pdf (accessed on October 31, 2020).
- Duffin Erin, Number of Mexican Immigrants in the United States 1850-2018 at https://www.statista.com/statistics/673350/mexican-immigrants-in-the-united-states/ (accessed on October 31, 2020).
- http://inicio.inai.org.mx/nuevo/GeneralRecommendationsfortheProcessingofPersonalDatainArti%EF%AC%81cialIntelligence.pdf (accessed on October 31, 2020).
- http://inicio.inai.org.mx/nuevo/SPECIFICGUIDELINESARTIFICIALINTELLIGENCEPROJECTS2019.pdf (accessed on October 31, 2020).
- http://omawww.sat.gob.mx/gobmxtransparencia/Paginas/documentos/estudio_opiniones/Evasion_en_IVA_Analisis_de_Redes.pdf (accessed on October 31, 2020).
- http://smia.mx/komputersapiens/ (accessed on October 31, 2020).
- http://www.cires.org.mx/sasmex_es.php (accessed on October 31, 2020).

- http://www.diputados.gob.mx/LeyesBiblio/pdf/LRITF_090318.pdf (accessed on October 31, 2020).
- http://www.ebizlatam.com/seguridad-e-inclusion-soluciones-conekta/ (accessed on October 31, 2020).
- http://www.ift.org.mx/sites/default/files/contenidogeneral/transparencia/1vision19-23.pdf (accessed on October 31, 2020).
- http://www.oecd.org/going-digital/ai/principles/ (accessed on October 31, 2020).
- https://adip.cdmx.gob.mx/proyectos (accessed on October 31, 2020).
- https://amqueretaro.com/queretaro/2020/10/22/convertiran-a-queretaro-en-ciudad-inteligente/
- https://bitso.com (accessed on October 31, 2020).
- https://blog.grillo.io/the-grillo-journey-e60b2dcee224 (accessed on October 31, 2020).
- https://buy.metricser.com/revelio/ (accessed on October 31, 2020).
- https://conekta.com (accessed on October 31, 2020).
- https://dapp.mx (accessed on October 31, 2020).
- https://data.oecd.org/mexico.htm#profile-innovationandtechnology (accessed on October 31, 2020).
- https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-cdmx/table/?q=FED%2FFECC%2FUNAI-CDMX%2F0000002%2F2020 (accessed on October 31, 2020).
- https://datos.gob.mx/blog/estrategia-de-inteligencia-artificial-mx-2018 (accessed on October 31, 2020).
- https://earthquaketrack.com/p/mexico/recent?mag_filter=7 (accessed on October 31, 2020).
- https://elceo.com/tecnologia/cnbv-alista-autorizaciones-para-80-empresas-fintech-entre-octubre-y-noviembre/ (accessed on October 31, 2020).
- https://english.elpais.com/elpais/2017/03/13/inenglish/1489403756_441981.html (October 31, 2020).
- https://ensanut.insp.mx/encuestas/ensanut2018/doctos/informes/ensanut_2018_presentacion_resultados.pdf (accessed on October 31, 2020).
- https://expansion.mx/tecnologia/2017/08/15/konfio-un-ejemplo-de-como-las-fintech-dejan-atras-a-los-bancos (accessed on October 31, 2020).
- https://gesedig.com (accessed on October 31, 2020).
- https://grillo.io/linux-foundation-hosting-openeew/ and https://grillo.io/open-source/ (accessed on October 31, 2020).
- https://gus.chat (accessed on October 31, 2020).
- https://home.inai.org.mx (accessed on October 31, 2020).
- https://homie.mx/h/ (accessed on October 31, 2020).
- https://inmobiliare.com/homie-la-plataforma-que-facilita-la-renta-de-departamento (accessed on October 31, 2020).
- https://inmobiliare.com/por-que-guadalajara-es-el-silicon-valley-mexicano/ (accessed on October 31, 2020).
- https://konfio.mx/ (accessed on October 31, 2020).

- https://kueski.com (accessed on October 31, 2020).
- https://kueski.com/blog/tecnologia/machine-learning-aprendizaje-supervisado/ (accessed on October 31, 2020).
- https://nearshoremx.com/cognitive_computing (accessed on October 31, 2020).
- https://news.microsoft.com/uploads/prod/sites/41/2018/11/IA-y-Crecimiento-MEXICO.pdf (accessed on October 31, 2020).
- https://nowports.com (accessed on October 31, 2020).
- https://panamericanworld.com/en/magazine/startups/meet-the-five-most-innovative-mexican-fintechs/ (accessed on October 31, 2020).
- https://publications.iadb.org/publications/english/document/FINTECH--Innovations-You-May-Not-Know-were-from-Latin-America-and-the-Caribbean.pdf (accessed on October 31, 2020).
- https://skyalert.mx/ (accessed on October 31, 2020).
- https://tec.mx/es/noticias/guadalajara/investigacion/asi-sera-el-innovador-hub-de-inteligencia-artificial-del-tec (accessed on October 31, 2020).
- https://www.adext.ai (accessed on October 31, 2020).
- https://www.banxico.org.mx/marco-normativo/normativa-agrupada-por-sujeto.html (accessed on October 31, 2020).
- https://www.cnbv.gob.mx/SECTORES-SUPERVISADOS/Fintech/Paginas/NORMATIVIDAD-FINTECH.aspx (accessed on October 31, 2020).
- https://www.consorcioia.mx/grupos-investigacion (accessed on October 31, 2020).
- https://www.eleconomista.com.mx/estados/Centro-de-Inteligencia-Artificial-en-NL-alista-inicio-de-operaciones-20180820-0020.html (accessed on October 31, 2020).
- https://www.eleconomista.com.mx/estados/Guanajuato-epicentro-de-la-cuarta-Revolucion-Industrial-20201029-0104.html (accessed on October 31, 2020).
- https://www.eleconomista.com.mx/finanzaspersonales/En-el-2019-70-de-las-nuevas-familias-compraria-una-casa-20181216-0053.html (accessed on October 31, 2020).
- https://www.eleconomista.com.mx/sectorfinanciero/Tenemos-mapeadas-mas-de-500-fintech-en-Mexico-CNBV-20190526-0072.html (accessed on October 31, 2020).
- https://www.eluniversal.com.mx/cartera/sat-va-por-mas-recaudacion-con-inteligencia-artificial (accessed on October 31, 2020).
- https://www.eluniversal.com.mx/estados/proyectan-puebla-como-ciudad-de-inteligencia-artificial (accessed on October 31, 2020).
- https://www.facebook.com/InteligenciaArtificial.Center/ (accessed on October 31, 2020).
- https://www.finnovista.com/radar/el-numero-de-startups-fintech-en-mexico-crecio-un-14-en-un-ano-hasta-las-441/ (accessed on October 31, 2020).
- https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers/ (accessed on October 31, 2020).
- https://www.gob.mx/cidge/articulos/crea-sfp-subcomision-de-inteligencia-artificial-y-deep-learning-de-la-cidge-161421?tab= (accessed on October 31, 2020).
- https://www.gob.mx/cms/uploads/attachment/file/520516/Comunicado_UIF_011.pdf (accessed on October 31, 2020).

- https://www.gob.mx/innovamx/articulos/inteligencia-artificial-131287 (accessed on October 31, 2020).
- https://www.gob.mx/misalud/ (accessed on October 31, 2020).
- https://www.gob.mx/shcp/prensa/comunicado-no-054-shcp-mediante-sat-y-pff-anuncia-acciones-en-contra-de-defraudacion-fiscal-a-traves-de-empresas-factureras (accessed on October 31, 2020).
- https://www.ibm.com/downloads/cas/JOEXYB08 (accessed on October 31, 2020).
- https://www.infobae.com/tecno/2020/04/01/asi-es-el-chatbot-mexicano-que-lucha-contra-la-desinformacion-sobre-coronavirus/ (accessed on October 31, 2020).
- https://www.innovatorsunder35.com/the-list/saiph-savage/ (accessed on October 31, 2020).
- https://www.insp.mx/resources/images/stories/2019/Docs/190213_LaObesidadenMexico.pdf, p. 28 (accessed on October 31, 2020).
- https://www.juniperresearch.com/new-trending/analystxpress/july-2017/chatbot-conversations-to-deliver-8bn-cost-saving (accessed on October 31, 2020).
- https://www.kubofinanciero.com/Kubo/Portal/productos/productos.xhtml (accessed on October 31, 2020).
- https://www.latamdigitalmarketing.com/blog/software-campanas-digitales/ (accessed on October 31, 2020).
- https://www.legalparadox.com/post/neo-banks-who-will-control-the-mexican-market (accessed on October 31, 2020).
- https://www.nanalyze.com/2017/07/7-chatbot-platforms-chatbots-easy/ (accessed on October 31, 2020).
- https://www.oracle.com/a/ocom/docs/dc/em/lpd100807811-impact-of-emerging-technology-on-cx-excellence.pdf?elqTrackId=d368a11a304041c8a7bf8e7a2f2a71e2&elqaid=82669&elqat=2 (accessed on October 31, 2020).
- https://www.oxfordinsights.com/mexico (accessed on October 31, 2020).
- https://www.redipd.org/es/la-red/historia-de-la-red-iberoamericana-de-proteccion-de-datos-ripd (accessed on October 31, 2020).
- https://www.researchgate.net/profile/Yasmin_Hernandez2/publication/340634786_Robotica_de_Servicio/links/5e96334a4585150839de623a/Robotica-de-Servicio.pdf (accessed on October 31, 2020).
- https://www.statista.com/study/45600/statista-report-fintech/ (accessed on October 31, 2020).
- https://www.uif.gob.mx/ (accessed on October 31, 2020).
- https://www.uv.mx/ciia/ (accessed on October 31, 2020).
- https://www.yalochat.com (accessed on October 31, 2020).
- https://www.youtube.com/watch?reload=9&v=1dv-gVKBRBA&ab_channel=UNAMGlobal (accessed on October 31, 2020). A technical report on the application can be found here: https://cscwcivictechnologies.files.wordpress.com/2020/10/asistente-virtual-saul-esparza.pdf (accessed on October 31, 2020).
- https://yucatanahora.mx/yucatan-busca-desarrollar-proyectos-de-inteligencia-artificial/ (accessed on October 31, 2020).

- Rosado-Tamariz E, Zuniga-Garcia M A, Santamaria-Bonfil G, Batres R, A machine-learning approach to speed-up simulation towards the design of optimum operating profiles of power plants. In Proceedings of the 8th International Conference on Informatics, Environment, Energy and Applications (IEEA '19). Association for Computing Machinery, New York, NY, USA: 194–199. DOI. https://doi.org/10.1145/3323716.3323735.

- Santamaría-Bonfil G, Reyes-Ballesteros A, Gershenson C, Wind speed forecasting for wind farms: A method based on support vector regression, Renewable Energy, 85, 2016: 790-809, ISSN 0960-1481, DOI. https://doi.org/10.1016/j.renene.2015.07.004.

- TOWARDS AN AI STRATEGY IN MEXICO: Harnessing the AI Revolution, White Paper Report, British Embassy in Mexico through the Prosperity Fund, Oxford Insights and C Minds, June 2018, p. 23. Electronic consultation at https://www.oxfordinsights.com/mexico (accessed on October 31, 2020).

- Wind Power Forecasting using Artificial Intelligence Tools, Ingeniería, InvestigacIón y Tecnología, 2018, 19-4: 1-11. DOI. http://dx.doi.org/10.22201/fi.25940732e.2018.19n4.033

- Yu Z, Han S, Zhu J, Sun X, Ji C, Guo X. Pre-pregnancy body mass index in relation to infant birth weight and offspring overweight/obesity: a systematic review and meta-analysis. PloS one. 2013;8:e61627. DOI. https://doi.org/10.1371/journal.pone.0061627 and Weng SF, Redsell SA, Swift JA, Yang M, Glazebrook CP. Systematic review and meta-analyses of risk factors for childhood overweight identifiable during infancy. Arch Dis Child. 2012;97:1019-26. DOI. https://doi.org/10.1136/archdischild-2012-30

Artificial intelligence (AI) systems are increasingly being used in our everyday life and in almost every kind of human activity, for instance in areas such as education and welfare, information society, judicial and law enforcement systems and recently to fight the Covid-19 pandemic. Very often referred to as "game changers", AI systems can bring about many benefits, but they can also raise complex and important legal, ethical, policy and economic issues.

This publication aims to feed the ongoing reflections within the CAHAI on the analysis of the challenges arising from AI systems and possible regulatory responses.

Firstly, it sets out to inform the reader of the progress of the work of the Ad Hoc Committee on Artificial Intelligence (CAHAI) and it presents several studies produced under the CAHAI.

Secondly, it brings in national perspectives of different observer States, from Israel, Japan and Mexico, to support the development of an international legal framework of artificial intelligence based on the standards established by the Council of Europe.

www.coe.int/cahai

PREMS 107320

ENG

www.coe.int

The Council of Europe is the continent's leading human rights organisation. It comprises 47 member states, including all members of the European Union. All Council of Europe member states have signed up to the European Convention on Human Rights, a treaty designed to protect human rights, democracy and the rule of law. The European Court of Human Rights oversees the implementation of the Convention in the member states.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE