COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 13 October 2020

CAHAI-PDG(2020)01

AD HOC COMMITTEE ON ARTIFICIAL INTELLIGENCE (CAHAI)

POLICY DEVELOPMENT GROUP

(CAHAI-PDG)

Draft Feasibility Study

V.0.3.

www.coe.int/cahai

1. General introduction

1. The Council of Europe is the continent's leading human rights organisation and the guardian of the rights of some 830 million Europeans. Throughout the transformations of our society since 1949, the Council of Europe has constantly ensured that human rights and fundamental freedoms, as well as democracy and the rule of law, guide development, including technological development. The Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data ("Convention 108[1]" and "108+"[2]), the Convention on Human Rights and Biomedicine ("Oviedo Convention"[3]) and the Convention on Cybercrime ("Budapest Convention"[4]), the Convention on Elaboration of a European Pharmacopeia[5] are some of the Organisation's legal instruments that have become recognised European or world standards, reconciling innovation and regulation for the benefit of human beings through common standards and frameworks. Although the European Convention on Human Rights (ECHR) does not specifically mention scientific and technological development, the dynamic interpretation provided by the European Court of Human Rights (ECtHR) in its case law has allowed it to address many different aspects related to this development.

2. Specifically, in the digital domain, the advances of the last decades have fundamentally transformed society by providing new tools for communication, information consumption, education, entertainment, commercial transactions and many other facets of daily life. Thanks to the detection of patterns and trends in large datasets using statistical methods, algorithmic systems now offer the possibility to recognise images or sound, streamline services and achieve huge efficiency gains in the performance of complex tasks. These services, commonly referred to as "artificial intelligence" (AI[6]), are presented as having the potential to promote human prosperity, individual and societal well-being by bringing about progress and innovation. At the same time, concerns are rising in public opinion in reaction to the media coverage of abuses or unintended consequences resulting from different types of AI applications, both in the private and public sectors. Discrimination, the advent of a surveillance society, the weakening of human agency, information disorders, electoral interference, attention economy, are just some of the concrete fears that are being expressed, in addition to the observation of the growing dependence on a technology developed and managed mainly by the private sector, which should be in line with the rule of law and the fundamental principle of democratic societies according to which all power must be accountable before the law.

3. Member states agree that economic growth is an important objective of public policies and consider innovation as one of its key components. However, innovation is nowadays developing at an increasingly rapid pace and causing profound and extraordinary societal changes. Until

[1] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No. 108

[2] Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, CETS No. 223

[3] Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine, ETS No. 164

[4] Convention on Cybercrime, ETS No. 185

[5] Council of Europe (1964) Conention on Elaboration of a European Pharmacopeia (CETS no. 050, adopted in Strasbourg on 22 July 1964 (entered into force on 22 May 1974)

[6] In order to avoid any form of customisation and to maintain a technologically neutral approach, the terms "AI systems, ", "AI applications" or "AI tools" will be preferred in this feasibility study to refer to algorithmic systems based, indifferently, on machine learning, deep learning, rule-based systems such as expert systems or any other form of computer programming and data processing. The notion of "algorithmic systems" is to be understood as defined in the appendix to Recommendation CM/Rec(2020)1 of the Committee of Ministers, as "applications which, often using mathematical optimisation techniques, perform one or more tasks such as collecting, grouping, cleaning, sorting, classifying and deriving data, as well as selecting, prioritising, making recommendations and taking decisions. By relying on one or more algorithms to perform their tasks in the environments where they are implemented, algorithmic systems automate activities to enable the creation of scalable, real-time services. »

now, this evolution in the field of AI applications has essentially been accompanied by a very large number of ethical, non-binding and frameworks which are not enforceable[7]. Based on this acquis, a legal and binding response can now be built where needed, as has already happened with innovative industrial processes such as pharmaceuticals, biomedicine or the automotive industry. Council of Europe member states have a duty to ensure that the fundamental values of human rights, democracy and the rule of law remain effectively anchored in appropriate legislative frameworks and applied throughout ongoing societal and technological developments. This is a response to their positive obligations under the European Convention on Human Rights[8] and a means to ensuring that individuals are not left without effective protection of their rights in the face of technological development and application.

4. Therefore, on 11 September 2019, the Committee of Ministers mandated an Ad hoc Committee on Artificial Intelligence (CAHAI) 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 Council of Europe standards in the field of human rights, democracy and the rule of law. This feasibility study takes into account Council of Europe standards for the design, development and application of digital technologies in the field of human rights, democracy and the rule of law, in particular on the basis of existing legal instruments, including relevant international - universal and regional - legal instruments. It also takes into account work carried out by other bodies of the Council of Europe as well as work in progress within other regional and international organisations (in particular within the United Nations – including UNESCO, ITU, WIPO and WHO - European Union, OECD, OSCE, the World Bank, and the World Economic Forum). Finally, this study takes into account a gender perspective and the building of cohesive societies and the promotion and protection of the rights of persons with disabilities.

[7] See in particular the work carried out by the Secretariat of the Council of Europe available on the dedicated website: https://www.coe.int/en/web/artificial-intelligence/national-initiatives

[8] See the conclusions of the Helsinki High-Level Conference "Governing the game changer - the impact of the development of artificial intelligence on human rights, democracy and the rule of law", co-organised by the Finnish Chairmanship of the Committee of Ministers of the Council of Europe and the Council of Europe

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

5. To date, there is no definition accepted by the entire AI scientific community. The term, which has become part of everyday language, appears to cover a very wide variety of sciences, theories and techniques of which the aim is to have a machine reproduce the cognitive capacities of a human being. This ambition has been expressed since the creation of computer science and the term can therefore cover, in a broad sense, any automation resulting from this technology, as well as, in a narrower sense, very precise technologies such as machine learning or deep learning based on neural networks.

6. Similarly, despite considerable efforts to propose some form of harmonisation, the work resulting from the various international organisations has also not led to a consensus on the definition of the term AI. The High-Level Group of Independent Experts mandated by the European Commission has therefore published a comprehensive document on the issue[9]. The European Commission's AI Watch Observatory has also published a very thorough study on an operational definition and taxonomy of AI[10]. The OECD Council Recommendation on AI also includes a preamble defining AI systems, the life cycle of an AI system, AI knowledge, AI actors and stakeholders[11]. UNESCO has also produced its own definition in a preliminary study on the technical and legal aspects related to the desirability of a normative instrument on the ethics of artificial intelligence[12] and in its draft Recommendation on the Ethics of Artificial Intelligence. The preliminary study thus refers to "AI-based machines" and "cognitive computing", which "have the potential to imitate or even surpass human cognitive abilities such as detection, linguistic interaction, reasoning and analysis, problem solving and even creativity". The paper also uses the term "intelligent machines" which "can demonstrate learning abilities comparable to those of humans, with self-relation and self-correction mechanisms, through machine or automatic learning algorithms, or even deep learning, using neural networks that mimic the functioning of the human brain".

7. When examining the legal instruments of the Council of Europe relating to scientific fields, it can be seen, for example, that the Convention on Human Rights and Biomedicine ("Oviedo Convention"[13]) does not define the subject matter it governs. The Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data ("Convention 108[14]" and "108+"[15]), for its part, defines the concept of "data processing", which derives from the application of information technology, without mentioning specific technical objects such as algorithms. The Convention does, however, precisely define its subject, "personal data", thus making it possible to determine whether or not a processing operation falls within its scope.

8. Among the non-binding instruments, Recommendation CM/Rec(2020)1 of the Committee of Ministers to member states on the impact of algorithmic systems on human rights[16] defines the notion of "algorithmic systems" which may cover all or part of AI applications. The Declaration

[9] AI HLEG, A Definition of AI: Main Capabilities and Disciplines, April 2019

[10] AI Watch, Joint Research Centre, Defining Artificial Intelligence : towards an operational definition and taxonomy of artificial intelligence, February 2020

[11] OECD, Council Recommendation on Artificial Intelligence, June 2019

[12] UNESCO, Preliminary study on the technical and legal aspects relating to the desirability of a standard-setting instrument on the ethics of artificial intelligence, March 2019

[13] Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine, ETS No. 164

[14] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No. 108

[15] Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, CETS No. 223

[16] Council of Europe, Committee of Ministers, Recommendation CM/Rec(2020)1 of the Committee of Ministers to member states on the human rights impacts of algorithmic systems, April 2020

of the Committee of Ministers on the Manipulation Capabilities of Algorithmic Processes (**Decl(13/02/2019)1**) does not include definitions and uses indifferently concepts such as "technologies", "data-based systems", "machine learning tools" depending on the specific objects to be considered[17]. The Commissioner for Human Rights[18], the Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (T-PD[19]) and the European Commission for the Efficiency of Justice (CEPEJ[20]) use a relatively similar generic definition referring to a set of sciences, theories and techniques..

9. With regard to a legal definition of AI for use in a new Council of Europe instrument, the CAHAI's work has also revealed different approaches among members, participants and observers resulting, in particular, from different legal traditions and cultures. While there seems to be a consensus on the need for technological neutrality, a more precise definition has been discussed. A balance was sought by the CAHAI between a definition that is too precise, much less technologically neutral and possibly obsolete in the short or medium term, and a definition that is too vague which, by leaving a wide margin of interpretation and better adaptability, would lead to some forms of AI applications escaping this regulation. The result is the possibility for the future legal framework to adopt in its preamble or first articles a simplified and technologically neutral definition of its purpose, focusing on the effects of AI systems on human rights, democracy and the rule of law and their socio-technical implications.

3. Opportunities and risks arising from the design, development and application of artificial intelligence on human rights, the rule of law and democracy.

A. Distinguishing AI technologies in risk-reward context

10. The term "Artificial Intelligence" or 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". In practical application the term AI is commonly used to refer to a loose collection of technologies which 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. Sometimes AI may be understood as "mimicking human cognition" or "conducting tasks normally requiring human intelligence" which equally applies to applications affecting human rights. For the purpose of identifying the idiosyncratic opportunities and risks associated with AI technologies, the focus may be on the features of systems, featuring AI technologies (AI systems, or AIS) which set them apart from the any other technologies in relation to human rights, democracy and the rule of law.

11. Such distinguishing features, which set AI systems apart, may include:

[17] Committee of Ministers, Declaration of the Committee of Ministers on the manipulative capabilities of algorithmic processes, February 2019

[18] Commissioner for Human Rights, Unboxing AI: 10 steps to protect human rights - Recommendation of the Commissioner for Human Rights, May 2019

[19] Data Protection Convention Advisory Committee on the Automated Processing of Personal Data (T-PD), Guidelines on Artificial Intelligence and Data Protection, January 2019

[20] CEPEJ, European Ethical Charter for the use of artificial intelligence in judicial systems and their environment, December 2018

- the capacity to achieve a certain goal [*Editorial note: technically it would be more accurate to say "enhance its performance towards a certain goal"*] by acquiring, processing and applying data, knowledge and skills;
- designed to use autonomous or human-assisted formulation and/or amendment of symbolic rules, or a model as the main tool(s) of achieving a certain goal;
- the capacity to purposefully interact and influence their environment, etc.

12. A system matching one or more of these criteria will be treated as AI system. The list of criteria can be amended and updated based on the identification of risks for human rights, democracy and the rule of law arising from particular properties of new technologies.

B. THE PROMISE OF ARTIFICIAL INTELLIGENCE

13. AI, just like many earlier technological developments is malleable and combinable, and this very flexibility enables immeasurable positive impact across the society, including areas such as medicine, environment, and social sciences. Many opportunities arise regarding welfare and competitiveness, education and training, research, Innovation etc. The use of AI systems may improve well-being of society and it may be even crucial during pandemics and other period of crises. AI technologies may encourage people to perform as active citizens, to enhance the political participation and to be conducive for the rule of law, democracy and human rights.

14. For example, AI may provide opportunities for human rights by eliminating illegal discriminatory bias in advisory AI systems.

15. However, opportunities may be accompanied by risks, or may (eventually) turn into risks themselves. Also there is the ever present risk of abuse/misuse of AI systems, even when it is meant 'for good'. The complexity and rapid development of AI technologies raise concerns among the people about their inherent safety and potential for abuse; technologies which fail to meet reasonable ethical expectations of people, including protection of basic human rights, will likely face resistance and may not be able to realize their positive potential. Effective regulation, protecting basic human rights and generally addressing the societal concerns about negative effects of AI technologies is, therefore, an essential element of supporting innovation for the good of the society.

16. It is of utmost importance to strike the right balance between managing risks on the one hand and supporting technological innovation on the other. The ecosystems of trust and excellence are closely interlinked and should reinforce each another. The regulation serves to align the interest of technology actors and the society, thus eliminating barriers to the acceptance of new technology.

17. Risks due to the use of AI manifest around the rights of non-discrimination, data protection, privacy rights, freedom rights, security, legal protection, social rights (access to public goods such social benefits, housing, health care (incl. insurance), education, credit etc. The following considers the risks which are inherent to AI technologies, and, separately, sources or risk which arise due to actions of agents using AI systems.

i) Precautionary principle

18. In this regard, risk-based approach towards AI systems can follow the best practice that mankind has in adopting new transformative technologies to secure positive impact of AI on society. As with the other cases where the impact of new technology cannot be reliably predicted, the precautionary principle appears to offer the best risk mitigation practice. The principle requires the agent implementing the action to assess the degree of danger (risk, threat), carefully weigh and address the risks of the worst case scenarios, or those scenarios where the potential magnitude of adverse consequences has no reliable limit. This includes both direct and possible indirect, secondary consequences of decisions or actions. The principle expresses the fundamental responsibility of decision-makers, politicians, scientists, designers, company managers, and so on for the impact of their actions on the society in terms of human rights, democracy and law.

19. According to the precautionary principle, if a certain action could cause damage with scientifically uncertain probability and magnitude, actions should be taken to prevent such damage and/or to limit its magnitude. If no effective mitigation measures can be designed, the potentially dangerous action should be avoided.

20. Application, and assessment of impact of AI systems in the area of human rights, rule of law and democracy requires the consideration of a technically complex and imperfectly determined intersection of interests of people, private sector, governments and society. Precautionary principle provides the best practice solution to resolving the inevitable uncertainties in this regard.

Ii) Principle of informed consent

21. The principle is widely used in medicine, including bioethics and Biomedicine, and personal data governance (OECD Recommendation on Health Data Governance, 2017). In the case of AI systems, the principle of informed consent may be understood as (1) informing the user that AI technologies are applied to make decisions which may impact her in any significant way; and (2) obtaining their consent to be exposed to the consequences of action by AI. Because of the complexity of AI technologies, the user should further receive appropriate information about the performance characteristics of AI system which determine its potential impact.

22. While this may appear as a considerable burden and a possible barrier in the application of AI technologies, it may be recalled that in the medical field the principle of informed consent serves to protect both the patient and the doctor. By preventing the penetration of «unseen» AI applications, the principle of informed consent will serve to address the long-term concerns and resistance to AI in the society. By clearly identifying those cases where AI regulations are be applied, it will also eliminate legal «gray zones» of technology and serve to protect the developers of new technologies from potential legal action due to unforeseen negative impact of their solutions.

23. The principle is also used when working with data for scientific research, various tests and experiments (the relationship between the researcher and the subject) – "Scientific research should only be carried out with the prior, free, explicit and informed consent of the person concerned. The information must be adequate, provided in an understandable form, and include an indication of how to withdraw consent. Consent may be withdrawn by the person concerned at any time and for any reason without negative consequences or damage." (Universal Declaration on bioethics and human rights, UNESCO, 2005). This is directly relevant for the preparation of training data sets for AI systems; the principle means that the collection of personal data (PD) and other information related to the private life of a citizen is carried out only with prior, free and informed consent.

24. Consent to the processing of PD is one of the most pressing issues in the field of digital technology regulation, and given the scale and scope of the use of AI systems, the issue becomes particularly important. In this context one further needs to recognize the difficulties associated with the right to revoke consent to the processing of PD in the case of AIS. The procedures used in training AI systems, primarily neural networks, do not imply a procedure for "de-learning", i.e., removing data from the structure of the trained algorithm. When mass AI is introduced, it is necessary to conduct appropriate comprehensive research to protect the rights of citizens, including the "right to forget" data.

25. It appears reasonable to propose that informed consent in the case of AIS would mean the obligation to provide full information about the nature and features of AI systems, the possible

consequences of their use, the procedure for processing and conditions for protecting PD will fully guarantee the rights of citizens, implement a human-cantered AI which duly reflects the primacy of human rights and freedoms.

C. Inherent risks of AI

i) Risks due to domain dependence of narrow AI

26. As recognized in [1], "narrow" AI systems differ from the General AI in that they can perform only very specific 'narrow' tasks. At this time all AI systems in practical use are narrow AI systems, which are technically capable of achieving their goals only if used in the appropriate environment, otherwise known as the "domain" of AI system. For example, an image recognition system may require certain minimal image resolution and colour depth to deliver its intended performance; a medical diagnostic system based on such image recognition will present a stability/robustness risk if applied to images of unsuitable quality. In general, application of narrow AI out of its context presents an inherent risk, particularly sensitive in areas where human rights may be affected.

27. Furthermore, a mistake in the specification of the domain where AI system is intended to operate can introduce a range of risks to human rights. Credit card fraud detection systems already use AI technologies today to "learn" typical user behaviour and restrict transactions which appear unusual. These systems may discriminate against people from vulnerable socioeconomic backgrounds due to their "unusual" spending behaviour; the use of ethnic data is usually very harmful where AI is used to detect fraud.

28. The "Question Zero", i.e. whether a particular AI application has a place in the life of a society if it is posing risks to human rights, the rule of law, and democracy, has a differentiated interpretation depending on the type of AI system and stage we address. What renders a particular technology "benevolent" or "malicious" is in many ways context-specific and, further, dependent on the intent of those who deploy it as shown in subsequent sections.

29. Making a distinction between so-called "green and red areas" requires careful specification and delimitation of the domains where AI technology might pose high or low risk to human rights. For example, the same image recognition technology may be considered low risk in amateur photo applications, and will be high risk in medical applications. A high-risk application needs to respect the domain, and performance characteristics of AI system to achieve its goals without creating undue risks. This means a more nuanced understanding of "red" and "green" lines including, as a minimum:

   a. identification of the domain of AI system,

   b. identification of the particular application task, including its impact on human rights, and

   c. identification of the performance characteristics required to perform the task without undue risk.

30. Rapid development of AI technologies also means that domains of AI systems are hardly ever static and domain misspecification may result in an incomplete picture of the benefits and risks that are at stake.

ii) Risks due to autonomous behaviour of AI

31. It is understood that AI systems can exhibit autonomous behaviour beyond what is be commonly expected from an engineered system. Under most practical applications the impact of such unexpected behaviour will be quickly detected and rectified by re-training or re-specifying the AI system, so that the negative impact will be contained. The degree of risk created in such cases where a failure would not create a cascading, scaling impact, can be predicted, modelled and mitigated.

32. However, in complex engineering systems, such as civil infrastructure, or airplanes, introducing AI elements creates the risk of cascading critical failure caused by unexpected and otherwise harmless behaviour of AI systems. For instance, while the failure of Boeing's "Maneuvering Characteristics Augmentation System" may or may not be considered a clear cut case of AI system, it may serve as the most recent illustration of cascading failure with catastrophic consequences caused by a fairly simple unexpected behaviour of an information system.

iii) Risks due to superior learning capacity

33. The superior learning capacity of the modern AI systems is already an accomplished fact at least in some practical applications, e.g. in the game of Go where Google's Alpha Go was able to develop superior strategies without human involvement, by playing and learning from itself.

34. As AI systems are increasingly used to organize and manage the information space and make it more accessible to human users, there is a risk that superior learning capacity of AI systems might overwhelm the user's own capacity to navigate the information space. Depending on the design and training of AI system, this carries numerous risks for the person and the society. For example, AI based personal news recommendation system may tend to evolve towards a biased and self-reinforcing presentation of news. The example of Microsoft's "Tay" chatbot which quickly learnt the least tolerant and abusive behaviour from Twitter discussions serves to illustrate the magnitude of the problem created by fast learning in the absence of inherent ethical principles in AI.

35. Superior learning capacity of AI creates a real risk that an AI would act in a harmful, manipulative or discriminatory way while adapting to human behaviour or it may learn a way to resolve a benign task in a harmful way to a human being in the process of learning.

iv) Risks due to lack of transparency

36. Risks due to lack of transparency mainly occur in the context were an AI system has an inclination towards opacity and bias in AI (including bias in data, in algorithms, in the analytics, in the result, etc).

D. Risks created in application of AI

37. What renders a particular technology "benevolent" or "malicious" is also dependent on the intent of those who deploy it. One may assume that that each agent involved in AI applications is a source of risk, either by mistake or by malicious design. Those risks are not unique to AI systems, and are present in the application of any complex technology. Historic experience in managing technology-related risks will be instructive in this context.

38. Malicious applications are understood as application of AI with intent to exploit the common good for the sake of private gain. As any new technology, penetration of AI in a society changes social structures around it, including people, organizations and institutions. New ways of abusing the new power structures in the society are likely to emerge and they will have impact human rights. This section explores the risks which are not inherent to AI but may emerge from the behaviour of agents using AI systems, i.e. not from the AI systems themselves. It may be recalled in this regard, that Ai systems do not have "common sense" or ethics and can only act in an ethical or unethical way if so directed by the actors of AI systems value chain.

39. Typically there are many agents involved at different stages of the value chain leading to practical applications of AI systems. Where AI appears in chains of processing, with many actors and layers involved, many questions revolving around accountability and responsibility

emerge. As these may not be specific to AI systems, they are urgent questions, given that harmful AI systems rapidly amplify harm.

40. In substance, this agent-related risk translates directly into the challenge of identifying clear responsibility of agents for any negative impact from their actions in the design, development, implementation and deployment of AI systems.

41. Because of the complexity of AI systems it is generally difficult to identify the responsibility on a «black box» basis ex post after the adverse event. The issue of identifying the responsibility for malicious use of AI is likely to remain a legal challenge for the foreseeable future; it will likely be addressed to a significant extent by refining the application of existing legal instruments to the cases of AI technologies. For example, medical devices in clinical use require certification, which assigns the responsibility for device (mis-)performance to the manufacturer. The same approach will be effective for devices using AI technologies, where the certification requirements will need to be updated accordingly, including the considerations given above, i.e. the precautionary principle, the principle of informed consent, and the inherent risks of AI technologies.

42. Where necessary, existing legal instruments may need to be supplemented with new instruments addressing the circumstances unique to AI systems. The latter may include, for example, the use of AI systems for creating and distributing deep fakes, manipulation of human behaviour using A systems, abuse of training data to bias AI system behaviour, and so on.

i) Responsibility of individuals and civil society

43. Individuals and civil society organizations may use AI systems to achieve certain goals which have impact on the welfare of others, or on the society as a whole. Risks to human rights may emerge from this application of AI through either a human error or malicious intent, e.g. through the distribution of fake news on social media. These risks generally point in the direction that Individuals and civil society organizations need to observe the applicable national laws and ethical standards of the society.

ii) Responsibility of corporations

44. Corporations are the main source of practical solutions applying AI technologies in the life of the society. They carry the primary responsibility for providing accurate and sufficient information about risks and benefits of their solutions to all members of the society who may be affected. In this regard, the general consumer protection legislation existing in Member States provides the basic framework regarding the responsibilities of corporations in mitigating the risks created by their solutions, including those using AI.

45. Corporations would also play important role in developing and applying industry standards, codes of ethics and codes of practice for responsible application of AI in the civilian market. As with other professions where a specialist makes decisions that have significant consequences for other people in an environment of asymmetric and incomplete information, ethics and codes of practice may work well to address the concerns in the society regarding the impact of corporate AI systems and practices on human rights.

46. A particular challenge for corporations lies in developing and systematically deploying “safety switches” that could be used to exercise human control over AI systems in cases these are found to operate to the detriment of a society, either by design mistake, or by malicious intent. While technically this would mostly not present a problem, it does increase the power of corporations in the society and poses additional issues in terms of legal and ethical responsibility that corporations should carry with this additional power of influence.

iii) Responsibiity of Governments

47. Any government use of AI which potentially might result in a breach of human rights should meet the demands of subsidiarity and proportionality and be based on the law. The general commitment of Governments to uphold the human rights, democracy and the rule of law applies in full to government uses of AI solutions.

48. The Governments further have to responsibility to update and apply the legal system to AI applications in a way that encourages fair market competition and innovation, while clearly attributing legal responsibility for the consequences of risks posed by AI to the actors in AI value chains.

49. The law should be accessible and should be so precise as to make it foreseeable when the infringement would be allowed. There may also be a positive obligation for government to protect human rights in market situations (where current laws/regulations do not already effectively do this).

50. A significant challenge of intergovernmental efforts relates to the challenge of regulatory arbitrage, whereby differences in national legal regimes may be exploited to the detriment of the society. The cross-border nature of cyberspace means that many AI applications can be applied across national boundaries and jurisdictions and coordinated intergovernmental effort will be required to maintain the effectiveness of legal instruments protecting human rights, democracy and the rule of law.

E. Risks to human rights, Risks to the rule of law, Risks to democracy by areas of impact

[*editorial note: this section is covered in detail in the CAHAI(2020)06-fin document. It may be summarised here, or dropped altogether in favour of discussion elsewhere in the document*]

51. Domain dependence of narrow AI stands in a sharp contrast with the universality of human rights. It is not technologically possible at this time to have an AI system which could maintain its performance in all situations when human rights are affected. Any application of AI systems in areas where human rights are affected may, under some circumstances, pose a risk. Some examples below are intended to illustrate how risks to human rights, rule of law, and democracy can be related to the sources of risk in a particular application context. Accurate identification of the sources of the AI application context and of the relevant risks is essential to identify the responsible actors, and the need for mitigation measures.

i) Respect for Human Dignity

52. As one example of AI application that can pose risks to human dignity, predictive policing and law enforcement needs to be mentioned. Vulnerability of AI systems to biased data, and inevitable statistical errors create the risk of violating the rights for Liberty and Security, Fair Trial, No Punishment without Law (art. 5, 6, 7 ECHR). These risks are mainly the responsibility of corporations, providing AI applications in this area, and the Governments.

53. The corporations have the responsibility to identify the application domain of the AI system, and to make sure that the system will maintain the essential performance standards when used within the domain. It is further incumbent on the corporations supplying Ai solutions in sensitive areas to ensure that those systems are robust and will not lose their performance characteristics under the influence of adverse factors.

54. The Governments have to take the responsibility that the use of AI for law enforcement is governed consistently with the standards of protection for human rights applicable to other sensitive technologies. In particular, it may be feasible to provide that any use of AI system in law enforcement must be consistent with the manufacturer’s specification of the environment where the minimum necessary performance of the system can be guaranteed.

ii) Freedom of the Individual

55. 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). It isi recognized that AI systems can have a ‘chilling’ effect on these freedoms, both in terms of freedom of expression, and in terms of freedom of assembly and association.

56. Both corporations and Governments have the capacity to use AI-assisted data processing capacity to extract information about individuals beyond what the individual might be willing to disclose about herself. This power may be matched with a commensurate responsibility for its use, e.g. responsibility not to act to the detriment to individual’s rights.

57. In terms of freedom of expression one may further note that media space is no longer populated with content created and disseminated solely by a limited number of media workers alone, who are bound by professional and ethical standards, but also by citizens. As a result, there is not necessarily any editorial control over a vast amount of published content. These processes have had a tremendous impact on audience behaviour and information consumption. On the one side, this increases the power of individual in reaching out; on the other side, it creates a severe impediment to being heard. Moderation of this many-to-many communication space to achieve a balanced protection of freedom of expression is also a matter of AI application.

58. The impact of AI moderation on freedom of expression, both positive and negative, is still severely under-explored, and does pose significant risks. Sources of risk may include inherent properties of AI systems (e.g. emergent bias due to incorrect domain specification), or intended or unintended abuse of AI systems in media space by individuals, corporations of Governments.

iii) Equality, Non-Discrimination and Solidarity

59. in many cases, AI systems have shown to perpetuate and amplify and possibly enshrine discriminatory or otherwise unacceptable biases. Also, AI systems can enlarge the group of impacted people, when it groups them based on shared characteristics. In terms of inherent properties of AI technologies, the risk relates to lack of transparency in AI systems, which can easily obscure the existence of biases, marginalising the social control mechanisms that govern human behaviour.

60. Any AI system has to be free from biases due to technical flaws, i.e. its operation should accurately reflect the properties of input data. However, beyond technically accurate interpretation of data, the decisions produced by AI system are also expected to agree with the norms of “fair” decision making accepted in the society. These expectations are changing, and it is up to the actors of AI systems value chain to impart this behaviour on the system. It also has to be recognized that 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 systems likely to remain wishful thinking, and bias will remain a constant risk in application of AI.

iv) Social and Economic Rights

61. AI systems can have major benefits when used for hazardous, heavy, exhausting, dirty, unpleasant, repetitive or boring work. However, wide adoption of AI systems in social and economic life create new risks to social and economic rights. Examples of some areas where these risks are most evident are given below.

62. Applications of AI for monitoring workers may 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). Constant monitoring by AI systems may also reduce the opportunity to organize (art. 5). Biases in AI-systems that assess and predict performance of workers would undermine

equal opportunities and equal treatment in matters of employment and occupation and may reinforce existing biases within the data or of their creator.

63. 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.

64. Expanding use of AI technologies will, in all likelihood, cause a disruptive change in patterns of employment, and will lead to a range of lasting socioeconomic consequences affecting the social and economic rights of workers. It can be expected that few, if any areas of social and economic activity will avoid exposure to this change, and related risks. One possible exception may be jobs that require direct human-to-human communication, were re-orientation and re-development of skills may offer some mitigation strategies for the disruptive change.

v) Democracy

65. AI systems do not have an impact only on individuals but on society and democratic processes as a whole. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals.

66. Well-functioning democracies require a well-informed citizenry, an open social and political discourse and absence of opaque voter influence. Yet 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 systems to determine which content is created and shown to users (information personalization). 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. As a consequence, it may undermine the shared under-standing, 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 meaningful decisions by voters. AI systems with manipulative capacities can covertly impact human behaviour especially on political and judicial issues or ethical and moral principles.

67. AI will not only lead to undesirable side effects but could also empower malicious actors ranging from cybercriminals to totalitarian states in their desire to control citizens. In particular, if decisions that previously were made by many decentralised actors are replaced by few centralised AI-driven systems, the systemic risks increase.

vi) Rule of Law

68. AI systems can increase the efficiency of institutions, but on the other hand it can also erode the procedural legitimacy of and trust in democratic institutions and the authority of the law.

69. AI systems can contribute to developing judicial systems that operate outside the boundaries and protections of the rule of law. While automated online dispute mechanisms provided by private companies can have the potential to enable consumers to act on their rights, when enforcing a claim in court is not feasible or too costly, they 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. 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.

F. Mitigating risk and aiming for positive impact of AI

[*Editorial note: this section can connect risks to general principles of impact assessment, positive and negative scenarios as introduced by delegations in their comments on Chapter 3. In general, this needs to be coordinated with Chapter 8*]

70. The challenge is not in classifying technologies, or even uses of technologies, as "positive" or "problematic", but rather in ensuring that policy makers, IT experts and regulators have the tools, knowledge and technological understanding necessary to harness AI for increasing public good, while enforcing existing laws and complying with their international obligations. Risk and impact are two sides of any technology, and potential impact generates risk. Risk is the likeliness of an impact to occur.

71. Impact assessments may fulfil an important role here. It may be necessary to pre-emptively identify the potential impact of AI systems on fundamental rights, security and safety (including vital infrastructure interests), and other public values (such as accountability and responsibilities, good governance). Sources of risk, and appropriate mitigation strategies can be identified. Predictive risk assessment starting with early design stages of AI systems for sensitive applications may be desirable.

72. Particular practical measures in this regard may include:

- introducing standards, requirements and best practices for developing, use and assessing AI systems;
- performing adequate risk assessment before launching an AI system in a sensitive application;
- developing of the national scientific-based prognostication systems for consequences of AI systems use for economic, labour, legal, cultural, medical, biological and psychological effects on human society and a human being itself;
- establishing rules on attribution of responsibility for the consequences of the use of AI systems;
- providing for sufficient human control over the actions of AI systems;
- using mitigation measures such as open datasets which can on their own reduce possible pre-visible risks of AI misbehaviour for training an AI;
- informing operators and users of AI system about its basic principles of operation and relevant performance characteristics;
- disclosing the fact of interaction with AI systems;
- recognizing and mitigating any possibility of degradation and other negative changes in human natural intelligence arising from the creation and use of AI technologies.

4. The Council of Europe's work in the field of artificial intelligence to date

73. The significant impact of the various applications of information technology on human rights, democracy and the rule of law has led the Council of Europe to build specific framework mechanisms, complementary to the European Convention on Human Rights. These mechanisms complement and reinforce each other; they form a coherent whole that is not limited or mutually exclusive. The binding (protection of personal data and the fight against offences committed or facilitated by computerised means) and non-binding frameworks emanating from the various sectors of the Organisation (recommendations, resolutions, charter, guidelines, reports and studies) will be examined in turn. The European Convention on Human Rights will be examined through its case law on new technologies.

4.1. Work in the area of personal data protection

74. Public concern about possible or actual excesses in the use of information technology by public authorities in the 1970s had provoked reactions from national legislators in Sweden[21], Germany[22] and France[23]. The Council of Europe has built on these initiatives by drawing up the very first international text on the "protection of individuals with regard to automatic processing of personal data" in 1981, which came into force on [1] October 1985[24]. "Convention 108", modernised by a protocol in 2018[25] ("Convention 108+"), continues to be a relevant defence of the protection of the privacy of individuals, regardless of the technological revolutions that have taken place since then. In particular, it prohibits the computer processing of sensitive data (e.g. political opinions or sexual life) and creates a right for everyone to know existing data about themselves, with a right of rectification. It also provides for a free flow of data between the parties that have ratified this Convention, well beyond the borders of the Council of Europe[26]. The protocol added new principles, such as transparency, proportionality, accountability, data limitation, respect for privacy by design and clarified certain definitions by replacing, for example, the notion of "automated file" used in the 1981 text by the notion of "data processing" (Article 2, b). The protocol also maintained the principles of *fair information* (fairness, purpose, data quality, rights of data subjects) with certain adjustments. Thus, fairness is no longer understood to apply only to the moment of collection, but now applies to all processing operations, with a general obligation of transparency (Article 8). With regard to quality, a principle of limitation (or minimisation) of data has been introduced, implying for data controllers an obligation to collect only the minimum necessary (Article 5). Finally, as regards the rights of individuals, the right not to be subject to an automated decision and the right to obtain knowledge of the reasoning underlying the processing of data, where the results of the processing are applied, has been introduced (Article 9). This new right is of particular importance in relation to the profiling of individuals[27].

75. Even if it is not specific to AI applications, the legal framework built around Convention 108 remains fully applicable to this technology as soon as the processed data fall within the scope of this text. Guidelines and a report in 2019 specified the guiding principles to be applied, both for legislators and decision-makers and for developers, manufacturers and service providers[28]. A legal instrument on AI applications will therefore have to take full account of these acquis to supplement them by including in its scope of application processing operations not involving personal data and by extending its scope, not only by preventing individual but also societal harm.

[21] Law of 11 May 1973 defining a protective status in computer matters

[22] Law of 10 November 1976 which came into force on 1 January 1978

[23] Law of 6 January 1978

[24] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No. 108

[25] Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, CETS No. 223

[26] Convention 108" has been ratified by the 47 member states of the Council of Europe and 8 non-member states (Argentina, Cape Verde, Mauritius, Mexico, Morocco, Senegal, Tunisia, Uruguay). Convention 108+" has already been signed at the time of writing by 34 member states and 4 non-member states (Argentina, Mauritius, Tunisia and Uruguay).

[27] See in this respect Recommendation (2010)13 on the protection of individuals with regard to automatic processing of personal data in the context of profiling, and the explanatory memorandum.

[28] Advisory Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (T-PD), Guidelines on Artificial Intelligence and Data Protection, January 2019 and Advisory Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (T-PD), Report on Artificial Intelligence (Artificial Intelligence and Data Protection: Challenges and Possible Solutions), January 2019

4.2. Work on criminal offences committed or facilitated by AI applications

76. The various applications of AI open up much-feared prospects in the field of cybercrime: automated and coordinated computer attacks, automatic searches for flaws in systems, AI constitutes a veritable arsenal of digital warfare. The Convention on Cybercrime ("Budapest Convention") remains today the international reference instrument, notably by creating a unique judicial co-operation framework adapted to the cross-border nature of these offences in addition to being guidelines for States wishing to adopt specialised legislation[29]. This legal text is fully applicable to acts carried out or facilitated by AI tools, it being specified that while the moral element of intentional offences in cyberspace is not legally complex to apprehend, the search for and exploitation of material elements to attribute the acts to a specific perpetrator is certainly more complex.

77. However, the purpose of this text does not concern unintentional offences, which do not systematically come under criminal law depending on the legal system. Unintentional damage caused, for example, by an autonomous vehicle falls within the scope of the liability regime, in which the relevance of the mechanisms for AI applications is currently being studied by the CDPC (European Committee on Crime Problems), which may propose the creation of a new specialised legal instrument[30]. This Committee has moreover expressed the wish to coordinate its activities with the activities of the CAHAI before initiating work in this field.

4.3 Work in the areas of Education, Children and Youth

78. The Committee of Ministers' Recommendation CM/Rec(2019)10 on developing and promoting digital citizenship education adopted in November 2019 was reinforced by a Ministerial declaration on citizenship in the digital age adopted at the Meeting of Education Ministers co-organised by the French Presidency of the Committee of Ministers and the Council of Europe on 26 November 2019. Building on this, the Committee of Ministers mandated the Steering Committee for Education Policy and Practice (CDPPE) to explore the implications of artificial intelligence and other emerging technologies for education generally and more specifically for their use in education.

79. Following the Committee of Ministers Recommendation CM/Rec(2018)7 on Guidelines to respect, protect and fulfil the rights of the child in the digital environment, the Steering Committee for the Rights of the Child has issued an implementation guide for policy makers, translating these principles into practice. The Guidelines provide an extensive range of measures that member states should include in their national legal framework and policies regarding AI and children's rights: if implemented, these will not only help children to thrive in this era of digital revolution, big data and machine learning systems, but also ensure accountability for and protection of their rights.

80. As a result of the Seminar on Artificial intelligence and its impact on Young People (4-6 December 2019), involving consultations with youth experts, several areas and proposals for further action by the Youth sector and its governmental and non-governmental partners have been identified. Improving institutional responses to emerging issues affecting young people's rights and their transition to adulthood, including artificial intelligence, with a focus on the

[29] Convention on Cybercrime, ETS No. 185 -

[30] A group of experts recommended to the CDPC the creation of an international legal framework and instrument to establish specific national legislation on automated driving - See meeting report of 10 July 2020

“pervasive influence of technologu and the digital space on the ways in which young people live their lives, are the focus of the Youth sector strategy 2030 launched in January 2020.

81. In December 2019, Eurimages published a study on the impact of predictive technologies and AI on access to a diversified cultural offer, in view of the feasibility of a new convention and a new fund to promote independent production in TV series. The new Council of Europe Youth sector strategy 2030, launched in January, includes a focus on the “pervasive influence of technology and the digital space on the ways in which young people live their lives”.

4.4. Non-binding frameworks for the supervision of AI applications

4.4.1. The Committee of Ministers

82. Since 1999, the Ministers' Deputies to the Council of Europe have appointed a Thematic Coordinator on Information Policy (TC-INF) from among their number, who has, among other things, monitored the implementation of the Council of Europe's Internet Governance Strategy 2016-2019,[31] in which the issue of dealing with the consequences of the development and use of AI has been included. The operational work of the different sectors of the Council of Europe was monitored in the framework of a transversal working group (*Internet Governance Task Force* - IGTF) in which a dedicated coordinator for AI, appointed by the Secretary General, sat. As a follow-up to these activities, work on a multi-annual strategy on digital governance is currently underway, with the objective to continue coordinating the production of instruments in the field.

83. With regard to the frameworks already produced, the Committee of Ministers adopted a Declaration on the Manipulation Capabilities of Algorithmic Processes[32] in February 2019 and a Recommendation on the Human Rights Impacts of Algorithmic Systems[33] in April 2020. The Declaration draws, inter alia, member States' attention "to the need to properly assess the need for stricter regulatory or other measures to ensure appropriate and democratically legitimate oversight of the design, development, deployment and use of algorithmic tools, with a view to implementing effective protection against unfair practices and abuses of economic power". The Recommendation, for its part, invites member States to "review their legislative frameworks and policies, as well as their own practices with regard to the ongoing acquisition, design, development and deployment of algorithmic systems to ensure that they are in line with the guidelines set out in the Appendix to this Recommendation".

4.4.2. The Parliamentary Assembly of the Council of Europe

84. The Parliamentary Assembly of the Council of Europe (PACE) adopted, on 28 April 2017, a Recommendation on "Technological convergence, artificial intelligence and human rights[34]". It sets out a number of working guidelines for the Committee of Ministers, inviting it in particular to adopt guidelines on the design and use of "artificial intelligence algorithms that fully respect the dignity and fundamental rights of all users". The Committee on Legal Affairs and Human Rights, meeting in Paris on 13 December 2018, decided for its part to set up a new sub-Committee on Artificial Intelligence and Human Rights. Several committees have begun work to identify the possible impact of AI applications in their respective fields. The report on the need for democratic governance of artificial intelligence, which will be presented by Mrs

[31] CM/Del/Dec(2016)1252/1.6

[32] Committee of Ministers, Declaration on the manipulation capabilities of algorithmic processes - Decl(13/02/2019)1, 13 February 2019

[33] Committee of Ministers, Recommendation to member states on the human rights impacts of algorithmic systems - CM/Rec(2020)1, 8 April 2020

[34] Recommendation 2102(2017)

Deborah Bergamini (Italy, EPP/CD) for adoption on 22 October 2020, proposes, in particular, that the Committee of Ministers support the drafting of a legally binding instrument governing AI applications, possibly in the form of a Convention[35].

4.4.3. The Commissioner for Human Rights

85. The Commissioner for Human Rights issued on 14 May 2019 a Recommendation Unboxing artificial intelligence: 10 measures to protect human rights[36]". It proposes a series of practical recommendations to national authorities in order to maximise the potential of AI systems while avoiding or mitigating their negative effects on people's lives and rights. The document focuses on 10 main areas for action: human rights impact assessment; public consultations; human rights standards in the private sector; information and transparency; independent monitoring; non-discrimination and equality; data protection and privacy; freedom of expression, freedom of assembly and association, and the right to work; avenues for redress; and promoting knowledge and understanding of AI.

4.4.4. The work of Committees, Commissions and expert groups

86. Since 2017, the Council of Europe has already implemented a large number of activities dealing with AI applications within its different sectors, In the current biennium (2020-2021) numerous activities of Council of Europe steering and ad hoc committees' terms of reference address artificial intelligence aspects within their scope of work.

87. First of all, a study produced by the Committee of experts on Internet Intermediaries (MSI-NET) under the authority of the Steering Committee on the Media and the Information Society (CDMSI) on the human rights' dimensions of automated data processing techniques and possible regulatory implications could be cited[37]. As a follow-up, the Committee of experts on Human Rights Dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT) on the "Human Rights Dimension of Automated Data Processing and Different Forms of Artificial Intelligence" produced, inter alia, a report on "Accountability and AI: Study on the Impact of Advanced Digital Technologies (including Artificial Intelligence) on the Concept of Accountability from a Human Rights Perspective[38]". The Committee of Experts on Freedom of Expression and Digital Technologies (MSI-DIG) is continuing its work in 2020 to prepare a standard-setting instrument (or other form of guidance from the Committee of Ministers to member States) on the impacts of digital technologies on freedom of expression.

88. The European Commission for the Efficiency of Justice (CEPEJ) adopted, at the beginning of December 2018, the very first European Ethical Charter for the use of artificial intelligence in judicial systems[39] by laying down 5 key principles (respect of fundamental rights, non-discrimination, quality and security, transparency, impartiality and fairness, "under the control" of the user). The CEPEJ prefigured an operational translation of its charter by proposing a self-evaluation checklist in its study in appendix. The Commission is currently studying the advisability and feasibility of a certification or labelling framework for artificial intelligence products used in judicial systems. The European Committee on Legal Cooperation (CDCJ) has

[35] Parliamentary Assembly of the Council of Europe, Political Affairs and Democracy Committee, Report on the need for democratic governance of artificial intelligence, Doc. 15150, 24 September 2020

[36] Commissioner for Human Rights, Recommendation "Unboxing AI: 10 steps to protect human rights", May 2019

[37] DGI(2017)12

[38] MSI-AUT, Accountability and AI: Study on the impact of advanced digital technologies (including artificial intelligence) on the notion of accountability, from a human rights perspective - DGI(2019)05, September 2019

[39] CEPEJ, European Ethical Charter on the use of artificial intelligence in judicial systems and their environment - CEPEJ(2018)14, December 2018

carried out a technical study on online dispute resolution[40] and is preparing draft guidelines to ensure the compatibility of these mechanisms with Articles 6 and 13 of the Convention on Human Rights.

89. The Venice Commission and the Directorate for the Information Society and Action against Crime of the Directorate General for Human Rights and the Rule of Law (DG1) have also published a report on digital technologies and elections[41].

90. The European Committee on Democracy and Governance (CDDG) is studying the impact of digital transformation – including AI – on democracy and governance, covering the interaction between such technologies and elections. This project is expected to finish by the end of 2020 and may result in the production of guidelines for a recommendation by the Committee of Ministers.

91. A learning course for equality bodies on the prevention and redress of AI-driven discrimination is currently under preparation. This follows a recommendation that human rights monitoring bodies should aim for better enforcement of non-discrimination norms in AI-related cases and a study on critical areas where AI can reinforce discrimination and inequality commissioned by the European Commission against Racism and Intolerance (ECRI) on "discrimination, artificial intelligence and algorithmic decision making[42].

92. In October 2018, the Division of Culture and Cultural Heritage organised a seminar on culture, creativity and artificial intelligence in the framework of the Croatian Chairmanship of the Committee of Ministers of the Council of Europe[43]. Eurimages published a study on the impact of predictive technologies and AI on the audio-visual sector, including possible specific measures to be put in place to guarantee freedom of expression and cultural diversity[44].A publication on the e-relevance of arts and culture in the age of artificial intelligence is due for release in 2020.

4.5. The case law of the European Court of Human Rights relating to the application of AI

93. The European Court of Human Rights has not yet developed case law specifically addressing the issue of AI applications. The case law produced on algorithms mainly concerns violations of Article 8 of the Convention (privacy) or Article 10 (freedom of expression) and, in a more residual way, Article 14 (discrimination - combined with other articles) on cases dealing with e.g. mass surveillance,[45]editorial responsibility of platforms[46][47]or electoral interference. As remedies in potential cases in member States may not yet have been exhausted, the CAHAI could not rely on more precise views of the ECtHR specifically on the technology which is the subject of its mandate. At the time of this feasibility study, however, a series of cases could be

[40] CDCJ, Technical Study on Online Dispute Resolution, 1 August 2018

[41] Venice Commission and DG1, Joint Report of the Venice Commission and the Directorate for the Information Society and Action against Crime of the Directorate General for Human Rights and the Rule of Law (DG1) on Digital Technologies and Elections - CDL-AD(2019)016, 24 June 2019

[42] ECRI, Study on Discrimination, Artificial Intelligence and Algorithmic Decision-Making, November 2018

[43] Division of Culture and Cultural Heritage, Conclusions of the Expert Seminar on Culture, Creativity and Artificial Intelligence, 12-13 October 2018 and see also the proposals for action arising from it.

[44] Eurimages, Study on the impact of predictive technologies and AI on the audiovisual sector, including possible specific measures to be put in place to ensure freedom of expression and cultural diversity, December 2019

[45] ECtHR, Big Brother Watch and others v. the United Kingdom, 13 September 2018 (Chamber judgment) - case referred to the Grand Chamber in February 2019

[46] ECtHR, Delfi AS v. Estonia, 16 June 2015 (Grand Chamber)

[47] ECtHR Court, Magyar Kétfarkú Kutya Párt v. Hungary, 23 January 2018 - case referred to the Grand Chamber in May 2018

cited in which algorithms based on statistical methods are used to process large amounts of information in order to establish evidence or prevent offences.

94. The first of the cases to be examined is *Sigurður Einarsson and others v. Iceland*[48], in which statistical data processing techniques were used by a prosecuting authority in an economic and financial case to process a very large quantity of documents. The question raised in this case concerned the availability to the defence of some or all of the data from which incriminating evidence was inferred. The applicants specifically argued that the defence had had no say in the electronic sorting of this data, which was carried out by the public prosecutor's office for the purpose of selecting the relevant information to be included in the investigation file. They argued that no one had controlled the selection by the public prosecutor's office of the documents to be submitted to the court and that they had been denied the opportunity to search using the electronic system employed ("Clearwell", an electronic evidence discovery system). It should be pointed out in this case that it was not disputed that the defence had been given access to the elements of the case and had been given the opportunity to consult the investigation file containing data not submitted to the national court. The question raised was whether the defence had been allowed to consult, on the one hand, the entire body of information collected on a non-selective basis by the public prosecutor's office and not included in the investigation file and, on the other hand, the "tagged" data obtained through the research carried out with Clearwell, in order to identify elements which could have been exculpatory. The Court unanimously concluded that there had been no violation of Article 6§1 of the European Convention on Human Rights, on the grounds that the defence had never sought access to the entire compilation or requested that further research be carried out. Furthermore, they never proposed any other investigative measures or proposed any other keywords capable of uncovering exculpatory material for their clients. The Court therefore considered that the lack of access to the data was not such as to deprive them of a generally fair trial.

95. Other decisions of the Court have dealt with the consequences of algorithmic mechanisms used to prevent the commission of infringements. Without going back as far as the *Klass v. Germany* judgment of 1978[49], the Court held in 2006 in the *Weber and Saravia v. Germany* judgment[50] that any abuse of the State's supervisory powers was subject to adequate and effective safeguards and that, in any event, Germany had a relatively wide margin of appreciation in the matter. This analysis is confirmed by a review of decisions in recent years, as the Court continues to grant States a wide margin of discretion in matters of national security, particularly in view of the current threats of international terrorism. The 2008 judgment in *Liberty and Others v. the United Kingdom*[51] appears to be an exception in sanctioning the law then in force on the interception of telephone and e-mail communications, in that it did not provide sufficient protection against abuse of power and the scope and modalities of the discretion enjoyed by the authorities were not clearly defined.

96. With regard to mass surveillance of the population using algorithms, potentially including AI tools, two cases could be considered, which have been referred to the Grand Chamber and where the last hearings took place on 10 July 2019: *Centrum För Rättvisa v. Sweden*[52] and *Big Brother Watch and others v. the United Kingdom*[53].

[48] ECtHR, Sigurður Einarsson and Others v. Iceland, 4 June 2019 (2nd section)
[49] ECtHR, dec. 6 September 1978, Klass and Others v. Germany, no. 5029/71
[50] ECtHR, Dec. 29 June 2006, Weber and Saravia v. Germany, no. 54934/00
[51] ECtHR, Dec. 1 July 2008, Liberty and Others v. the United Kingdom, no. 58243/00
[52] ECtHR, Dec. 19 June 2018, Centrum För Rättvisa v. Sweden, no. 35252/08 referred back to the Grand Chamber in February 2019 - hearing held on 10 July 2019
[53] ECtHR, Dec. 13 September 2018, Big Brother Watch and others v. the United Kingdom, nos. 58170/13, 62322/14 and 24960/15 referred back to the Grand Chamber in February 2019 - hearing held on 10 July 2019

97. The Swedish foundation Centrum För Rättvisa alleged that the legislation allowing mass interception of electronic signals on their territory for foreign intelligence purposes had infringed its right to privacy. In its Chamber judgment of 19 June 2018, the Court concluded, unanimously, that there had been no violation of Article 8 (right to respect for private life) of the Convention. The Chamber considered that, although certain aspects required improvement, the Swedish mass interception system offered adequate and sufficient safeguards against arbitrariness and the risk of abuse. In reaching this conclusion, the Chamber took into account the State's discretion to protect the national security, specifically in view of the current threats of international terrorism and cross-border criminality. In light of these findings, the Chamber found that there were no separate issues under Article 13 (right to an effective remedy) of the Convention and rejected any consideration of the Foundation's applications on this basis. The case is still pending before the Grand Chamber.

98. Big Brother Watch's claims were filed after Edward Snowden (former contract agent for the US National Security Agency) revealed the existence of surveillance and intelligence sharing programmes between the USA and the UK. The case concerns complaints by journalists and advocacy organisations about three surveillance regimes: mass interception of communications, intelligence sharing with foreign states, and obtaining communications' data from communication service providers. In the view of the Chamber, the principle of this mass surveillance system is not subject to criticism and confirms the wide margin of appreciation left to member States in this area, which would potentially allow for the inclusion of AI-based systems. Only the implementation modalities are sanctioned. In its judgment, the Chamber therefore develops an argument limited to finding that section 8(4) of the United Kingdom Regulation Investigatory Powers Act 2000 (RIPA) infringes Article 8 of the Convention, as it does not meet the requirement of "quality of the law" arising from the Convention. The Chamber also finds a second violation of Article 8 of the Convention by the lack of safeguards surrounding the system for acquiring data from communication service providers, such as making access subject to prior review by an independent court or administrative body. Finally, the Chamber considered that the system of mass interception of communications violates Article 10 of the Convention (freedom of expression), in so far as it applies to journalists. This system did not provide any specific guarantees to protect the confidentiality of journalists' information, which could be selected, voluntarily or involuntarily, for examination. The case is still pending before the Grand Chamber.

5. Mapping of instruments applicable to artificial intelligence

i. International legal instruments applicable to artificial intelligence

99. General international and regional human rights instruments, including the European Convention on Human Rights, are applicable in all areas of life and are therefore also applicable in the context of AI systems. The important question is whether these different instruments, separately or applied together, are able to sufficiently meet the challenges posed by AI systems and to ensure adherence to the Council of Europe's (CoE) standards on human rights, democracy and the rule of law in the design, development, deployment and use of AI systems. Currently, no international legal instrument exists that specifically applies to the challenges to democracy, human rights and the rule of law raised by AI systems – or by automated decision-making processes more generally – in a comprehensive way. There are, however, a number of international legal instruments that deal with certain aspects pertaining to AI systems indirectly.

100. In this regard, the CAHAI took note, during its 2nd plenary meeting, of the analysis of some relevant international legally binding instruments made by an independent consultant.[54] This analysis was based on a review of international legally binding and non-binding instruments in four core areas (data protection, health, democracy and justice) and was complemented by an overview of CoE instruments in other fields and their main underlying values. It noted that various international legal instruments already exist to safeguard human rights more generally[55], to safeguard the rights of specific groups in light of vulnerabilities that can also be relevant in an AI-context[56], and to safeguard specific human rights that can be impacted by AI. The latter encompass, for instance, the right to non-discrimination[57] and the right to the protection of privacy and personal data[58], in particular in the context of automated personal data processing.

101. Furthermore, in addition to horizontally applicable instruments, a number of international legal instruments deal with specific sectors or domains that may indirectly pertain to AI or automated decision-making processes. These instruments cover areas as diverse as cybercrime[59], biomedicine[60], and aviation.[61] Finally, some legal instruments concern procedural rights – such as transparency[62] and access to justice[63] – that might be helpful in monitoring and safeguarding the protection of the substantive rights mentioned above, or in addressing aspects relating to liability for harm caused by certain products.[64]

102. The CAHAI-PDG acknowledges that these different legal instruments may be relevant in the context of AI regulation. However, the CAHAI-PDG also supports the conclusions drawn in the analysis that these instruments are not tailored to the specific challenges posed by Artificial Intelligence and therefore they do not always provide adequate safeguards in this context. This will be the subject of further analysis under sub-section iv) below.

[54] See CAHAI (2020)08-fin, Analysis of internationally legally binding instruments, final report prepared by Alessandro Mantelero, Associate Professor of Private Law and Data Ethics & Data Protection, Polytechnic University of Turin.

[55] Such as, for instance, the European Convention on Human Rights and Fundamental Freedoms (ETS No. 5) and its Protocols; the European Social Charter (ETS No. 163); the International Covenant on Civil and Political Rights; the International Covenant on Economic, Social and Cultural Rights; and the Charter of Fundamental Rights of the European Union.

[56] See for instance, the Convention on the Rights of the Child and the Convention on the Rights of Persons with Disabilities. See also the European Charter for Regional or Minority Languages (ETS No. 148) which could for instance indirectly help safeguard sufficient attention to minority languages when developing (speech- or text-oriented) AI-applications.

[57] See for instance, 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 on Cybercrime and its Additional Protocol concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems.

[58] See for instance, the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (ETS No. 108) or the EU's General Data Protection Regulation.

[59] See for instance the Convention on Cybercrime (ETS No. 185).

[60] See for instance the Convention for the protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine (ETS No. 164).

[61] See for instance the Chicago Convention on International Civil Aviation.

[62] See for instance the Council of Europe Convention on Access to Official Documents (ETS No. 205).

[63] See for instance the European Convention on the Exercise of Children's Rights (ETS No. 160) and the European Convention on Mutual Assistance in Criminal Matters (ETS No. 30).

[64] See for instance the European Convention on Products Liability in regard to Personal Injury and Death (ETS No. 91) and the European Union's Product Liability Directive and Machinery Directive.

103. The growing need for a more comprehensive and tailored governance framework to address the new challenges and opportunities raised by AI has been acknowledged by a number of intergovernmental actors at the international level. To date, most of these initiatives are limited to non-binding recommendations.[65] It is worth mentioning that the European Commission has announced the preparation of a legislative proposal to tackle some of AI's ethical challenges, which is scheduled for publication in the first quarter of 2021.[66]

ii. Ethical Guidelines applicable to artificial intelligence

104. In recent years, private companies, academic and public-sector organizations have issued principles, guidelines and other soft law instruments for the ethical use of AI[67]. In this regard, the CAHAI took note, during its 2nd plenary meeting, of the mapping work by two independent consultants[68] who reviewed 116 documents on "ethical AI", primarily developed in Europe, North America and Asia. This mapping revealed that current AI ethics guidelines converge on some generic principles, but they sharply disagree over the details of what should be done in practice. Notably as regards transparency, the most frequently identified principle, it was not clear whether transparency should be achieved through publishing source code, the algorithmic training data or some other means. Hence, resolving semantic ambiguities and conflicting characterisations of this and other principles was considered an important issue to be addressed by policy makers.

105. The mapping showed that overall, a growing agreement has emerged around five ethical principles: transparency, justice, non-maleficence, responsibility, and privacy. These could be considered as priority areas of oversight and possible intervention at both the governmental and intergovernmental level. Compared to the rest of the world, soft law documents produced within CoE Member States appear to place greater emphasis on the ethical principles of solidarity, trust and trustworthiness, and refer more sporadically to the principles of beneficence and dignity. The principles of privacy, justice and fairness showed the least variation across CoE-Member States, CoE-observer countries and the rest of the world, hence the highest degree of cross- geographical and cross-cultural stability.

106. In terms of key policy implications, it was noted that ethical guidelines are useful tools to exert practical influence on public decision making over AI and steering the development of AI systems for social good. However, it was also underlined that soft law approaches should not be considered substitutes for mandatory governance. Due to conflict of interest, there is a particular risk that self-

[65] For instance, the OECD adopted a Council Recommendation on AI listing a number of ethical principles, which provided inspiration for the human-centered AI principles endorsed by G20 in a Ministerial Statement. See the OECD Recommendation of the Council on Artificial Intelligence, May 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 and the G20 Ministerial Statement on Trade and Digital Economy, June 2019, https://www.mofa.go.jp/files/000486596.pdf. Also UNESCO is preparing a (non-binding) Recommendation on ethical AI, of which a first draft is currently subject to consultation. See UNESCO, First Draft of the Recommendation on the Ethics of Artificial Intelligence, September 2020, https://unesdoc.unesco.org/ark:/48223/pf0000373434. It should also be noted that, while UNESCO's current draft does mention AI's impact on human rights and the rule of law, it does not focus on the challenges posed by AI on democracy or the democratic process.

[66] The European Commission particularly emphasizes risks for fundamental rights, for safety and for the effective functioning of the liability regime. See the European Commission's White Paper on Artificial Intelligence, published in February 2020, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[67] Amongst recent initiatives feature the Ethics Guidelines for Trustworthy AI published in April 2019 by the Independent High-Level Expert Group on Artificial Intelligence, set up by the European Commission, and its "Assessment List for Trustworthy AI" (ALTAI) for self-assessment published in July 2020.

[68] See CAHAI (2020)07-fin, AI Ethics Guidelines: European and Global Perspectives, report prepared by Marcello Ienca and Effy Vayena.

regulation by private AI actors will be promoted to bypass or avoid mandatory governance by governmental and intergovernmental authorities.

107. The CAHAI-PDG agrees with the general findings of the mapping study and that the most common principles identified in the mapping of ethical guidelines could be considered for inclusion in a future legal instrument on AI. Human rights, which were mentioned in only just over half of the soft law documents reviewed, should be the focus of any future legal instrument on AI based on the CoE's standards.

iii. Overview of national instruments, policies and strategies related to artificial intelligence

108. The analysis of the electronic consultation carried out among CAHAI members, observers and participants on this issue[69] has indicated that three Member and two Observer states have adopted specific legally binding frameworks on AI systems. In all three Member States, these legal frameworks concern the testing and using of autonomous cars. In two Member States, legal frameworks on the use of AI systems in the fields of recruitment and automated decision-making by public authorities are under development.

109. Domestic ethical charters and soft law documents such as reports, opinions, guidelines, resolutions and white papers appear to be more widespread and regard issues such as robotics, facial recognition, the use of "ethical AI" in the public service and in electoral processes, and the use of personal and non-personal data. In one Member State, a voluntary AI certification programme has been launched; two Member States have formally endorsed international or European non-binding ethical frameworks on AI. A total of eleven Member and four Observer States have adopted one or more of the above-mentioned instruments. Different types of institutions such as National Councils, Committees, public institutions specialised on AI and Government entities have been responsible for their development.

110. Strategies and policies on AI systems have been put in place in thirty Member and four Observer States. Built on multi-annual action plans, accompanied in some cases by ambitious funding programmes, they pursue the objectives of increasing the trust in this technology and promoting its uptake, strengthening skills for its design and development, supporting research and boosting business development. States have very often involved experts from the public and private sectors, as well as academia, in the making of these plans.

111. In most cases, AI systems are the subject of targeted strategies, whilst in other cases they have been integrated into broader sector policies concerning economy and digital technologies. The development and use of AI systems has also been considered in sectorial strategies concerning agriculture, e-justice, the efficiency of public services, improving health, environment, education, security and defence, mobility or data.

112. Finally, the need for promoting the development of AI in line with ethical requirements and international human rights standards has been underlined in seven national strategies.

iv. Advantages, disadvantages and limitations of existing international and national instruments and ethical guidelines on artificial intelligence

113. The overview of instruments has demonstrated that a number of more broadly applicable provisions already extend to the development and use of AI systems. Due to their broad character, and in the

[69] See the document CAHAI (2020) 09 rev 2, on the electronic consultation of CAHAI members, observers and participants, which includes replies until 30 September 2020.

absence of a specific legal response in terms of international legally binding instruments focused on AI, significant efforts have been put into interpreting existing legal frameworks in the light of AI, and/or in formulating non-binding rules to contextualise the principles provided by existing binding instruments.[70] However, the fact that existing legal instruments were adopted in a pre-AI era often tends to reduce their effectiveness in providing an adequate and specific response to the challenges brought by AI systems. For instance, a report of the CoE European Commission against Racism and Intolerance (ECRI) on "Discrimination, artificial intelligence, and algorithmic decision-making" has highlighted the limitations of existing international and domestic legal instruments in the field of non-discrimination.[71] The independent expert's analysis prepared for CAHAI regarding the impact of AI on human rights, democracy and the rule of law provided similar conclusions as regards other rights.[72]

114. Besides the fact that the existing instruments are not tailored to the specific challenges raised by AI systems, their number and diversity make it more difficult to interpret and apply them to the AI-context in a consistent and comprehensive manner. While certain soft-law instruments (including AI ethics guidelines) do set out more tailored principles regarding the development and use of AI systems, these are non-binding and hence limited in their effectiveness. Moreover, they do not clearly define any binding obligations for Member States and private actors with regards to the respect of human rights, democracy and the rule of law. Furthermore, ethical guidelines do not have the same universal dimension as human rights-based standards and are characterised by a variety of theoretical approaches[73], which limits their utility. The CAHAI-PDG therefore notes that while there is no legal vacuum as regards AI regulation, a number of substantive and procedural legal gaps nevertheless exist.

115. First, the rights and obligations formulated in existing legal instruments tend to be articulated too broadly or generally to secure their effective application to the challenges raised throughout the lifecycle of AI systems. It has been indicated that a translation or concretisation of existing human rights to the context of AI systems through more specific provisions could help remedy this issue.[74] For instance, the right to non-discrimination could be further concretised in terms of the right not to be subject to (intentional or unintentional) bias arising from the biased design, implementation or use of AI systems. The CAHAI-PDG believes that CoE standards on human rights, democracy and the rule of law could provide an adequate basis for the elaboration of more specific provisions to regulate AI.

[70] See for instance 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.

[71] These instruments do not apply if an AI system invents new classes which do not correlate with protected characteristics under such instruments (i.e. gender or race), to differentiate between people. Such differentiation can nevertheless be unfair: By way of illustration, AI-driven price discrimination could lead to certain groups in society consistently paying more.

[72] See CAHAI(2020)06-fin, The Impact of AI on Human Rights, Democracy and the Rule of Law, report prepared by Catelijne Muller.

[73] As pointed out in the independent expert's report CAHAI (2020)08-fin, cited above.

[74] See CAHAI(2020)06-fin and CAHAI (2020)08-fin, cited above. See also Raso, Filippo and Hilligoss, Hannah and Krishnamurthy, Vivek and Bavitz, Christopher and Kim, Levin Yerin (Harvard University), Artificial Intelligence & Human Rights: Opportunities & Risks, Berkman Klein Center Research Publication, 2018, http://dx.doi.org/10.2139/ssrn.3259344; Nathalie A. Smuha (KU Leuven), 'Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea', in Philosophy and Technology, 2020, https://doi.org/10.1007/s13347-020-00403-w; Karen Yeung, Andrew Howes, and Ganna Pogrebna (University of Birmingham), 'AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing', in The Oxford Handbook on Ethics of AI (eds. M. D. Dubber, F. Pasquale, and S. Das), 2020, DOI: 10.1093/oxfordhb/9780190067397.013.5.

116. Secondly, adequate legal protection is currently lacking for a number of principles that are relevant to protect human rights, democracy and the rule of law in the context of AI, even though there is agreement as to their importance. These gaps concern, for instance, the necessity to ensure *human control and oversight* over AI-applications and to secure the *transparency* and *explainability* of AI-systems and of the human decisions pertaining to the design- and use-choices of AI-systems, in particular when they are likely to produce legal or other significant effects on individuals. The inadequate protection of those principles in existing legal instruments was also pointed out in the European Commission's White Paper on AI.[75] Safeguarding these principles is often a necessary precondition for safeguarding substantive rights, given the opacity of AI-systems and the human choices made to design and use them.[76] The existence of asymmetries of information between those who may be negatively impacted by AI systems and those developing and using it, also stresses the need to reinforce mechanisms of *accountability* and *redress*, to render AI systems *traceable* and *auditable*, and to ensure their *robustness*. Furthermore, current instruments typically lack sufficient attention to the dimensions of *effectiveness* and *competence* in the AI governance dialogue, and to the *societal dimension* of AI's risks that surpasses the impact on individuals, such as the potential impact on the democratic decision-making process.

117. Besides inadequate safeguards, these gaps also lead to legal uncertainty for stakeholders, and in particular AI designers, developers, deployers and end-users, who lack a predictable and sound legal framework in which AI products and services can be designed, implemented and used. This uncertainty also risks hampering beneficial innovation, and can hence stand in the way of reaping the benefits provided by AI for citizens and society at large. A comprehensive legal framework for AI systems can help provide the contours in which beneficial innovation can be stimulated and enhanced, and AI's benefits can be optimised, while ensuring – as well as maximising – the protection of human rights, democracy and the rule of law via effective legal remedies and safeguards for the customers and citizens.

118. Based on the analysis of existing instruments, it can be concluded that a regulatory approach to AI systems must aim to address those gaps. Building on existing legal frameworks, such regulatory approach could contain concrete provisions to safeguard human rights, democracy and the rule of law in the context of AI more generally, regardless of the sector concerned, and provide guidance on the elaboration of certain sector-specific provisions that are only relevant in a given field or application in which AI is used and where certain principles must be contextualised.[77] These conclusions have been supported by the CAHAI-PDG, which has stressed that a co-regulatory approach – with a binding instrument establishing horizontal principles overarching all different sectors, to be combined with tailored rules set out in additional non-binding sectoral instruments – could provide the necessary flexibility and ability to adapt to an evolving context, while providing

[75] See the European Commission White Paper on AI, 19 February 2020, COM(2020) 65 final, at p 9: "*A key result of the feedback process is that, while a number of the requirements are already reflected in existing legal or regulatory regimes, those regarding transparency, traceability and human oversight are not specifically covered under current legislation in many economic sectors*".

[76] It is only when the traceability of AI is ensured, for instance through the documentation or logging of relevant information, that a system can be audited and that it can be verified to which extent it may for instance infringe the right to non-discrimination. Furthermore, the lack of explanation of the decision-making process hinders the possibility for individuals to challenge a decision and seek redress. In this regard, the European Commission White Paper on AI noted more generally, at p12, that "*the specific characteristics of many AI technologies, including opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour, may make it hard to verify compliance with, and may hamper the effective enforcement of, rules of existing EU law meant to protect fundamental rights*". This also applies to the human rights provisions in other existing legal instruments, given that they are currently not tailored to the specific challenges raised by AI.

[77] In this regard, the CAHAI-PDG recognized the context-specificity of certain risks. The wide-scale use of AI-based remote biometric identification, for instance, does not raise the same impact on human rights as the use of an AI-based system to recommend a song.

legal certainty. The Group underlined its expectations that this approach could provide for the necessary level of guidance to private actors developing self-regulatory frameworks.

119. As mentioned in the CAHAI progress report, the work undertaken by the CAHAI provides an opportunity to contribute and complement other international initiatives in this area (e.g. by the OECD, the European Union – in particular the European Commission, UNESCO and the United Nations in general, with whom coordination and synergies are being sought on a regular basis[78]) by enacting a concrete instrument based on the Council of Europe's standards on human rights, the rule of law and democracy, as part of a global legal mechanism for the regulation of digital technologies. In this regard, the CAHAI-PDG has underlined that part of the added value that the Council of Europe can provide when elaborating a legal instrument on AI is that, besides the protection of human rights, it can also address the societal and environmental challenges posed by AI to democracy and the rule of law.[79] Developing a legally binding instrument based on standards on human rights, the rule of law and democracy – should this option be supported by the CAHAI – would contribute to making the CAHAI initiative unique among other international initiatives, which either focus on the elaboration of a different type of instrument or have a different scope or background.

v. International legal instruments, ethical guidelines and private actors

120. CoE instruments are typically addressed to the Member States rather than to private actors. Nevertheless, private actors can be addressed indirectly, by virtue of the rights granted to states and obligations assumed by states under such instruments. To this end, States may have a duty to ensure that private actors act in line with certain legal provisions by implementing and enforcing these provisions in their national laws, and by making sure that effective legal remedies are available at national level. In turn, private actors, in line with the UN Guiding Principles on Business and Human Rights, have the corporate responsibility to respect the human rights of their customers and of all stakeholders, throughout all stages of the process.[80]

121. A number of international instruments directly focus on the need for businesses to comply with human rights standards and ensure responsible technological research and innovation.[81] Over the past years, private actors have shown a strong interest in advancing the responsible development and use of AI systems, acknowledging not only the opportunities but also the risks raised thereby. Private actors have not only contributed to the proliferation of AI ethics guidelines, but some have also explicitly argued in favour of a regulatory framework to enhance legal certainty in this domain.[82]

[78] During its second plenary meeting, the CAHAI heard updates from FRA, the European Union, the OECD, the United Nations High Level Panel on Digital Cooperation and UNESCO. See the report of the second plenary meeting of the CAHAI, paragraphs 78-84.

[79] It can be noted that, while the European Commission White Paper on AI focuses on the impact of AI on fundamental rights, it does not specifically address AI's impact on democracy and the rule of law.

[80] See Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic system), available at https://rm.coe.int/09000016809e1154. See also Recommendation CM/Rec(2016)3 of the Committee of Ministers to Member States on human rights and business, available at https://rm.coe.int/human-rights-and-business-recommendation-cm-rec-2016-3-of-the-committe/16806f2032.

[81] Most notably the UN Guiding Principles on Business and Human Rights, particularly Articles 18 and 19. See also the OECD Due Diligence Guidelines for Multinational Enterprises and the OECD Due Diligence Guidelines for Responsible Business Conduct.

[82] Besides the statements of individual companies, such as, for instance, Microsoft (https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/) or IBM (https://www.ibm.com/blogs/policy/ai-precision-regulation/), the Policy Recommendations of the European Commission's High-Level Expert Group on AI – which were drafted by 52 experts, including over 20 private companies, and which were specifically addressed to EU Member States – specifically call upon the consideration of adopting new legislation. The document states, for

122. Should a co-regulatory approach to AI systems be recommended, private actors, civil society organisations, academia and other stakeholders would have an important role not only in assisting states in the development of a binding legal framework, but also in contributing to the development of sectorial soft law instruments that concretely implement the binding provisions in a context-specific manner (for instance through sectorial guidelines, certifications and technical standards). An effective regulatory framework for AI systems will require close cooperation between all stakeholders, from states and public entities who must secure public oversight, private actors who can contribute their knowledge and secure socially beneficial AI innovation, and civil society organizations who can represent the interests of the public at large. The CAHAI-PDG acknowledges that the CoE is uniquely positioned to lead this effort and – by building further on existing frameworks – to guide the alignment of AI systems with its standards on human rights, democracy and the rule of law.

6. Main conclusions of the multi-stakeholder consultations

6.1 Feasibility Study Table of Contents

6.2 Main conclusions on the type and content of a legal framework for the design, development and application of artificial intelligence, based on Council of Europe 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

7.1 Key values, rights and principles deriving - in a bottom-up perspective - from sectoral approaches and ethical guidelines; in a top-down perspective - from the requirements of human rights, democracy and the rule of law.

123. In line with the CAHAI mandate, a CoE legal framework on AI should ensure that the development, design and application of this technology is based on the Council of Europe's standards on human rights, democracy and the rule of law. This can be done by: (a) formulating fundamental principles and fundamental rights that must be guaranteed in the context of AI, (b) translating these fundamental rights and principles key requirements that AI systems should meet, and into concrete rights that individuals should be able to invoke (see boxes under section 1.), and (c) set out some red lines for AI systems or applications that pose significant risks for human rights, democracy and the rule of law.

124. The aim of the legal framework should particularly consist in addressing the substantive and procedural legal gaps identified under Chapter 5 of this Feasibility Study, to ensure both its relevance and effectiveness amidst existing legal instruments. This framework of principles, requirements and rights that apply transversally could be combined with sectoral instruments that

instance, at p. 40 that: "*For AI-systems deployed by the private sector that have the potential to have a significant impact on human lives, for example by interfering with an individual's fundamental rights at any stage of the AI system's life cycle and for safety-critical applications, consider the need to introduce: a mandatory obligation to conduct a trustworthy AI assessment (including a fundamental rights impact assessment which also covers for example the rights of children, the rights of individuals in relation to the state, and the rights of persons with disabilities) and stakeholder consultation including consultation with relevant authorities; traceability, auditability and ex-ante oversight requirements; and an obligation to ensure appropriate by default and by design procedures to enable effective and immediate redress in case of mistakes, harms and/or other rights infringement".* The document also stresses the need to enhance legal certainty.

provide context-specific guidance to AI-practitioners, for instance through sectoral guidelines or concrete and non-exhaustive AI assessment lists to operationalise the requirements.

1. **Fundamental principles and human rights in the context of AI to be included into the CoE legal framework**[83]

a) Freedom and Human Autonomy[84]

125. Freedom and autonomy are core human values which are reflected in various human rights laid down in the ECHR, and have a strong connection with human integrity and dignity. In the context of AI, these values refer to the ability of every human being to act truly self-determinedly, by autonomously deciding on the effects that AI systems can have on themselves, others and society, in a deliberative way. More generally, individuals should be able to make informed autonomous decisions regarding AI systems. AI systems should not subordinate, coerce, deceive, manipulate, condition or hurt humans, but rather augment, complement and empower human cognitive, social and cultural skills.

126. Human freedom and autonomy can be impacted by different AI-applications. AI-driven (mass) surveillance, for example through the use of facial recognition technology[85], enables situations whereby individuals are constantly watched, followed and identified. This can cause psychological 'chilling' effect that might drive individuals to adapt their behaviour to certain (perceived) norms. More subtly, the indiscriminate on- and offline tracking of all aspects of our lives through by collecting data on our online behaviour, our location, our IoT applications (such as smart watches, health trackers, smart speakers, smart homes, smart cars, etc.) could have the same impact. Other forms of AI-driven biometric recognition can have an even greater impact on our integrity and autonomy. The use of AI systems to identify and analyse micro-expressions, gait patterns, voice tones, heart rates and other types of biometric data can be used to assess or even predict our behaviour, mental states and emotions. While the inferences made by these systems can be spurious or pseudoscientific, the (ab)use thereof can nevertheless seriously impact individuals' lives.

Key substantive rights and obligations:

- ❖ Key substantive rights:
  - o The right to autonomy, agency and human oversight over AI systems;
  - o The individual's right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them.
  - o The right to challenge a decision and to seek redress.

- ❖ Key obligations:
  - o Requirements regarding human oversight[86]:
    - ▪ An appropriate level of involvement of human beings in the operation of AI systems must be ensured.

[83] The principles and rights mentioned under this chapter are not stated in any specific order.

[84] Reason for combining i. and vi: This was suggested in the majority of comments we received on this issue.

[85] See also the report of the Fundamental Rights Agency of the European Union, Facial recognition technology: fundamental rights considerations in the context of law enforcement, November 2019, accessible at: https://fra.europa.eu/en/publication/2019/facial-recognition-technology-fundamental-rights-considerations-context-law.

[86] Care must be taken that to ensure that the 'human in the loop' does not become a moral or legal 'crumple zone', which can be used to "describe how responsibility for an action may be misattributed to a human actor who had limited control over the behaviour of an automated or autonomous system. Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become simply a com-ponent—accidentally or intentionally—that bears the brunt of the moral and legal responsi-bilities when the overall system malfunctions

- In some instances, previous validation of a particular output by a human being will be necessary, while in other instances constant monitoring or control of the general functioning of the AI system can be sufficient.
- It must be ensured, that a human is able to disable the AI system or change its functionality if necessary at all times.

- It must be made sure that humans can challenge the results of the AI system. This means the possibilities of intervention in usage processes have to be explicitly mapped.
- Further detailed requirements for human oversight of AI systems can either be developed on a horizontal or sectoral level as appropriate.
- Red lines should be drawn for certain AI systems or uses that are considered to be too impactful with regard to human freedom and autonomy (see below under "red lines").

b) Non-Discrimination and Diversity

127. AI systems can have a negative impact on the right to non-discrimination and the right to equal treatment. Various studies[87] have pointed to the fact that the use of these systems can perpetuate and amplify discriminatory or unjustifiable biases, which has an adverse impact not only on the individuals subjected to the technology, but on society as a whole.[88] Indeed, reliance on biased AI systems could increase inequality and segregation, thereby threatening the necessary level of economic and social equality required for a thriving democracy. The risk of discrimination can arise in multiple ways, for instance due to biased training data (e.g. when the data-set is not sufficiently representative or inaccurate), due to a biased design of the algorithm or its optimisation function, due to the exposure of a biased learning environment, or due to the biased use of the AI system by its human users.

128. While the right to non-discrimination is already set forth in numerous international legal instruments[89], Chapter 5 has shown that it needs to be contextualised to the specific challenges raised by AI. Particular attention must be given to individuals and groups with an increased risk of being disproportionately impacted by AI, such as women, children, older people, economically disadvantaged persons, members of the LGBTI community, persons with disabilities, and "racial", ethnic or religious groups. The active participation of and meaningful consultation with a diverse community that includes effective representation from these groups in all stages of the AI-lifecycle can help prevent and mitigate adverse human rights impacts more generally, including in the context of the prohibition to discriminate. In addition, it is important to duly consider the risk of intersectional discrimination arising from the use of AI systems.[90]

[87] See, inter alia, F. Zuiderveen Borgesius, Discrimination, artificial intelligence, and algorithmic decision-making, Study undertaken for the Council of Europe, 2018, accessible at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73; Joy Buolamwini, Timnit Gebru; Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018; Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, et al. AI Now 2019 Report. New York: AI Now Institute, 2019. https://ainowinstitute.org/AI_Now_2019_Report.pdf.

[88] See CAHAI-PDG 1st meeting report, p.5.

[89] See also Chapter 5 above.

[90] Intersectional discrimination – which is rarely addressed in the discussion on discriminatory AI – takes place on the basis of several personal grounds or characteristics/identities that operate and interact with each other at the same time in such a way as to be inseparable. Current AI systems are particularly susceptible to such discrimination as they merely look for correlations between different features taken from the training data. A CoE-legal framework should take a special interest in this issue, as intersectional discriminations are rarely covered by national discrimination law which tends to focus on one ground of discrimination at a time.

Key substantive rights and obligations:

- ❖ Key substantive right:
  - A right not to be subject to bias arising from the biased design, development, implementation and use of AI, whether used in the public or private sector

- ❖ Key obligations:
  - Member States should, if necessary, impose binding requirements on the representativeness/balance of data sets[91], to effectively counter the discriminatory potential of AI systems.
  - Consistent requirements for test and evaluation data should also be considered. Depending on the classification of the AI system, this may also include quality parameters and requirements for training, testing and evaluation data, so that appropriate AI systems can be developed from quantitatively sufficient and high-quality data sets.
  - Member states must refrain from using AI systems that discriminate or lead to discriminatory outcomes
  - Member States must protect individuals from the consequences of use of such AI systems by third parties
  - Transparency of AI systems must be ensured in order to determine or detect any risk of discrimination or biases that occur throughout the AI system's lifecycle.
  - The principle of non-discrimination should be particularly taken into account in the use of AI by public authorities. Member states should apply the highest level of scrutiny when using AI systems in the context of law enforcement, especially when engaging in methods such as predictive or preventive policing. Such systems need to be independently audited prior to deployment for any discriminatory effect that could indicate de facto profiling of specific groups. If any such effects are detected, the system cannot be used.
  - The CoE legal instrument should incorporate an intersectional perspective in the legal framework to avoid intersectional discrimination. By using current AI systems which only look at correlations in the data, all kinds of unacceptable biases can easily surface. The problem with these systems is that, even if they would excel at identifying patterns, the system has no understanding of the meaning of the patterns. It will only be able to provide a label to a specific pattern. A legal framework has to deal with this restriction. An example of this would be mandatory use of intersectional training datasets, mandatory creation of intersectional benchmarks, and introduction of intersectional audits for all machine learning systems.

c) Principle of Transparency and Explainability of AI systems

129. At present, it is often not clearly recognisable whether a product or service uses an AI system in the first place, and if so, according to which criteria the system operates. Therefore, it is difficult or often impossible to know whether there the system impacts human rights, democracy and the rule of law in the first place. Transparency is hence crucial to ensure the enforcement of other principles or substantive rights., such as the right to challenge a decision or to seek redress.

130. Therefore, processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions must be explained to those directly and indirectly affected. To

[91] The basis for assessing whether these requirements have been taken into account can also be the results produced by the respective AI system, which means that access to the training, test and evaluation data set as such is not always necessary. This requires, however, that suitable procedures be available for the operational scenario to review results in terms of representativeness and balance.

strengthen trust and confidence, the manner in which AI systems work must be as transparent and traceable as possible. This can be either in the form of public disclosure of information on the system in question, the training data used, its processes, the human/organisation's decisions relating to the manner in which the system is used, its direct and indirect effects on human rights, and measures taken to identify and mitigate adverse human rights impacts of the system, or in the form of an independent, comprehensive, and effective audit. In all cases, the information made available should allow for meaningful assessment of the AI system. Without such information, a decision cannot be duly contested. Those who are affected by a decision that is solely or significantly informed by an AI system should be notified and be promptly provided with the aforementioned information, especially when this decision is made by a public authority. The use of an AI system must not only be made public in clear and accessible terms, individuals must also be able to understand how decisions are reached and how those decisions have been verified.

131. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and necessitate special transparency requirements. Indeed, in those instances, the system's auditability (through documentation or logging obligations regarding the training and testing processes for instance) must be ensured. While, a high standard on explainability could lead to a reduced performance and accuracy of the AI system, systems that cannot be subjected to appropriate standards of transparency and accountability should not be deployed in situations where they can negatively impact individuals and society.

132. While business secrets and intellectual property rights should be respected, AI systems should be auditable by public authorities in order to verify compliance with existing legislation. Moreover, the use of AI systems in public services should be held to higher standards of transparency. Public authorities should not acquire AI systems from third parties unwilling to waive restrictions on information (e.g. confidentiality or trade secrets) where such restrictions impede or frustrate the process of (i) carrying out human rights impacts assessments (HRIAs) (including carrying out external research/review), and (ii) making HRIAs available to the public.

- ❖ Key substantive right:

  A right to an explanation of how the AI functions, what logic it follows, and how its use affects the interests of the individual concerned

- ❖ Key obligations to ensure transparency and explainability:

  Key requirements to ensure transparency can be divided into different groups, depending on the specific aim of the requirement. These groups are (a) communication, (a) traceability including information rights and (c) documentation. In circumstances of 'black box' algorithms, specific transparency measures may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.

  - o Communication
    Above and beyond the labels and markings already required under data protection law, additional obligations to state that people are interacting with an AI system and not with a human being are necessary. In particular, but not limited to, if an AI system is used for interaction with individuals in the context of public services, especially justice, welfare, and healthcare, the user needs to be notified and the possibility of recourse to a professional upon request and without delay must be communicated.

Those who have had a decision made about them by a public authority that is solely or significantly informed by the output of an AI system should be notified and be promptly provided with the aforementioned information.

- Traceability including information rights
  It is important to include a requirement that certain information on AI systems is made available to those with a legitimate interest - consumers, citizens, operators of AI systems and supervisory authorities, amongst others. At the same time, actors obligated to provide information should not be burdened with unreasonably high compliance costs and trade and business secrets should be protected without creating undue barriers to due process. For entitled parties, it must be ensured that information is comprehensible and accessible at low thresholds. Such information could include (but should not be limited to) the types of decisions or situations subject to automated processing, criteria relevant to a decision, information on the training data sets used, description of the types of data and information on the collection of the data, description of the methodology, description of potential legal or other significant effects or consequences of the result produced by the AI system.

- Documentation
  Furthermore, the legal framework on AI should include requirements with regard to documentation, including recording and storage of data. This can make an important contribution to the transparency of AI systems and can also facilitate effective monitoring and enforcement by the supervisory authorities in charge. Access to, and use of, stored data sets must be linked to formally verifiable legal requirements. Furthermore, organisational and technical measures should be taken to ensure that only authorized persons have access to training data, algorithmic models, protocols and possible evaluations. Requirements should be set for the type of documentation.

d) Prevention of harm

133. AI systems can impact and potentially harm individuals, societies as well as the environment. The prevention of harm is a fundamental principle that should be upheld, in both the individual and collective dimension. Accordingly, human dignity as well as mental and physical integrity must be protected, with adequate safeguards for persons and groups who are more vulnerable. Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. Preventing harm also entails consideration of the natural environment and all living beings, and the manner in which the AI systems can have an adverse impact thereon.

134. Member states should therefore ensure that adequate safeguards are put in place to minimise and prevent harm stemming from the development and use of AI, whether this concerns physical, psychological, economic, social or legal harm. Attention must therefore be given to the safety and security of AI systems, including safeguards for their technical robustness, reliability, and measures that prevent the risk of adversarial attacks or malicious uses.

Key substantive rights and obligations:

❖ Key substantive right:

The right to physical, psychological and moral integrity in the light of AI

- ❖ Key obligations:
  - o Requirements regarding safety, security and robustness
  - o Specific requirements to ensure resilience to attack and security (e.g. security-by-design), general safety, accuracy and reliability should be considered. Related technical standards could be approved, possibly incorporating appropriate certification mechanisms.
  - o Member States have to ensure that AI systems are accountable so that individual human responsibility can be attributed in the case of harm or other adverse effects. Clear distribution of liability of the relevant actors involved is one important aspect to look at in realization of accountability. Accountability is a means to achieve the realization of all the above mentioned principles and requirements.

e) Data protection and Privacy

135. Privacy forms part of the right to private life and is fundamental to the enjoyment of other human rights. Thus the development, training, testing and use of AI systems that rely on the processing of personal data must fully secure a person's right to respect for private and family life under Article 8 of the European Convention on Human Rights, including the "right to a form of informational self-determination" in relation to their data.

136. Member states should effectively implement the modernised Council of Europe Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data ("Convention 108+") as well as any other international instrument on data protection and privacy that is binding on the member state.

137. Not all AI systems process personal data. But even where AI systems are not designed to process personal data, instead relying on anonymised, anonymous, or non-personal data, the line between personal data and non-personal data is increasingly becoming blurred. Thus, the interplay between personal and non-personal data must be further examined. Machine learning systems in particular can infer sensitive personal information about individuals from anonymised or anonymous data, or even from data about other people. In this regard, special consideration must be given to protecting people against inferred personal data.[92]

Key substantive rights and obligations:

- ❖ Key substantive right: A right to private life

- ❖ Key obligations:
  - o The processing of personal data at any stage of an AI system lifecycle must be based on the principles set out under the Convention 108+
  - o In particular (i) there must be a legitimate basis laid down by law for the processing of the personal data at the relevant stages of the AI system's lifecycle; (ii) the personal data must be processed lawfully, fairly and in a transparent manner; (iii) the personal data must be collected for explicit, specified and legitimate purposes and not processed in a way incompatible with those purposes; (iv) the personal data

[92] See, for instance, S. Wachter and B. Mittelstadt, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, Columbia Business Law Review, 2019(2).

must be adequate, relevant and not excessive in relation to the purposes for which they are pro-cessed; (v) the personal data must be accurate and, where necessary, kept up to date; (vi) the personal data should be preserved in a form which permits identification of data subjects for no longer than is necessary for the purposes for which those data are processed.

- Member states should also introduce a legislative framework that provides appropriate safeguards where AI systems rely on the processing of genetic data; personal data relating to offences, criminal proceedings and convictions, and related security measures; biometric data; personal data relating to "racial" or ethnic origin, political opinions, trade-union membership, religious or other beliefs, health or sexual life. Such safeguards must also provide protection against this data being processed in a discriminatory or biased way. Consideration must also be given to how to protect data subjects from personal data being inferred about them from anonymised, anonymous, or non-personal data.

f) Democracy

138. AI does not have an impact only on individuals but on society and democratic processes as a whole. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. Well-functioning democracies require a well-informed citizenry, an open social and political discourse and the absence of opaque voter influence. Yet in modern information societies, citizens are only able 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).

139. This technologically-enabled informational infrastructure of society could amplify hyper-partisan content one is likely to agree with and provide an unprecedented powerful tool for individualised influence, as well as the rapid spread of targeted disinformation. As a consequence, these systems can undermine the shared understanding, mutual respect and social cohesion required for democracy to thrive. In some instances, personalised AI-predictions can also undermine the agency and autonomy required for meaningful decisions by voters. AI systems can be embedded with with manipulative capacities so as to covertly impact human behaviour and thereby affect political and judicial decision-making process.

140. Effective, transparent and inclusive democratic oversight mechanisms are hence needed to ensure that the democratic decision-making process – and the correlated values of pluralism, access to information, and autonomy – are safeguarded. Member States should ensure a meaningful participatory approach, with special attention to the inclusion of vulnerable individuals and groups, which is key to ensure trust in the technology and its acceptance by all stakeholders. Consultations should provide an opportunity for all stakeholders, including state actors, private sector representatives, academia, civil society organisations and the media, to provide input on the role of AI in society. In any consultation process, the option should always be present of abandoning or not implementing a system if it is found to be incompatible with the protection of people's rights or when it is unsuitable to achieve its intended aim – guided by the principles of necessity and proportionality. Moreover, the timely and prior publication of all relevant information on the AI system that facilitates a proper understanding of its operation, function, and potential or measured impacts, should be enabled.

Key substantive rights and obligations:

[To be elaborated]

g) Rule of Law

141. AI can increase the efficiency of institutions, but it can also erode the procedural legitimacy of – and trust in – democratic institutions and the authority of the law.

142. The challenges arising from the use of AI in judicial systems were thoroughly examined by the European Commission for Justice (CEPEJ) of the Council of Europe, which in 2018 adopted the “European Ethical Charter on the use of AI in the judicial systems and their environment”, setting out 5 principles which should guide AI deployment in this field. The CEPEJ analysis regarded AI use in judicial adjudication processes and in online, non-judicial dispute resolution mechanisms. It was noted that while automated online dispute mechanisms provided by private companies can have the potential to enable consumers to act on their rights, when enforcing a claim in court is not feasible or too costly[93], they are governed by the terms of service and are not always fully aligned with the requirements set forth by law, substantively and procedurally. 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. The challenge is therefore to establish an AI governance framework which allows AI companies to act responsibly and in compliance with relevant legal requirements, while allowing for proper remedies and intervention by state authorities when this does not happen.

143. By using AI law enforcement and public administrations could become more efficient, yet at the cost of being more opaque and involving less human agency, autonomy and oversight. 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. In this context the tendency to outsource service delivery to the private sector must also be accounted for, as this raises questions of accountability, independent oversight and public scrutiny, which might be amplified by the use of opaque AI systems with little recourse to independent oversight.

144. 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, potentially including court orders and ombudspersons for complaints.

145. Applied in the AI context, remedies should not only aim at ensuring individual’s victim redress, but also collective redress and thus ‘affirm, reinforce, and reify the fundamental values of society’. Thus addressing harm or injuries arising from the design, development and application of AI should not depend solely on the initiative and capacity of individuals to access judicial remedies. Effective judicial remedies, such as class actions for AI should be considered.

146. 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

[93] Contribution of Israel

applications. Proper remedies should include cessation of unlawful conduct and guarantees of non-repetition. The obligation to repair the injury or damage caused by the violation, either to an individual or to a community, should exist.

147. Furthermore, member states should provide access to an effective remedy to those who suspect that they have been subjected to a measure that has been solely or significantly informed by the output of an AI system in a non-transparent manner and without their knowledge. Effective remedies should involve prompt and adequate reparation and redress for any harm suffered by the development, deployment or use of AI systems, and may include measures under civil, administrative, or, where appropriate, criminal law.

148. Member states must ensure that individuals have access to information in the possession of a defendant or a third party that is relevant to substantiating their claim that they are the victim of a human rights violation caused by an AI system, including, where relevant, training and testing data, information on how the AI system was used, meaningful and understandable information on how the AI system reached a recommendation, decision or prediction, and details of how the AI system's outputs were interpreted and acted on.

149. When national authorities consider challenges to human rights violations caused by the development, deployment or use of AI systems, they must show appropriate scepticism towards the "allure of objectivity" presented by AI systems and ensure that individuals challenging human rights abuses are not held to a higher standard of evidence compared to those responsible for the measure being challenged.

Key substantive rights and obligations:

[To be elaborated]

2. Red lines

150. Red lines should 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. I.a. the following AI-applications could give rise to the necessity of a ban, moratorium and/or strong restrictions or conditions for exceptional and/or controlled use:

- Facial recognition and other forms of remote 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
- DeepFake and DeepEmotion AI systems.
- Human-AI interfaces
- AI agents in charge of lethal force (be it military or non-military force).
- Autonomous AI agents in charge of distributing vital resources and allocating scarcity and priorities (health, food, water, air, land, energy, internet, mobility)
- Massive financial automated trade agents.

151. Exceptional use of such technologies, such as for national security purposes or medical treatment or diagnosis, should be specifically foreseen by law, evidence based, necessary and proportionate and only be allowed in controlled environments and (if applicable) for limited periods of time. A general prohibition to harm a human being or covertly manipulate his or her behaviour should be installed.

7.2 Role and responsibilities of member States and private actors in the development of applications complying with these requirements

152. According to international law, the CoE member states are responsible for ensuring that private actors respect human rights standards. The legal responsibility of private actors usually depends on the national legal frameworks in the respective member states.

153. Using AI systems that are capable of determining and perhaps altering our social interaction and democratic discourse, the impact of large private companies on human rights becomes more prevalent. Access to justice might be challenged when many AI-applications are developed and deployed by only a handful of large private actors and these companies dominate both the development of AI as well as the (eco)systems AI operates in and on. In this respect, experts suggest to 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[94].

154. It is important that national authorities carry out an evidence-based assessment of domestic legislation to check its compliance with human rights and prepare new legislation if gaps should be detected; they are equally responsible for establishing control mechanisms – also a duty of the private sector - and set effective judicial remedies[95]. Oversight bodies at national level shall have the powers to audit and assess the functioning of algorithmic systems, if needed. Such oversight powers could complement the existing obligations arising from European data protection law (accountability principle, impact assessment, prior consultation with supervisory authorities, etc) in an attempt to increase transparency. The auditing procedure itself could be strictly confidential in order to respect trade secrets or other conflicting commercial rights, as well as minimise security threats.

155. Specific procedures for certification and risks assessments are to be defined at the national level, which would fall within the remit of the authority in charge for the application of the convention. National legislation should regulate the questions of safety, security, privacy, responsibility and insurance, as well as issues related to creation, certification and use of databases.

7.3 Liability for damage caused by artificial intelligence

156. The development and use of AI raises new challenges in terms of product liability and safety.[96] Views differ however as to whether existing liability regimes should apply, or whether specific regimes should be developed for the context of AI.[97] Whether or not the question of liability will be set in a future legal framework at the CoE-level, the following principles should be pursued:

[94] C. Muller, p. 16; 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.

[95] Contribution Bulgaria

[96] Contribution Germany

[97] the Netherlands in favour („At this point in time the Netherlands is not in favour of the inclusion of liability for damage caused by artificial intelligence as a separate element. Only if research shows that the current liability laws are inadequate, should separate liability regulation for AI be considered.“); Mexico, CCBE, AI Transparency Institute in favour of specific liability regulations for AI.

- A proper and balanced liability regime in each sector, is important for both consumers and manufacturers[98].
- It is essential to guarantee the same level of protection to actors harmed by AI as traditional technologies[99].
- The liability should comprise the whole life cycle of the AI-system
- There should be a clear allocation of liability between actors involved in the development and operation of AI (creators, developers, deployers, operators, utilizers and users[100]), as well as certification bodies[101].
- It is important to regulate the issue of trans-border responsibility, such as, for instance, when the company using an AI is registered in one State, the developer of that AI in another State, and the user who suffered from the actions of the AI is a national of a third State[102].
- The rules for liability may consist of "setting the core principles and of establishing obligations for public authorities to create detailed mechanisms for remedying damages"[103].

8. Possible options for a Council of Europe legal framework for the design, development and application of artificial intelligence based on human rights, democracy and the rule of law

For each option: content, addressees, added value, role of private actors, member States' expectations arising from the written comments submitted

Preliminary remarks

157. Before outlining the possible options for a legal framework, it should be stressed that any framework to be adopted should be as technology-neutral as possible and take into account the need for cross national and cross domain harmonization, as well as cultural diversity and national legal system pluralism. Coherence between any general/horizontal instrument to be adopted and the existing or upcoming sectoral instruments produced in the various CoE pillars has to be ensured, which explains why this chapter also addresses possible complementarity mechanisms. Due attention should be paid to avoiding major inconsistencies with initiatives at the EU, OECD or UN(ESCO) levels.

158. The previous chapters have highlighted gaps in existing international legal instruments and the need to base the regulatory approach to AI on a legally binding instrument for which the existing framework based on human rights, democracy and the rule of law can, or even should, provide an appropriate and common context. Indeed, as pointed out by the analysis of the international legally binding and non-binding instruments commissioned by CAHAI, "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" (CAHAI (2020)08-fin, page 3). Metavalues as human dignity, solidarity and living in harmony and peace, which are roots and aims of the ECHR's guaranteed human rights and freedoms are underrepresented in soft law and ethical codes (CAHAI (2020)07-fin). Currently, competition on technical standards and ethical frameworks for AI takes place on a global level. However, self-regulation cannot effectively protect and further human rights, democracy and rule of law, and does not suffice to properly and adequately govern the huge impact AI may have on e.g. human autonomy, human integrity and human oversight.

[98] Israel

[99] Contribution Germany

[100] Contribution Poland

[101] Contribution of the Russian Federation

[102] Contribution of the Russian Federation

[103] Contribution of Bulgaria

159. This motivates the need for establishing an adequate, international legal framework to address AI's challenges and benefits for human rights, democracy and rule of law, that creates the right incentives to stimulate responsible and human-centric AI innovations, but is able to block irresponsible innovation in a timely manner. Harmonization of national legal frameworks for AI development and operation on the basis of an international standard will ensure safety, security, protection of human rights, while facilitating technological progress, mutually beneficial trade and scientific exchange in a rapidly developing and innovative area.

160. The legal framework should contain both common principles and operational solutions, as well as more specific and sectoral provisions (that contextualize the common principles). In accordance with CAHAI(2020)08-fin's main findings and conclusions on possible options, a number of guiding principles can be extracted from existing binding and non-binding international and regional instruments. These principles should be contextualized and clarified with regard to the specific AI environment, and form the basis for an initial set of provisions for future AI regulation focusing on the most challenging issues. In particular, the legal instrument should introduce specific limitations to AI when developed or used in a way that is not consistent with respect for human dignity, human rights, democracy and the rule of law. This should include the possibility for banning certain applications of AI that violate human rights in a way which cannot be mitigated. At the same time, the legal instrument should be carefully crafted so as to allow for socially beneficial innovation and research, while effectively blocking irresponsible and harmful implementation and deployment of innovation and research, that was not expectable after sandboxing test. As highlighted in Chapter 3, AI may serve humankind in unprecedented ways, from solving the problem of traffic accidents over providing people with improved healthcare, to tackling climate change, but it can equally have a negative impact on these same domains if unconstrained.

i. Updating of existing legally binding instruments

161. A first scenario to be considered is the need and desirability of updating currently available legal (binding) instruments. Those might require adaptation to the particularities of AI systems that should take into account its specific characteristics and the sectors in which it is to be applied.

162. An Additional Protocol to the European Convention on Human Rights could be adopted to enshrine new or adapted human rights in relation to AI, as suggested a.o. by PACE (Rec.2101(2017)) and by experts (CAHAI(2020)06-fin). The following list is based on their suggestions, but is evidently open for discussion: 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; a separate right to physical, psychological and moral integrity in light of AI-profiling and affect recognition; a right to refuse to be subjected to profiling, to have one's location tracked, to be manipulated or influenced by a "coach"; a strengthened right to privacy to protect against AI-driven mass surveillance; the right to have the opportunity, in the context of care and assistance provided to elderly people and people with disabilities, to choose to have contact with a human being rather than a robot. It is not unlikely that, under the dynamic and evolutive interpretation adopted by the ECtHR, existing Convention rights, such as the right to private life (Article 8), freedom of thought (Article 9) and of expression (Article 10), and right to non-discrimination (Article 14, Protocol 12), can be interpreted so as to include the aforementioned rights. The advantage, however, of an Additional Protocol is that the recognition of certain rights in relation to AI does not depend on a ruling by the ECtHR, and hence, offers more clarity and legal certainty (also avoiding possible criticism on the ECtHR for interpreting Convention rights too expansively).

163. The AI framework could be inserted in a Protocol of the Convention 108+, which is an international legally binding instrument. The added value would be to benefit from an existing effective instrument at international level, as well as an existing framework of data protection

independent authorities, whose scope of regulatory activities could be expanded to artificial intelligence. Drawback of this approach, however, is that it may not capture all concerns in relation to AI, given the specific focus of Convention 108+ on the protection of individuals, and the processing of personal data. AI's enablers lie specifically in larger aggregation of data, including non-personal data, and interoperability standards, talents, and access to infrastructure and investments.

164. The following remarks apply to all international treaties drafted under the eagis of the Council of Europe, whether they are under the form of protocols to existing conventions, or new (framework) conventions (cf. 8.ii and 8.iii):

- The addressees of the treaties are member States, but signature can be opened to other states. Private actors could sign a legally binding agreement with the Council of Europe to report on an annual basis all technical and organisational measures taken to design, develop and apply an artificial intelligence based on human rights, rule of law and democracy.

- The strength of treaties lies in their formality and the fact that they are legally binding on those states which have accepted them. States becoming parties to a convention incur legal obligations which are enforceable under international law.

- Whereas no general conclusion can be drawn as to the speed of preparation and entry into force of a treaty (periods may range from a couple of months – like in the case of the European Convention on Spectator Violence and Misbehaviour at Sport Events and in particular at Football Matches (ETS No. 120, 1985) or the Additional Protocol to the Oviedo Convention on a Prohibition of Cloning Human Beings (ETS No.168, 1998) to several years, depending on the nature and degree of the problems to be solved, but also on the political will of member States), a potential weakness of international treaties is the slowness of the ratification process. Even after having voted in the Committee of Ministers in favour of the text, there is no obligation to ratify, and there is no way of guaranteeing that all member states will ratify the treaty.

ii. Convention

165. The AI framework could result from an ad hoc Convention of the Council of Europe on AI, which could, for instance, focus on algorithmic decision-making, building further on Recommendation CM/Rec(2020)1. This would be an international legally binding instrument on the responsible and beneficial design, development and application of artificial intelligence for the setting up of automated decision-making systems, based on the Council of Europe's standards on human rights, rule of law and democracy.

166. The AI Convention would focus on the protection of democratic values and natural and legal persons. Given that (potentially) harmful AI systems (such as emotion detection and 'criminality' prediction systems) are already being rolled out across the world with little to no oversight, the Convention could provide a robust answer by contraining such applications. It would stress the importance of a speedy accession by the maximum number of Parties in order to facilitate the formation of an all-encompassing legal regime of artificial intelligence under the Convention and urge member States and other Parties to the AI Convention to initiate the process under their national law leading to ratification, approval or acceptance of the AI Convention.

167. The added value would be to get a specific legally binding instrument on the design, development and application of artificial intelligence based on the Council of Europe's standards on human rights, rule of law and democracy. It would create harmonization of rules and obligations across states on AI deployment, as well as a clear agreement regarding AI research and

development procedures, which seems crucial for the way forward. Successful examples of such innovative legal frameworks developed in related areas by the Council of Europe in the past are the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108), and the Budapest Convention on Cybercrime (CETS No.185).

168. It could also be valuable to combine the idea of an additional protocol to the European Convention on Human Rights with the content of the Convention 108+. This Convention could state positive and negative obligations of States (obligations to take measures to protect, to inform, to efficiently deal with the adverse consequences of a damage resulting from an AI system, etc) and propose an effective network of independent competent authorities to ensure the effective implementation of those safeguards. These authorities could deal with acts or omissions of States in the field of AI and engage the State's responsibility under the Convention under some circumstances. The Convention could also support the States in taking adequate steps to secure high standards for the use of AI.

169. At the same time, attempting to draft a full-fledged convention containing detailed legal obligations concerning AI might be considered premature. An overly prescriptive and rigid approach could cause states to discourage developments in the field, out of a concern that their approach risks being considered a violation of their international obligations. Rigid rules could stymie regulatory innovation and curtail research, development and deployment of new technologies and cutting-edge solutions to existing problems, many of which could save lives and benefit society as a whole.

iii. Framework Convention

170. Under the so-called "framework convention and protocol approach", parties agree on a more general treaty, the framework convention, and more detailed protocols to fill out the room left for specific regulations. The framework convention formulates the objectives of the regime, the establishment of broad commitments for its parties and a general system of governance. More detailed rules and the setting of specific targets are left to either parallel or subsequent agreements between the parties. This regulatory technique, which has certain benefits compared to single "piecemeal" treaties in international law, could be particularly appropriate in the field of AI.

171. A multilateral "Framework Convention on Responsible AI" would significantly contribute to safeguard European values and standards in requiring the States to mutually agree on the scope of the legal framework on Responsible AI and the procedure to be complied with to offer effective safeguards in the design, development and application of an artificial intelligence based on the Council of Europe's standards on human rights, rule of law and democracy. It could contain the commonly agreed upon core principles and rules for AI research, development and implementation, in the interests of human society. It could also contain specific rules on safeguards (including data protection), preventive measures, jurisdiction, liability, international cooperation. For instance, a model agreement could be developed for the exchange of information. This could be useful to mobilise a network of already existing independent competent authorities like the ones dedicated to data protection or competition supervision a national level. The Framework Convention on Responsible AI could also set forth the rules and procedures necessary for States to implement the Convention.

172. Whereas the addressees of the Framework Convention are states and state bodies, private actors have an important role to play in the implementation of specific regulations implemented at the national scale on the basis of broad international commitments. In particular, they could take up a prominent role in the design of co-regulatory mechanisms by which States would, in close interaction with private actors, give further shape to their international commitments. In this way, the Framework Convention could be an adequate mechanism to regulate a dynamic area like AI,

as it would not only establish clear binding rules and duties for Member States, but also indirectly for stakeholders, like public and private companies, researchers and civil society organisations, by engaging them in co-regulatory efforts to establish more granular rules both horizontally and intersectorally. Failing to establish effective co-regulation in the expected timeframe would imply that Member States are no longer in compliance with their international commitments. Private actors could sign a legally binding agreement with the Council of Europe to report on an annual basis all technical and organisational measures taken to design, develop and apply an artificial intelligence based on human rights, rule of law and democracy. Such mechanism would require a monitoring system and a compliance assessment list with publicly accessible lables of results of the assessments.

173. The added value of such Framework Convention is that it allows parties to agree on the core principles and values to be respected, irrespective of the context, in relation to the design, development and application of artificial intelligence, whereas more detailed protocols and/or existing context-specific legal instruments (such as Convention 108+), as well as soft law instruments, could fill out the room left for detailed provisions that apply to specific contexts. An existing example of such Framework Convention at CoE level is the Framework Convention for the protection of national minorities (FCNM). The FCNM is a legally binding instrument under international law and provides for a monitoring system, but the word “Framework” highlights the scope for member states to translate the Convention’s provisions to their specific country situation through national legislation and appropriate governmental policies. Another example, albeit not officially carrying the term "Framework" in its title, is the Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine, in short, the Convention on Human Rights and Biomedicine (so-called "Oviedo Convention"). This Convention was adopted in 1997 to respond to potential misuses of scientific advances in the field of biology and medicine. It draws on the principles established by the European Convention on Human Rights and aims at protecting the dignity and identity of all human beings. It sets out fundamental principles applicable to daily medical practice and also deals specifically with biomedical research, genetics and transplantation of organ and tissues. It is further elaborated and complemented by Additional Protocols on specific subjects, for instance, on the prohibition of cloning human beings.

iv. Soft law instrument(s)

Council of Europe

174. A distinction should be made between soft law instruments at the level of the Council of Europe, on the one hand, and at the national level, on the other hand. The former are already heavily relied upon in the vertical pillars (cf. Chapter 4), but could be complemented with a general instrument, such as a recommendation or declaration, that consolidates the common principles. This soft law instrument could operate as a stand-alone document, or complement a binding instrument in order to further operationalize its provisions. Other options include the drafting of guidance documents with a view of increasing the understanding of the relationship between the protection of human rights, democracy and rule of law, and artificial intelligence (e.g. by providing information about case law of the ECtHR), and hence, contributing to strengthening protection at the national level. Such ‘manuals’ or ‘guides’ can be developed through broad multi-stakeholder consultation with governments, private companies, civil society organisations and representatives of the technical community and academia and should be evolving documents that are updated periodically and fleshed out collaboratively, in light of new developments. Precedents include the Manual on Human Rights and the Environment) and the Guide to Human Rights for Internet Users.

Member State level

175. Soft law mechanisms at the national level could be encouraged by the AI framework adopted at CoE level. Such mechanisms are preferably approved by national competent authorities, to further operationalize the AI framework and demonstrate compliance to it. Such soft law instruments could consist of approved guidelines, Codes of Conduct, labeling, marks and seals for AI, as well as certification mechanisms.

176. Whereas soft-law measures cannot, by themselves, meet the objectives of ensuring that AI applications protect and advance fundamental values and rights, they can make important contributions. Advantages of a soft law approach include flexibility, adaptability, immediacy of implementation, broader appeal and global reach, and capacity to be reviewed and amended quickly.

177. Technical standards and certifications in particular are highly versatile and can serve, depending on the domain and allowances made for national regulatory regimes, both as 'hard law' and 'soft law'. This specific property of standards and certifications must be recognized and utilized as part of an effective but also efficient and nuanced legal framework. Effective incentives can be created for corporations to adopt such soft-law instruments promptly, including through the procurement practices of intergovernmental organizations and of national public sector entities (even absent, or pending, broader system-wide regulatory requirements).

178. Standards and certifications can be developed for all stages of AI development and operations and may engage all agents involved. Hence, the addressees include all categories of agents in the design and operation of AI in the private and public sectors: designers, developers, procurement agents, deployment agents, operators, and validation agents (i.e.: those whose responsibility is to test or audit the effectiveness of the application). The added value is evident. Sound technical standards and certifications offer evidence-based instruments upon which both "hard law" and "soft law" regimes can be flexibly developed to meet the needs of different domains and the allowances of national regulatory regimes. Even when used solely as soft-law instruments, standards and certifications are extremely effective at propagating best practices across industries and economies globally. Such standards and certifications also can serve to create incentives for beneficial innovation, for example through incorporation in the procurement practices of intergovernmental organizations and of national public sector entities (even absent, or pending, broader system-wide regulatory requirements). They efficiently serve the needs of all categories of agents involved in the design, development, procurement, and operation of AI, and help empower ordinary citizens by serving as the "currency of trust" that both expert and non-expert can relate to (as we do with nutritional labels or car safety crash-tests). The underlying evidence sought by such standards and certifications can also be used to spur, accelerate, and reward innovation through open, recurring, AI innovation benchmarking initiatives such as those envisioned in Section 8.v. Absent sound evidence-based technical standards and certifications, the goal of ensuring that AI-enabled systems protect and advance the fundamental values codified in ECHR and Convention 108+ is unlikely to be met. (As adoption practices will rest on ad-hoc decisions or unsound evidence / ersatz science.)

179. Private actors, in particular (but not limited to) academic institutions and prominent standards-setting bodies, can help ensure that such soft-law instruments are in fact, theoretically sound and practically effective. Private actors, in particular corporations, can incorporate such soft-law instrument into their governance, procurement, operation, and auditing practices (as they already do with many standards and certifications related, for example, to security). It is worth noting that, building on the 2011 UN Guiding Principles on Business and Human Rights, the Committee of Ministers of the Council of Europe adopted Recommendation CM/Rec(2016)3 on human rights and business. This Recommendation provides more specific guidance to assist member States in preventing and remedying human rights violations by business enterprises and also insists on measures to induce business to respect human rights.

180. Rating Agencies could also play a role in providing an annual ranking of private organisations on Responsible AI, based on a compulsory reporting on Digital Responsibility made by the operators engaged in AI. Soft Law instruments could contribute to safeguard European values and standards.

181. However, it should be stressed that, while self-regulation might be a complementary method of implementing certain principles and rules, it cannot substitute for the positive obligations that Member States have under the ECHR to effectively protect and safeguard human rights in relation to AI. Voluntary, self-regulatory and ethics-based approaches lack effective mechanisms of enforcement and accountability, and should therefore, on their own, not be considered as a sufficient and effective means to regulate AI. Moreover, certification mechanisms are not immune from errors and mistakes and, in order for these mechanisms to be effective, a number of conditions should be fulfilled.

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

182. While like-minded common values are being promoted by some Member States or international organisations in which they are active, the cross-border agreement of international legal framework on AI systems is needed because of virtual impact of AI on Member State's jurisdictions and challenges of a global competition on the market.

183. There are numerous ways in which best practices can be identified or encouraged (many of which are familiar to, or implemented by member states or economic actors). A **European Benchmarking Institute** could be a highly effective, efficient, and trustworthy source of identification, definition, and consensus around the underlying evidence that should guide sound best practices. Such evidence can, in turn, serve as the basis for a wide-range of best practices that can be efficiently and effectively propagated by sound technical standards and certifications.

184. Also a uniform model developed at the level of the Council of Europe to carry out a human rights, democracy and rule of law impact assessment could be extremely helpful in harmonising member States' implementation of common values in relation to AI systems.

185. Another approach that could be supported at the national level is the set-up of pilot projects and regulatory sandboxes. This would require a balanced legal framework wherein such experimentation can take place while addressing legal challenges as they arise.

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.

186. In order to ensure complementarity between the horizontal/general instrument and vertical/sector-specific instruments of the Council of Europe (e.g. in relation to media, cybersecurity, justice, democracy, education...), the former could include explicit references to the existing or to-be-developed instruments in the different pillars of the Council of Europe (described in Chapter 4). Another mechanism to ensure complementarity could be the setting up of a joint certification scheme/body, comparable to the one existing in the pharmaceutical sector (the European Directorate for the Quality of Medicines & HealthCare and its Pharmacopoeia). Such joint certification mechanism/body could, firstly, be tasked with providing more detailed guidelines regarding human rights impact assessments and common quality standards at European level. Subsequently, it could be responsible for supporting the implementation and monitoring the application of quality standards for AI systems, just like EDQM does for safe medicines and their safe use.

vii. Concluding remark.

187. Giving the evolving nature of AI, a co-regulatory approach is desirable. A robust legal framework will therefore likely consist of a combination of binding and non-binding instruments, that complement each other. A binding instrument, a Convention or Framework Convention, should establish the 'general' legal framework for AI consolidating general common principles – contextualised to apply to the AI environment– and include more granular provisions addressing specific issues and sectors, like those identified in CAHAI(2020)08-fin (i.e., healthcare, data protection, democracy and justice). This instrument, which should include appropriate follow-up mechanisms and processes, could be combined with additional (binding or non-binding) sectoral CoE instruments establishing further sector specific principles and detailed requirements on how to address specific sectoral challenges of AI, for instance by incentivising private actors to develop self-regulatory initiatives. This approach would allow for the flexibility required for technological development.

9. Possible practical mechanisms to ensure compliance and effectiveness of the legal framework

9.1 - The Role of Compliance Mechanisms

188. The ultimate effectiveness of any [legal framework] will depend on the breadth of its adoption and compliance. Practical mechanisms (such as impact assessments, lifecycle auditing, and monitoring, certification methods, and sandboxes) are one way of driving such compliance, helping member states to understand and monitor adherence to the [legal framework]. Such mechanisms confer further benefits beyond compliance, for example, increasing transparency around the use of AI and creating a common framework for promoting trust.

189. A [legal framework] should formulate the abstract requirement to develop compliance mechanisms at a general level as well as what principles need to be fulfilled by any practical mechanisms to ensure compliance. It is a matter for Member States to decide how to enforce this through their legislative framework, including which practical mechanisms they choose to make mandatory or which actors or institutions they empower to provide independent, expert, and effective oversight. This enables implementation to account for existing local institutions' roles, regulatory culture, and legal requirements. Rather than mandating a single solution, this approach further enables the creation of an AI assurance ecosystem, which creates the potential for diverse participation and the emergence of novel and innovative approaches to compliance. That said, collaboration between Member States should be considered paramount to protect against the risk of diverging approaches and the fragmentation of markets.

190. Compliance mechanisms might be used to assess both the use case and the design of the AI-enabled system.

[Placeholder link to other chapters about:

1. The red-green types of AI applications

2. The use in different contexts, e.g. public sector vs private sector applications

3. A potential two-tier assessment to determine which use-cases require a more robust assessment]

191. On the question of when AI should be subject to such assessment, we agreed on the fundamental importance of ex-ante assessment and continuous assessment at various milestones throughout the AI project lifecycle. Compliance mechanisms should also evolve over time to account for the evolving nature of the system. The ongoing assessment approach presents three salient advantages. 1) allows for a better understanding of the AI project implications (from design, development, and deployment); 2) facilitates decision-making to reconsider future unforeseen uses of the AI project and; 3) monitors any changes in the behavior of the model ex-post (particularly crucial in e.g., reinforcement learning contexts)[104]. The procurement of pre-built AI-enabled solutions and technical advancements such as transfer learning scenarios present challenges that need to be considered.

9.2 - The Role of Different Actors

192. As outlined above, each Member State should ensure national regulatory compliance with any [framework convention]. Different actors should play contribute in a complementary way to bring about a new culture of AI applications that are compliant with the legal framework principles and local regulations to generate adequate compliance and oversight incentives, either as Assurers, Developers, or Operators and Users.

Assurers of systems

193. Member States should also be responsible for identifying and empowering independent actors to provide oversight. [105] These independent actors could be an expert committee, academics, sectoral regulators or private sector auditors[106]. Where they do not exist already, Member States might consider setting up independent oversight bodies equipped with appropriate and adequate inter-disciplinary expertise, competencies, and resources to carry out their oversight function. Such bodies might be equipped with intervening powers and be required to report to parliament and publish reports about their activities regularly[107]. They might also resolve disputes on behalf of citizens or consumers. For example, states could extend the mandate of existing or create new ombudsman to assess and resolve any complaints or appeals, complementing more binding mechanisms[108]. It is unreasonable to expect that such a body might be able to cover all AI-based products and systems and so consideration to scope would be important.

194. Many AI systems are deployed across multiple jurisdictions. It is vital for adequate oversight and fast policymaking to share information among the Member States. There is a role to develop mechanisms of information sharing and reporting[109] about AI Systems under each Member State's regulatory framework (e.g. sharing certified AI systems, banned AI applications or the current status of a specific AI application). This could be convened by an expert committee operating at an international level[110], appointed by agreement between relevant national regulators. Notification platforms could further support collaboration, providing international counterparts with visibility of local violations of the [legal framework][111].

[104] Backed by Germany, Netherlands at PDG meeting & 2nd plenary, p.11
[105] CINGO drafting group contribution
[106] IEEE drafting group contribution
[107] Unboxing AI
[108] Backed by Sweden, Council of Europe & Switzerland at the PDG meeting and highlighted in Catelijne Muller report, Sweden drafting group contribution
[109] Israel written contribution to feasibility study
[110] Israel written contribution to feasibility study
[111] Turkey drafting group contribution

195. Private sector actors can also play a role in assuring systems. In addition to auditing services, certification schemes can support the [legal framework] promoting an active role for the private sector in preventing and managing the risks of adverse human rights impacts associated with AI-enabled systems. Within certification schemes, professional training could include the [legal framework] as part of the training curricula. In broader terms, universities and civil society could be part of education policy to disseminate, research, and instruct on the [legal framework] and technical developments. This approach would also confer further benefits in a global market economy.

Developers of systems

196. Actors building AI-enabled systems (both private and public sector) should also consider actions they can take to increase compliance with a [legal framework]. For example, there could be policies to increase the visibility of where such technologies are being deployed (e.g., public sector contracts should be published / public registers[112] or notification systems[113] might be another way of doing this) or developing norms and standardized tools for internal audit and self-certification (acknowledging the limitations of this approach)[114]. Liability considerations should also be taken into account[115].

Operators and Users of systems

197. Operators and users of AI could generate demand for AI applications that comply with the [legal framework]. This is particularly true of the public sector and its relative procurement power. The promotion of trust carriers, such as certification labels on AI systems' lifecycles, and periodic auditing and reporting, are market responses pushed by operators and users of AI systems preferences and expectations. When operators and users of AI systems become better informed of their rights and redress mechanisms, the transaction cost of oversight is significantly reduced.

9.3 - Examples of Types of Compliance Mechanism

198. There are many contexts where organizations are already required to meet standards or regulations, for example, financial services and healthcare. Each of these systems has evolved into ecosystems of services that allow organizations to prove to themselves, their customers, and regulators that they have met a required standard. Different mechanisms will work best in different contexts, depending on sector as well existing infrastructure, sectoral mechanisms and institutions[116]. It should also be considered which components of an AI-enabled system might be subject to compliance, for example, the training data used, the algorithm construction, the weighting of different inputs, or the accuracy of any outputs. Inclusive participatory processes should be conducted to establish the relevant regulatory and enforcement mechanisms in each case[117].

199. A [legal framework] might specify that practical mechanisms adhere to a set of principles that promote the framework's core values. These might include:

[112] CINGO written contribution to feasibility study
[113] CINGO written contribution to feasibility study, Sweden drafting group comments
[114] Bulgaria written contribution to feasibility study
[115] Mexico written contribution to feasibility study
[116] Germany written contribution to feasibility study Backed by Netherlands, 2nd plenary, p. 11, CDDG comments to first draft
[117] CDDG comments to first draft

- **Dynamic (not static):** assessment ex-ante and at various points throughout the AI project lifecycle[118] to account for any changes in the behaviour of learning models.
- **Technology adaptive:** to support the future-proofing of any compliance mechanisms[119].
- **Differentially Accessible:** understandable to experts and non-experts, in turn simplifying the process of any potential appeals and redress[120].
- **Independent**: conducted, or overseen, by an independent party.
- **Evidence-based**: supported on evidence produced by technical standards and certifications. For example, including data collected through best practices such as borderless, standardization or key metrics developed through benchmarking[121].

200. Any mechanisms need to be practically implementable, accounting for existing governance infrastructure and technical limitations. The practical mechanisms outlined below should therefore be considered as a toolkit that presents ample opportunity for further regulatory innovation[122] and refinement:

(1) Human rights impact assessments[123] - The European Commission's High Level Expert Group on AI has recommended that organisations perform a fundamental rights impact assessment. These assessments might explicitly validate conformity with principles outlined in the [legal framework]. **[124]** In specific contexts, 'integrated impact assessments' might be deemed more appropriate to reduce the administrative burden on development teams (bringing together, for example, human rights, data protection, transparency, accountability, competence, and equalities considerations). Conducting human rights due diligence is also recommended by the United Nations Guiding Principles on Business and Human Rights (UNGPs)[125].

(2) Certification & Quality Labelling - Ex-ante obligations, administered by recognized bodies and independently reviewed, would help build trust. An expiration date would ensure systems are re-reviewed regularly **[126] .** Such schemes could be made voluntary or mandatory, depending on the maturity of the ecosystem [127]. Legal safeguards must ensure certifications are not used by companies to shield themselves from potential liability claims associated with their conduct. The certification process should subject to regulation regarding auditors' qualifications, the standards adopted, and how conflicts of interests are managed [128]. The certification process should strive for continuous improvement and be responsive to complaints[129]. Ongoing multi stakeholder standards development work would support this[130] led by standard setting bodies[131].

[118] 2nd plenary, p.11, Poland drafting group contribution
[119] 2nd plenary, p.11
[120] 2nd plenary, p.11
[121] IEEE drafting group contribution
[122] Israel written contribution to feasibility study
[123] Sweden drafting group contribution; Bulgaria, Netherlands and Conference INGOs written contribution to feasibility study, Catelijne Muller report
[124] Backed by Netherlands in PDG meeting
[125] IBA drafting group submission
[126] Germany written contribution to feasibility study, Telefonica drafting group contribution
[127] Telefonica drafting group contribution
[128] International Bar Association written contribution to feasibility study
[129] IBA drafting group submission
[130] Turkey, Poland & IEEE drafting group contributions
[131] CDDG comments to first draft

(3) Audits - Regular independent assessments or audits of AI-enabled systems by experts or accredited groups is also a mechanism that should be exercised throughout the lifecycle of every AI-enabled system to verify their integrity, impact, robustness, and absence of bias. Audits will facilitate a move towards more transparent and accountable use of AI-enabled systems[132]. Audits could certify organisations as a whole, rather than just specific use cases.

(4) Regulatory Sandboxes - Regulatory sandboxes, particularly those that enable closer regulatory support, present an agile and safe approach to testing new technologies and could be used in order to strengthen innovative capacity in the field of AI. Sandboxes could be of particular use where a timely, possibly limited market introduction appears warranted for public welfare reasons, e.g. in extraordinary crises such as a pandemic[133], or in cases where current legal frameworks have not been tested in practice that could lead to constrained innovation. Cross-jurisdictional sandboxes present further opportunities for collaboration, building on the model of the Global Financial Innovation Network[134].

9.4 - Conclusion

201. Mandating practical mechanisms to enforce compliance should be considered only one part of a broader package of initiatives required to drive change. Member States could reinforce compliance mechanisms with several initiatives. For example, to invest in digital literacy, skilling up and building competencies and capacities of developers, policymakers[135] and wider society[136] to understand the human rights implications of AI-enabled systems; to drive the widespread adoption of norms such as open access to source code[137]; or engaging with human rights civil society organizations as key stakeholders at various stages of development[138].

202. This more comprehensive work to develop best practices and norms within existing legal and regulatory regimes should be accompanied by ongoing discourse, collaboration, and best practice sharing between actors at a national and international level. Centres of Expertise would be well placed to facilitate collaboration on innovative solutions to inter-sectoral regulation projects[139].

10. Final considerations

[132] IEEE drafting group contribution
[133] Germany written contribution to feasibility study, Greece drafting group contribution
[134] https://www.fca.org.uk/firms/innovation/global-financial-innovation-network
[135] Netherlands written contribution to feasibility study
[136] CDDG comments to first draft
[137] Mexico written contribution to feasibility study
[138] CAHAI(2020)21 rev PDG contributions pg 45-46
[139] CAHAI(2020)21 rev PDG contributions pg 32-33