COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 3 December 2021

CAHAI(2021)09rev
Restricted

AD HOC COMMITTEE
ON ARTIFICIAL INTELLIGENCE (CAHAI)

Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law

www.coe.int/cahai

Part I: Introduction

I Background

1. This paper contains the outcomes of the work of the Council of Europe Ad hoc Committee on Artificial Intelligence (CAHAI) on the potential elements of a legal framework for the development, design and application of artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law.

2. The paper has been drafted by two Working Groups of the CAHAI, namely the CAHAI Policy Development Group (CAHAI-PDG) and the CAHAI Legal Frameworks Group (CAHAI-LFG) while taking into account the outcomes of the multi-stakeholder consultation conducted in the Spring of 2021 by the third Working Group, the CAHAI Consultations and Outreach Group (CAHAI-COG). It was examined and adopted by the CAHAI at the occasion of its sixth Plenary meeting on 30 November – 2 December 2021, and consequently submitted to the Committee of Ministers for further consideration in line with the terms of reference of the CAHAI.

II General remarks

3. The CAHAI observes that the application of artificial intelligence (AI) systems has the potential to promote human prosperity and individual and social well-being by enhancing progress and innovation, yet at the same time certain applications of AI systems give rise to concern, as they potentially pose risks to human rights, democracy and the rule of law.

4. To effectively prevent and/or mitigate these risks, the CAHAI considers that an appropriate legal framework on AI based on the Council of Europe's standards on human rights, democracy and the rule of law, should take the form of a legally binding transversal instrument. The CAHAI notes that – in addition to the proposed legally binding transversal instrument that sets out general principles and specific legal norms – existing or future legally binding and/or non-legally binding instruments may be needed at sectoral level, to provide more detailed guidance on ensuring that the design, development and application of AI occurs in line with human rights, democracy and the rule of law in specific domains.

5. The legally binding transversal instrument should focus on preventing and/or mitigating risks emanating from applications of AI systems with the potential to interfere with the enjoyment of human rights, the functioning of democracy and the observance of the rule of law, all the while promoting socially beneficial AI applications. It should be underpinned by a risk-based approach: the legal requirements to the design, development and use of AI systems should be proportionate to the nature of the risk they pose to human rights, democracy and the rule of law. Basic principles that enable the determination of such risk (e.g. transparency requirements) should be applicable to all AI systems.

6. In accordance with Article 1 d of the Statute of the Council of Europe, matters relating to national defence fall outside the scope of a legal framework of the Council of Europe and are therefore not covered in the scope of a legally binding (or non-legally binding) instrument of the Council of Europe. The CAHAI is of the opinion that the issue of whether that scope could cover "dual use" and national security should be further considered in the context of developing a Council of Europe legal framework on AI, taking into account possible difficulties in this respect.

7. The various legal issues raised by the application of AI systems are not specific to the member States of the Council of Europe, but are, due to the many global actors involved and the global effects they engender, transnational in nature. The CAHAI therefore recommends that a legally binding transversal instrument of the Council of Europe, though obviously based on Council of Europe standards, be drafted in such a way that it facilitates accession by States outside of the region that share the aforementioned standards. Not only will this significantly increase the impact and efficiency of the proposed instrument, but in addition it will provide a much-needed level playing field for relevant actors, including industry and AI researchers which often operate across national borders and regions of the world. The standards of the Council of Europe on human rights, democracy and the rule of law are sufficiently universal in nature to make this a realistic option. There are several precedents of Council of Europe treaties being applied beyond the European region, cf. notably the Budapest Convention (Cybercrime) and the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108), which currently have 66 and 55 Parties respectively, many of which are not member States of the Council of Europe.

8. It is further recommended that, to ensure both global and regional legal consistency, a legally binding transversal instrument of the Council of Europe should take into account existing and upcoming legal and regulatory frameworks of other international and regional fora, in particular the United Nations system (including UNESCO), the European Union, and the Organisation for Economic Co-operation and Development – all of which are currently involved in developing various forms of standards related to AI systems.

9. The CAHAI notes that the purpose of an international legal framework should not be to lay down any detailed technical parameters for the design, development and application of AI systems, but to establish certain basic principles and norms governing the development, design and application of AI systems and regulate, in a consistent and deliberate manner, if and on what conditions AI systems potentially posing risks to the enjoyment of human rights, the functioning of democracy and the observance of the rule of law may be developed, designed and applied by all types of organisations, including public and private actors alike.

10. In Part II (*chapters III – XI,* developed by the CAHAI-LFG), the elements which could be part of a legally binding transversal instrument are set out. Part III (*chapters XII and XIII,* developed by the CAHAI-PDG*)* outlines the elements which could be part of possible additional legal instruments.

PART II: Elements for a legally binding transversal instrument

III Elements relating to object and purpose, scope, and definitions

11. Concerning the *object and purpose* of the legally binding transversal instrument, the CAHAI recommends that it should, in particular, be stated that the aim of the instrument is to ensure full consistency with respect for human rights, the functioning of democracy and the observance of rule of law in the developing, designing and applying of AI systems, irrespective of whether these activities are undertaken by private or public actors. Further, it should be stated that the instrument shall facilitate cooperation to this end by its Parties, both at international and domestic levels, and that the necessary follow-up mechanisms shall be established. Finally, the object and purpose should underline the need for establishing a common legal framework containing certain minimum standards for AI development, design, and application in relation to human rights, democracy and the rule of law.

12. The CAHAI considers that the legally binding transversal instrument should contain a provision defining its *scope*. This provision should clarify that the instrument shall be applicable to the development, design and application of AI systems, irrespective of whether these activities are undertaken by public or private actors, with a particular focus on such systems which are assessed to pose potential risks to the enjoyment of human rights, the functioning of democracy, and the observance of the rule of law. As necessary, potential exceptions to the scope should also be addressed.

13. In so far as *definitions* are concerned, the CAHAI considers that, as a minimum, the following definitions should be included in a legally binding transversal instrument: *"Artificial intelligence system"; "lifecycle"; "AI provider"; "AI user"; "AI subject"; "unlawful harm"*. The CAHAI recommends that all definitions used should, in so far as possible, be compatible with similar definitions used in other relevant instruments on AI. Furthermore, definitions should be carefully drafted to ensure, on the one hand, legal precision, while, on the other hand, being sufficiently abstract to remain valid despite future technological developments concerning AI systems.

IV Elements relating to fundamental principles of protection of human dignity and the respect of human rights, democracy, and the rule of law

14. The CAHAI considers it necessary that a legally binding transversal instrument contains certain *fundamental principles of protection of human dignity and the respect of human rights, democracy, and the rule of law*, which should apply to all development, design, and application of AI systems, irrespective of whether the actor is public or private.

15. At the same time the CAHAI, recognising the risks of duplicating or even fragmenting existing general standards of international law, including human rights law, recommends that such fundamental principles be drafted in such a way that the risks of unwarranted duplication or fragmentation are duly minimised. This entails, *inter alia*, further tailoring rights and obligations relating to human rights, democracy and the rule of law for the purpose of this instrument only where and when, after careful examination, the conclusion is reached that existing standards in their current form cannot provide sufficient protection of the rights of individuals in the specific context of the development, design and application of AI systems.

16. Concerning the concept of "human dignity", the CAHAI notes that the dignity of the human person is universally agreed to constitute the real basis of human rights, cf. also the prominence given to the concept in the preamble of the 1948 Universal Declaration of Human Rights. In the view of the CAHAI, it makes particularly good sense to use this concept in a legally binding transversal instrument on the potential adverse impacts on fundamental human rights of individuals caused by the development, design, and application of AI systems.

17. The CAHAI further notes that some of the provisions related to these particular elements may be formulated as positive direct rights of individuals, or alternatively as obligations on Parties to ensure the introduction in their domestic law and practice of measures aimed at protecting the rights of individuals in relation to AI systems. Based on its deliberations the CAHAI would, where feasible and necessary, tend to favour a combination of both the establishment of certain direct, concrete and *positive rights of individuals in relation to the development, design and application of AI* systems, as

well as the *establishment of certain obligations upon Parties*, to ensure a more uniform application of the legally binding transversal instrument among Parties.

V Elements relating to risk classification of artificial intelligence systems and prohibited applications of artificial intelligence

18. The CAHAI recommends that a legally binding transversal instrument should provide for the establishment of a methodology for *risk classification* of AI systems with an emphasis on human rights, democracy, and the rule of law. The criteria used for assessing the impact of application of AI systems in this regard should be concrete, clear, and with an objective basis and the assessment itself done in a balanced manner, thus providing for both legal certainty and nuance.

19. In particular, the CAHAI considers that the risk classification should include a number of categories (e.g., "low risk", "high risk", "unacceptable risk"), based on a risk assessment in relation to the enjoyment of human rights, the functioning of democracy and the observance of the rule of law. The risk classification will be based on an initial review to determine if a full HUDERIA ("Human Rights, Democracy and Rule of Law Impact Assessment") is required (cf. chapter XII) just as the impact assessment itself may have an impact on whether to uphold or change the initial risk classification of the AI system in question. This impact assessment is considered as an element of the overall legal framework on AI systems proposed by the CAHAI. However, the specific HUDERIA model need not necessarily form a constituent part of a possible legally binding instrument.

20. As regards the criteria which could be considered for the purpose of the risk assessment, reference is made to the elements listed under paragraph 51 in chapter XII below. Some of these criteria may need to be enshrined in the legally binding instrument, to ensure they are duly considered and consistently applied.

21. Regarding *prohibited applications of AI (the so-called "red lines" or "unacceptable risk")*, the CAHAI considers that a legally binding transversal instrument should provide for the possibility of putting a full or partial moratorium or ban on the application of AI systems, which in accordance with the aforesaid risk classification are deemed to present an unacceptable risk of interfering with the enjoyment of human rights, the functioning of democracy, and the observance of the rule of law. Such possibility should also be considered for the research and development of certain AI systems that present an unacceptable risk. Notably, the CAHAI wishes to draw the attention to, for instance, some AI systems using biometrics to identify, categorise or infer characteristics or emotions of individuals, in particular if they lead to mass surveillance, and AI systems used for social scoring to determine access to essential services, as applications that may require particular attention, taking into account possible legitimate exceptions. A moratorium or ban should, however, only be considered, where on an objective basis an unacceptable risk to human rights, democracy or the rule of law has been identified and, after careful examination, there are no other feasible and equally efficient measures available for mitigating that risk and given the specific sphere of application. Review procedures should be put in place to enable reversal of a ban or moratorium if risks are sufficiently reduced or appropriate mitigation measures become available, on an objective basis, to no longer pose an unacceptable risk.

VI Elements relating to the development, design, and application of artificial intelligence systems in general

22. The CAHAI recommends that a legally binding transversal instrument should include a number of provisions applicable to all development, design and application of AI systems, so as to enable their appropriate classification in terms of potential risk to the enjoyment of human rights, the functioning of democracy, and the observance of the rule of law, and to ensure their compliance therewith by setting out minimum safeguards. These can include, for instance, provisions regarding the transparency of AI systems. In line with the risk-based approach mentioned above, further provisions should be rendered applicable to AI systems based on and in proportion with their risk classification, in order to ensure that the risks they pose to human rights, democracy and the rule of law are duly mitigated.

23. A legally binding transversal instrument should, as a general rule, state that, subject to certain limitations, *the development and design of, as well as the research in, AI systems should be carried out freely, with due consideration for safety and security*, and in full compliance with the Council of Europe standards on human rights.

24. Furthermore, the CAHAI recommends the inclusion of a provision encouraging Parties to establish *"regulatory sandboxes"* to stimulate responsible innovation in AI systems by allowing for the testing of AI systems under the supervision of the competent national regulator, all the while ensuring compliance with the standards set out in the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No 108) and its amending Protocol (CETS No 223), as well as with the standards set out in this legally binding transversal instrument on the design, development and application of AI, and any other applicable standards.

25. To promote a multi-stakeholder approach, and in order to raise awareness in society about the impact of the development, design and application of AI systems, the CAHAI considers it useful to include a provision calling for Parties to promote evidence-based *public deliberations* on and inclusive engagement with this topic. Inspiration for the wording of such a provision may be found in Article 28 of the Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine (CETS No 164).

26. The CAHAI proposes to include a provision on *prevention of unlawful harm* potentially stemming from the development, design, and application of AI systems, including clarifying the concept of "unlawful harm" for the purpose of the transversal instrument on AI, human rights, democracy and the rule of law.

27. The CAHAI further proposes to include a provision on respect of *equal treatment and non-discrimination* of individuals in relation to the development, design, and application of AI systems to avoid unjustified bias being built into AI systems and the use of AI systems leading to discriminatory effects.

28. For the same reasons, a legally binding transversal instrument should contain provisions on ensuring that *gender equality* and rights related to *vulnerable groups and people in vulnerable situations,* including *children,* are being upheld throughout the lifecycle of artificial intelligence systems.

29. The CAHAI also considers it prudent to include a provision on *data governance* for AI systems, in accordance with and building on the Council of Europe Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No 108) and its amending Protocol (CETS No 223). This can include the requirement to establish data governance mechanisms to assess and ensure the data accuracy, integrity, security and representativeness in a manner that is suitable for the intended purpose of the system and proportionate.

30. Finally, the CAHAI recommends the introduction of provisions on *robustness, safety and cybersecurity*, *transparency, explainability, auditability and accountability* throughout their lifecycles. It should be noted that the concepts of "transparency", "explainability" and "accountability" are considered by the CAHAI to be of paramount importance for the protection of the rights of individuals in the context of AI systems. In addition, the CAHAI recommends that the issue of *sustainability* in relation to AI systems throughout their lifecycles be considered in a suitable manner.

31. Last, but not least, a legally binding transversal instrument should include a provision aimed at ensuring the necessary level of *human oversight* over AI systems and their effects, throughout their lifecycles.

VII Elements relating to the development, design, and application of artificial intelligence systems in the public sector

32. The development, design, and application of AI systems in the public sector give rise to some concerns about how to ensure the respect for human rights, democracy, and the rule of law when AI systems are used to take or inform decisions that impact the rights and obligations of individuals and legal persons. That said, the CAHAI underlines that not all public sector AI applications pose risks to the enjoyment of human rights, the functioning of democracy and the observance of the rule of law. It is accordingly important to carefully examine the potential for risk posed by a given application of an AI system on a case-by-case basis. Accordingly, a distinction should be made between, on the one hand, AI systems which can interfere with human rights, democracy or the rule of law, and on the other hand, AI systems which though operated by the same public authorities do not present any such risks.

33. Based on the assumption that a legally binding transversal instrument should be general in nature, the CAHAI recommends that such instrument should focus on the potential risks emanating from the development, design, and application of AI systems for the purposes of *law enforcement*, *the administration of justice*, and *public administration.* Concerning "public administration", in particular, the CAHAI notes that a legally binding transversal instrument should not address the plethora of specific administrative activities undertaken by public authorities, such as health care, education, social benefits etc, but be limited to general prescriptions about the responsible use of AI systems in public administration. Issues related to the various sectors of public administration may, as necessary, be addressed in appropriate sectoral instruments.

34. The CAHAI finds that a legally binding transversal instrument when addressing the development, design, and application of AI systems in the public sector should, as a minimum, include provisions on *access to effective remedy,* a mandatory *right to human review* of decisions taken or informed by an AI system except where competing legitimate overriding grounds exclude this*,* and an *obligation for public authorities to implement adequate human review for processes which are informed or supported by AI systems and to provide relevant individuals or legal persons with*

meaningful information concerning the role of AI systems in taking or informing decisions relating to them, except where competing legitimate overriding grounds exclude or limit such review or disclosure. Furthermore, Parties should be obliged to ensure that *adequate and effective guarantees against arbitrary and abusive practices* due to the application of an AI system in the public sector are afforded by their domestic law.

35. The CAHAI also notes that Parties should ensure compliance with the standards concerning AI systems as regards human rights, democracy and the rule of law insofar as private actors acting on their behalf are concerned.

VIII Elements relating to democracy and democratic governance

36. While recognising that AI may play a positive role in the functioning of democracy and democratic governance to foster inclusive and participatory processes, the CAHAI is also concerned about the potential use of AI to unlawfully or unduly interfere in democratic processes. The shaping of public opinion through AI, as well as potential chilling effects arising through the use of AI, should therefore be considered in the context of a possible legally binding instrument, whereas more specific issues regarding election manipulation such as *micro-targeting, profiling,* and *manipulation of content* (including so-called "deep fakes") could be dealt with in more sectoral instruments.

37. The role of private entities, for instance, online platforms that help shape the public sphere, should also be considered in this respect, insofar as the growing concentration of economic power and of data could undermine democratic processes.

38. In this context, the CAHAI underlines the need for respecting *the right to freedom of expression, including the freedom to form and hold opinions and to receive and impart political information and ideas, and the right to freedom of assembly and association*, with the aim of ensuring that all parties and interest groups have access to democratic processes in equal conditions, and that a free space for public debate can be ensured.

IX Elements relating to safeguards

39. The CAHAI recommends that a legally binding transversal instrument should include a series of provisions on legal safeguards to be applied to all applications of AI systems used for the purpose of deciding or informing decisions impacting the legal rights and other significant interests of individuals and legal persons.

40. These safeguards should, at least, include the following: the right to *an effective remedy before a national authority* (including judicial authorities) against such decisions; the right to *be informed about the application of an AI system in the decision-making process*; and the right to *choose interaction with a human in addition to or instead of an AI system,* and the right to *know that one is interacting with an AI system rather than with a human.* Other safeguards may be relevant depending on the specificities of the AI systems being used. The modalities of the exercise of these rights should be foreseen by national law. Legitimate exceptions to these rights may be foreseen by law, where necessary and proportionate in a democratic society.

41. Finally, the CAHAI is of the opinion that a legally binding transversal instrument should also include a provision on the *protection of whistle-blowers* in relation to the development, design, and application of AI systems which potentially could adversely impact the enjoyment of human rights,

the functioning of democracy, and the observance of the rule of law. Such provision should respect legitimate legal limitations on disclosure.

X Elements relating to civil liability

42. Though recognising that issues related to civil liability and the development, design, and application of AI systems would in general be covered by existing domestic law of the Parties to a possible legally binding instrument, the CAHAI nevertheless considers it useful to examine the issue in more detail to explore the need to ensure that all parties share a common basic approach to civil liability in relation to AI.

XI Elements relating to supervisory authorities, compliance, and cooperation

43. The CAHAI considers that a legally binding transversal instrument should include provisions obliging Parties to take all necessary and appropriate measures to *ensure effective compliance* with the instrument, in particular through the *establishment of compliance mechanisms and standards*. Furthermore, provisions on the establishment or designation of *national supervisory authorities*, defining their powers, tasks and functioning as well as ensuring their expertise, their independence and impartiality in performing their functions, and the allocation of sufficient resources and staff, should be considered for inclusion. In addition, the legally binding transversal instrument should contain provisions regulating the *cooperation between Parties* and *mutual legal and other assistance, including exchange of data and other forms of information* while ensuring coherence with other already applicable instruments of the Council of Europe in the field of international mutual legal assistance.

44. A legally binding transversal instrument should also contain provisions on the establishment of a "*committee of the parties"* to support the implementation of the instrument. In this regard, the CAHAI refers to the standard provisions used in other Council of Europe legally binding instruments, which may, if and as necessary, be amended to better suit the purposes of the present legally binding instrument.

PART III: Elements for possible additional legal instruments

XII Human rights, democracy, and rule of law impact assessment

45. The CAHAI considers it useful and necessary to supplement a legally binding transversal instrument with a *non-legally binding model for assessing the impact of AI systems on the enjoyment of human rights, the functioning of democracy, and the observance of the rule of law*.

46. A well-conducted human rights, democracy, and rule of law impact assessment can advance the assessment of how the deployment of AI systems can affect the enjoyment of human rights, the functioning of democracy, and the observance of the rule of law. It should though be noted that this type of impact assessment is not designed to balance negative and positive impacts, something which may depend on the specificities of the legal system in the jurisdiction in which the AI system is intended to be applied. In a subsequent stage, it can then be examined if and how risks identified through the HUDERIA can be mitigated, and if and how a legitimate interest can legitimize the system's use despite interference with human rights, democracy and rule of law standards, when such limitations are prescribed by law, proportionate, and necessary in a democratic society.

47. Indeed, a HUDERIA should *not stand alone,* but be supplemented, at the level of domestic or international law, by other compliance mechanisms, such as certification and quality labelling, audits, regulatory sandboxes and regular monitoring, as pointed out in the Feasibility Study. It is important that the impact assessment is aligned with such other compliance mechanisms, as it would be unjustifiably costly and burdensome to require human rights, democracy, and rule of law impact assessments that diverge from public supervisory or regulatory approaches laid down under domestic law. In addition to compliance mechanisms, it must also be ensured that effective remedies remain available for those who may be adversely impacted by the deployment of AI systems.

48. Given the time and resources necessary to undertake such an assessment, and in order to safeguard the proportionality of a risk-based approach, the CAHAI believes that, as a rule, a formalised extensive human rights, democracy, and rule of law impact assessment should only be mandated if there are clear and objective indications of relevant risks emanating from the application of an AI system. This requires that all AI systems undergo an initial review in order to determine whether or not they should be subjected to such a formalised assessment. It is recommended that indications as to the necessity for a more extensive assessment be further developed. It should also be considered that using an AI system in a new or different context or for a new or different purpose or otherwise relevant changes would require a reassessment.

49. The CAHAI underlines that adopting a risk-based approach entails that any relevant impacts by the application of an AI system on the enjoyment of human rights, the functioning of democracy, and the observance of the rule of law should be duly assessed and reviewed on a systematic and regular basis with a view to identifying mitigating measures tailored to the risks at hand, and if such mitigating measures are not deemed sufficient, applying prohibitive measures, as necessary. Furthermore, given the need for an iterative assessment process, such assessment should in any case be carried out again whenever a given AI system undergoes substantial changes.

50. The CAHAI recommends that, at least, the following *main steps* be included in a human rights, democracy, and rule of law impact assessment, subject to an initial review having been conducted, and including stakeholder involvement, where relevant:

(1) *Risk Identification*: Identification of relevant risks for human rights, democracy and the rule of law;

(2) *Impact Assessment*: Assessment of the impact, taking into account the likelihood and severity of the effects on those rights and principles;

(3) *Governance Assessment*: Assessment of the roles and responsibilities of duty-bearers, right holders and stakeholders in implementing and governing the mechanisms to mitigate the impact;

(4) *Mitigation and Evaluation*: Identification of suitable mitigation measures and ensuring a continuous evaluation.

51. As regards the *Impact Assessment* step, the CAHAI further recommends that the assessment of an AI system, at least, could include the following elements: assessment of the *context and purpose* of the AI system, *level of autonomy* of the AI system, *underlying technology* of the AI system, *usage* of the AI system (both intended and potentially unintended use), *complexity* of the AI system (part of multiple deep neural networks/building on other AI systems), *transparency* and *explainability* of the system and the way it is used, *human oversight and control mechanisms* for the AI provider and AI user*, data quality*, *system robustness/security*, involvement of *vulnerable persons or groups*, the *scale*

on which the system is used, its *geographical* and *temporal* scope, assessment of *likelihood* and *extent* of potential harm, the potential *reversibility* of such harm, and whether it concerns a "red line" application as established by domestic or international law.

52. Moreover, the CAHAI notes that whereas the impact assessment of AI systems is relatively straightforward in relation to human rights, due to the existence of clearly defined and universal obligations in this area, the impact assessment of AI systems on democracy and the rule of law may prove more difficult in some cases. Nevertheless, given the strong interlinkage between human rights on the one hand and democracy and the rule of law on the other hand, in some situations a negative impact on the former can also provide an indication of a negative impact on the latter. For instance, when the right to freedom of assembly and association or the right to free elections is hampered, it hampers the functioning of democracy. In the same vein, an interference with the right to a fair trial negatively impacts the rule of law. Furthermore, other elements can also be considered, such as the purpose and function of the system within a democratic society, its application domain (with particular attention to the use of AI systems in the public sector or the public sphere), and the way it can hamper certain democratic- and rule of law-principles (such as the principle of legality, the prevention of misuse of power, or judicial impartiality and independence).

53. Finally, the CAHAI is of the opinion that *stakeholder involvement* in the impact assessment should be assured. The more severe the impact is deemed to be, or the larger its scale, the more extensive the stakeholder engagement should be. In this regard, particular attention should be paid to involving external stakeholders and members of society (i.e., those who are not covered by the categories of "AI providers" and "AI users", as listed in Chapter III) who could potentially be adversely affected by the deployment of the AI system.

XIII Complementary elements relating to artificial intelligence in the public sector

54. As set out in Chapter VII, the development, design and application of AI systems in the public sector should be addressed in a legally binding transversal instrument, covering the most important transversal rights and obligations that should be respected in this domain. Additionally, the CAHAI is of the opinion that, given the context specificity of the risks posed by AI in the public sector in light of its specific role in society, such a transversal framework may be supplemented by additional legally binding or non-legally binding instruments at sectoral level.

55. These instruments could for instance elaborate further principles and requirements, specifically for the public services, regarding *transparency*, *fairness*, *responsibility*, *accountability, explainability,* and *redress* to ensure the responsible use of AI. The CAHAI recommends that the use and design, procurement, development and deployment of AI systems in the public sector is subject to adequate *oversight* mechanisms in order to safeguard *compliance* with human rights, democratic principles and the rule of law, and foster *public trust* by rendering the use of AI systems trustworthy, i.e *intelligible*, *traceable* and *auditable*.

56. Additionally, considering that the distinction between public and private sector involvement is often ambiguous, and considering the liability issues relating to the contracting out of public services to private actors any provisions applying to the design, development, and application of AI in the public sector should also apply to private actors that act on behalf of the public sector.

57. The CAHAI considers that the following elements relating to the design, procurement, development and deployment of an AI system by a public entity could, in addition to those elements

already described in Chapter VII, be addressed as part of a legally or non-legally binding instrument on AI in the public sector:

58. In the *design phase* of the system, the CAHAI is of the opinion that a legally or non-legally binding instrument could address how due consideration could be given to the analysis of the problem which the public entity intends to solve, in order to assess whether an AI system is the appropriate fit for the problem and, if so, which characteristics it should have. A legally or non-legally binding instrument could furthermore address the following issues: the data sets to be used for the AI system should be clearly identified, and the protection of such data and their origin respected. The design choices of the system should then be rendered explicit and documented. The intended users of the system, both civil servants and the public, as well as those potentially affected by the system should be involved early on, and their capabilities in using the AI system in question should be considered. An open and transparent co-design approach should be favoured. Finally, a human rights, democracy and rule of law impact assessment should be carried out to anticipate, prevent and mitigate potential risks. This also requires putting in place risk management and mitigation frameworks, which are relevant throughout all phases.

59. In the *procurement* phase, a thorough review of applicable legislation and policy measures in place should be conducted. Where necessary, public procurement processes should be adapted and public procurement guidelines for AI should be adopted, to ensure that procured AI systems comply with human rights, democracy and rule of law standards. A multidisciplinary and multi-stakeholder approach should be ensured in order to involve various perspective and angles, including those of vulnerable groups. Because public entities are responsible for the systems they adopt and apply, careful attention should be paid to the potential impact on public accountability.

60. During the particularly sensitive phase of *development* of the system, documentation and logging processes should be meticulously kept to ensure transparency and traceability of the system. Adequate test and validation processes, as well as data governance mechanisms should be put in place. Amongst other risks, the potential risk of unequal access or treatment, various forms of bias and discrimination, as well as the impact on gender equality should be assessed.

61. Risk management and mitigation frameworks set up in previous phases should be evaluated, adapted and maintained during the *deployment* phase. Taking into account the nature of the risk, human involvement may need to be guaranteed in order to ensure appropriate oversight over the system. Where appropriate, the AI system should be initially and regularly audited by an independent actor, and the results rendered publicly available to foster public trust. To this end, the CAHAI considers that the establishment of public registers listing AI systems used in the public sector, containing essential information about the system such as, its purpose, actors involved in its development and deployment, basic information about the model, and performance metrics, where appropriate, and the result of a HUDERIA, should be addressed in the context of a legally binding or non-legally binding instrument on AI in the public sector. In addition, the aforesaid instrument could address the establishment of a feedback mechanism in order to collect input on how to improve the system directly from its users and those potentially affected thereby. Further, the instrument could address the need for the AI system to be subjected to regular evaluation and update, including by taking into account the feedback. The evaluation process could be a periodic one. Transparency and communication towards users and citizens should likewise be addressed, as should the possibility of access to accountability and individual and collective redress mechanisms. Last but not least, the instrument should address the right of the public to be informed about the fact that they are

interacting with an AI system rather than a human being, as well as the right to interact with a human being rather than *only* an AI system, in particular when the rights and interests of individuals or legal persons can be adversely impacted. Legitimate exceptions to these rights may be foreseen by law where necessary and proportionate in a democratic society.

62. Finally, a legally binding or non-legally binding instrument on AI in the public sector could address measures to increase digital literacy and skills among both civil servants and the general public, notably through investment in capacity building (initial and continuous training and education) of public officials and awareness raising about the benefits, risks, capabilities and limitations of AI systems, and through enabling public interest research. Such skills should encompass theoretical as well as practical knowledge on the interplay between the design, development and application of AI systems on the one hand, and human rights, democracy and the rule of law on the other hand. Furthermore, the aforesaid instrument could also address the way in which these systems should be supervised and the risks arising therefrom should be managed.