COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 30 March 2021

CAHAI-LFG(2021)02
Restricted

# AD HOC COMMITTEE
# ON ARTIFICIAL INTELLIGENCE (CAHAI)

## Legal Frameworks Group (CAHAI-LFG)

### Outcome from Sub-Working Group 1

Prepared by CAHAI-LFG Sub-Working Group 1

www.coe.int/cahai

**LFG Subworking Group 1: Scope, Purpose, Definitions, Basic Principles, General Criteria for a Risk-Based Approach (relevant parameters, e.g. sector, use)**

*1st DRAFT*

**Preamble**

*Commentary: the preambulatory clauses of the future legal instrument are intended to reflect the overall purposes and considerations driving the conclusion of this legal instrument. Preambles of Conventions or Recommendations may clarify the context and the circumstances in which these instruments have been negotiated and adopted. They should remain concise and typically give the following information: a clause formally recalling the context of the instrument's adoption; the reasons for which it was adopted; the relationship with existing standards; the main objectives to be achieved. The exact wording and structure of the preambular part would of course depend on the nature of the legal instrument – namely, whether it will be a legally binding treaty (convention) or a non-binding act (recommendation of the Committee of Ministers). The following text is suggested for possible inclusion in the preambular part, with some optional elements suited for a binding instrument.*

[The member States of the Council of Europe, and the other signatories hereto,]

- Considering that the aim of the Council of Europe is to achieve greater unity between its members, based in particular on respect for human rights and fundamental freedoms, as well as democracy and the rule of law;

- Recognising the value of fostering co-operation with the other States parties to this Convention given the global nature of the challenges for the protection of human rights and fundamental freedoms, democracy and the rule of law that arise from the development, deployment and use of Artificial intelligence (AI) systems;[1]- Conscious that AI systems have the potential to promote human prosperity and individual and societal well-being by enhancing progress and innovation, but at the same time also raise new challenges and risks for human rights and fundamental freedoms as well as democracy and the rule of law, that need to be properly addressed;[2]- Underlining the particular need to ensure that discrimination and inequalities that persist in our societies are not perpetuated by AI Systems;[3]- Bearing in mind the Convention for the Protection of Human Rights and Fundamental Freedoms (ETS No. 5, 1950) and its Protocols, in the light of the relevant case law of the European Court of Human Rights, the European Social Charter (ETS No. 35, 1961, revised in 1996, ETS No. 163); and the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No.108, 1981, amended by its amending Protocol CETS No. 223);[4]

Have agreed as follows:]

---

1 *Motivation: Feasibility Study, Chapter 10, Rz.*

2 *Motivation: also underlined in the feasibility study, part 1: General introduction, p. 2.*

*3 To be discussed on LFG level, whether to keep or whether it might be too specific for implementing in preamble.*

4 *Suggestion from CEG to include reference to the "Council of Europe Convention on Preventing and Combating Violence against Women and Domestic Violence (CETS No. 210, Istanbul Convention, 2011)".*

## Chapter I – General provisions

### Article 1 - Object and purpose

The purpose of this [instrument] is to set a legal framework that ensures that AI systems are designed, developed and used in a way that guarantees full respect for all persons'[5] human rights and fundamental freedoms, for democracy and the rule of law. Each Party shall take in its internal law the necessary measures to give effect to the provisions of this [instrument].

### Article 2 - Definitions

### a) Definition "AI system"

*Commentary: There is no single definition of Artificial Intelligence ("AI") accepted by the scientific community. "AI" is used as a blanket term for various computer applications based on different techniques, which exhibit capabilities commonly and currently associated with human intelligence [rationality]. AI systems act in the physical or digital dimension by recording their environment through data acquisition, analysing certain structured or unstructured data, reasoning on the knowledge or processing information derived from the data, and on that basis decide on the best course of action to reach a certain goal. They can be designed to adapt their behaviour over time based on new data and enhance their performance towards a certain goal. AI systems should be defined in a technologically neutral (i.e. regardless of the underlying technology being used) way, comprising all the various automated decision-making technologies that fall under this umbrella term, including their broader socio-technical context. A simplified and technologically neutral definition of its purpose, covering those practices or application cases where the development and use of AI systems, or automated decision-making systems more generally, can impact on human rights, democracy and the rule of law, and taking into account all of the systems' socio-technical implications. To ensure greatest possible consistency the chosen definition should take into account existing definitions developed under other auspices, on a European, global and national level.*

*To sum up, the key elements of the CoE AI definition should be: Sufficient precision to identify and separate AI systems that we intend to regulate from other meanings of "AI"; Sufficient breadth to include various types of AI systems, also accounting for "black boxes"; Technological neutrality and future proofing, as far as possible; Simplicity, legal certainty and ease of practical application by stakeholders; "Compatibility" with other definitions[6] is a plus, but not an "end goal": after all, there are no international "AI standards" yet, and CoE is on track to become one of the first "standard-setting" body; "Permanence" whereas continuous updating should be avoided]*

---

*5 Suggestion from ALLAI (Catelijne Muller) to add „both individually as well as collectively"; Rationale: "AI impact' is to be considered both at individual and at societal/collective level whereas AI can impact both the individual as well as larger parts of our collective society" (Muller report "Impact of AI on Human Rights Democracy and the Rule of Law").*

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

6 According to the Feasibility Study (par.5, p.3), "The term, which has become part of everyday language, covers a wide variety of sciences, theories and techniques of which the aim is to have a machine reproduce the cognitive capacities of a human being." The following definitions have been taken into account when developing the AI definition: - "Default" definition used by the Council of Europe: "A set of sciences, theories and techniques whose purpose is to reproduce by a machine the cognitive abilities of a human being. Current developments aim, for instance, to be able to entrust a machine with complex tasks previously delegated to a human." (https://www.coe.int/en/web/artificial-intelligence/glossary).

- Definition proposed by the EU High-Level Expert Group on AI (HLEG) (https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines): "Artificial intelligence (AI) systems are software (and possibly also

**bb) Suggestion for CoE Definition of AI systems**

*Commentary: We provide 2 options to be discussed on LFG level. The 2 options are intended to represent the diversity between popular approaches to definitions of AI: either through a more detailed description of its functioning, or through a comparison to human intelligence/cognition.*

***Option 1:***

For the purposes of this [instrument]:

***"Artificial intelligence (AI) Systems" - machine based systems that act in the physical or digital dimension, perceiving their environment through the [acquisition of structured or unstructured data, analysing the data, reasoning on the knowledge or][7] processing and interpreting information [derived from the data][8] and formulating an output (recommendations, predictions or decisions), [9][to reach a given (set of) goal(s)][10] with a varying degree of autonomy and adaptiveness. [AI systems are not limited to systems with certain technical features. They can, e.g., either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by the systems previous actions. The term covers both stand-alone systems and any AI-based components embedded in larger systems. It covers data-driven as well as non-data driven systems (expert systems, knowledge reasoning and representation, reactive planning, argumentation, etc.).][11]"***

***Option 2*:**

For the purposes of this [instrument]:

---

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

- Definition proposed by OECD in its principles on AI (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449): "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."

- Definition proposed by OECD network of Experts on AI (One AI) (https://www.oecd.ai/network-of-experts): "An AI system is a machine-based system that is capable of influencing the environment by producing an output (recommendations, predictions or decisions) for a given set of objectives. It uses ma-chine and / or human-based inputs / data to perceive environments, abstract these perceptions into models and interpret the models to formulate options for outcomes / output. AI systems are designed to operate with varying levels of autonomy."

- Definition proposed by the UNESCO Ad Hoc Expert Group (AHEG) for a draft recommendation on the ethics of artificial intelligence (https://unesdoc.unesco.org/ark:/48223/pf0000373434): "AI systems as technological systems which have the capacity to process information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control. Three elements have a central place in this approach:

(a)AI systems are information-processing technologies that embody models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in real and virtual environments. AI systems are designed to operate with some aspects of autonomy by means of knowledge modelling and representation and by exploiting data and calculating correlations. AI systems may include several methods, such as but not limited to:

(i) machine learning, including deep learning and reinforcement learning,

(ii) machine reasoning, including planning, scheduling, knowledge representation and reasoning, search, and optimization, and

(iii) cyber-physical systems, including the Internet-of-Things, robotic systems, social robotics, and human-computer interfaces which involve control, perception, the processing of data collected by sensors, and the operation of actuators in the environment in which AI systems work."

7 ALLAI (Catelijne Muller) suggests to delete this part, since it focusses solely on data, while at this time ever more 'data-poor' AI systems are being researched. Also in the technical world 'reasoning' and 'interpreting' are understood solely in the realm of AI, while in the wider world, reasoning is a typically human exercise.

8 See previous footnote.

9

10 ALLAI (Catelijne Muller) suggests to delete this part, since not all AI is given an explicit goal. The trick here is to shape or translate the technical definition of AI into a legal definition, to guarantee legal certainty.

11 To be discussed on LFG level, whether such specification is needed.

***"Artificial intelligence (AI) Systems" - technological systems which have the capacity to process information and perform cognitive tasks in a way resembling human intelligence[12], including capacity for autonomous learning and decision-making. These systems typically include software and hardware elements, and typically exhibit aspects of reasoning, learning, perception, prediction, planning or control. [The term covers both stand-alone systems and any AI-based components embedded in larger systems. It covers data-driven as well as non-data driven systems (expert systems, knowledge reasoning and representation, reactive planning, argumentation, etc.).]***

**b) Further Definitions for CoE legal instrument on AI systems**

*Commentary: At this stage, we cut this sub-chapter to a minimum and suggest to further elaborate it at the end, when knowing the outcome submitted by the other subgroups (SGs 2-7) and the terms used therein. In any case, any definitions to be included should be technically neutral and future proofing as far as possible.*

**- AI Actors:** AI actors are those stakeholders in and of the private and public sector who play an active role throughout the entire AI System Lifecycle as defined in this Chapter, including AI suppliers (e.g., human content moderators, data labellers or persons working in data enrichment services more broadly)

- [**AI Operator:** Legal or physical person certified as the operator of an AI system by a national AI certification authority.[13]]

**- AI System Lifecycle:** AI system lifecycle phases involve: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection, processing, storing, training and labelling, as well as model building; ii) 'verification and validation'; iii) 'deployment and use'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.

**[- Certification State:** the State of the national AI certification authority which certified the AI operator.][14]

12 Questions posed by ALLAI (Catelijne Muller) to be discussed on LFG level: First of all, many (if not most) AI-systems do not resemble human intelligence. In the same way that an airplane does not resemble a bird. Again looking at it from an 'interpretation' angle, I am concerned of the granular discussions that could arise on what would be covered by the framework and what would not be. A party in a trial could easily state that due to the particular workings of the system, there is no resemblance with human intelligence. For example, an AI system is able to recognise a car when it is on a road, but will be completely confused when it flies in the air, floats in the water, etc. Humans would recognise the car without any problem. In fact, merely due to the workings of AI, many AI systems make mistakes that humans would never make. I would avoid any 'resemblance with human intelligence', also to avoid antropomorphising AI. Looking at it from another angle, the part on 'resembling human intelligence' could even over-include automated technologies such as the calculator.

13 Proposed by RUS;(reasoning: since "AI actors" appears to be too wide to properly identify persons holding legal responsibility/liability for AI systems' activities, ensure "single entry point for litigation" and otherwise protect the interests of the victim, as well as provide a link to risk assessment and oversight mechanisns). Proposal by other group members to use a more „function oriented" definition, and not to link it to the term „certification".

14 Proposal by RUS to accompany definition of AI operator. As no consensus was reached within the group, this is left to be discussed on LFG level, whether to keep this definition.

**Article 3 – Scope**

**1.** Each Party undertakes to apply this **[Name of the binding Instrument] [subject to its jurisdiction]** in the **public and private sectors**, thereby securing every **[Subject]**'s right to **[right/rights indicated in the title of the Legal Document]**.[15]

2. This [Instrument] shall not apply to military [or dual-use][16] AI systems[17] .

**Chapter II – Basic principles for the regulation of AI systems**

*Commentary: While SG 1 is mandated to draft the section on Basic Principles, SGs 2-7 are working on concrete principles and requirements. In order to avoid duplication, we suggest to return to this chapter and further elaborate and commit ourselves to concrete language after looking at the discussions in SGs 2-7.*

*However, this is the set of basic principles and first drafting proposals SG 1 suggests:*

**Article 4 - General provisions**

Each Party shall take the necessary measures in its law to ensure that the design, development, and use of AI Systems is compatible with the Council of Europe standards on human rights (including economic and social rights), democracy and the rule of law. AI systems that interact with vulnerable groups must be designed, developed and used respectful to their rights. AI systems that interact with children must be designed, developed and used respectful to children's rights, be child-centered and transparent in a way that children and / or their caregivers can understand the interaction.[18]

In particular, the Parties shall take the necessary measures to provide the following:

**Article 5 - Protection of Human Dignity**

Each Party shall provide that the design, development and use of AI Systems respect the dignity of the human beings interacting with or impacted by the AI system.

**Article 6 - Prevention of harm and principle of precaution**

The Parties shall take the necessary measures to ensure that adequate safeguards are put in place to minimise and prevent harm stemming from the design, development and use of AI Systems in both the individual and collective dimension concerning the negative impact on human rights, democracy and the rule of law, whether this concerns physical or psychological harm, economic, environmental, social or legal adverse consequences, providing additional safeguards for persons and groups who are more vulnerable.

**Article 7 - Non-discrimination, (Gender-)Equality, Fairness and Diversity**

The Parties shall provide that the design, development and use of AI Systems respect the right to non-discrimination, equal treatment and equality [before the law[19]]. [The right to non-discrimination

---

15 These provisions will have to be further elaborated in order to make them more legible at a later stage.

16 Including "dual-use" AI systems proposed by RUS (reasoning: to include AI systems that can be used for military purposes). As no consensus was reached within the subgroup, to be further discussed on LFG level.

17 Other exclusions may be added here, perhaps via individual declarations.

18 2nd half sentence has been controversially discussed by SG 1. RUS: This requirement is not present in the feasibility study and seems too far-reaching (as, for instance, not all children might even be able to understand interactions with AI).

19 Addition „before the law" proposed by RUS with following reasoning: The ECHR or its Protocols do not contain a "right to equality". Protocol 12 refers to a "fundamental principle according to which all persons are equal before the law and are entitled to the equal protection of the law". With regard to non-discrimination, the Court and the Committee of Ministers commonly refer to "equality before

underlying this [Instrument] shall extend to all differentiation grounds that can lead to direct or indirect discrimination, including intersectional discrimination.][20]

The parties should encourage gender balance and diversity in the AI workforce and periodic feedback from a diverse range of stakeholders.

**Article 8 - Transparency and Explainabililty**

The Parties shall ensure that the design, development and use of AI Systems meets minimum standards of explainability and traceability as further determined under Chapter [•] of this [Instrument] that are necessary to enable the right holders to effectively protect their rights.

**Article 9 - Protection of Human Freedom and Human Autonomy, including the Protection of Personal Data**

a) The Parties must ensure that individuals and society can decide in an informed and autonomous manner on the use of an AI systems and on their consequences.

b) The Parties must provide that human oversight mechanisms are established throughout the entire lifecycle of an AI System, ensuring that human intervention is possible whenever needed to safeguard human rights, democracy and the rule of law.

c) The Parties must ensure to individuals and groups the right to effectively contest and challenge decisions informed and/or made by an AI system and demand that such decision be reviewed by a person (right to opt out).

**Article 10 - Accountability, responsibility, liability**

**[•]**

**Article 11 - Human Oversight**

**[•]**

**Article 12 - Cooperation**

**[•]**

**Chapter III - General criteria for a risk-based approach[21]**

*Commentary: The risks of AI systems to human rights, democracy and the rule of law depend on the application context, technology and stakeholders involved. To counter any stifling of socially beneficial AI innovation, and to ensure that the benefits of this technology can be reaped fully while adequately tackling its risks for human rights and fundamental freedoms as well as democracy and the rule of law, a Council of Europe legal framework on AI should pursue a risk-based approach targeting the specific application context. Such risk-based approach should follow the principle that the greater the potential of an AI system to negatively affect human rights, democracy and the rule of law, the more stringent*

---

the law", „equal protection of the law" and "right to equality of arms in crim-inal proceedings". Likewise, Article 14 of the ICCPR covers "right to equality before courts and tribunals and to a fair trial".

20 This sentence has been controversially discussed by SG 1 members; For further background and information SG 1 refers to SG 3.

21 SG 1 is expected to identify and determine the parameters / criteria that are relevant for determining the risk level of an AI system, e.g. sector in which the AI sector is used, rights and (number) of people affected; impact on the society as w whole). SG5, on the other hand, shall elaborate on the procedures to assess this risk, i.e. whether the assessment shall be done through self-assessment, whether external audits are needed and how this does impact questions of accountability). The CAHAI-PDG SG 1 is responsible for the development of a HRIA methodology.

*the requirements and the more far-reaching the intervention by means of regulatory instruments, which might include a moratorium or a ban as a last resort.*

**Article 12 – Risk-based Approach**

This [Instrument] pursues a risk-based approach defining graduate requirements for the design, development and use of AI Systems that shall depend on the potential impact of the application of an AI System on human rights, democracy and the rule of law.

The Parties shall implement the necessary procedural measures defined under [Chapter 5] of this [Instrument] for an effective assessment of the impact of the application of an AI system on human rights, democracy and the rule of law[22], taking into account the following assessment criteria:

a) the **likelihood of a negative impact** on human rights, democracy or rule of law;

b) the **severity of the impact, including its scale**, relating *inter alia* to:

i. the gravity of the negative impact;

ii. the number of people and characteristics of groups that likely to be affected;

iii. its geographical and demographic reach;

iv. its temporal extension;

v. the extent to which the potential adverse effects are reversible;

vi. the interrelatedness of human rights and possible simultaneous impacts of an AI system on more than one protected right and freedom.[23]

vii. The likelihood of exacerbating existing biases, stereotypes, discrimination and inequalities with respect to protected grounds of discrimination and segments of the population in vulnerable situations";

c) When assessing the potential adverse impact, the following aspects are also taken into consideration:

i. AI-specific factors increasing the risk level, such as complexity of the used AI system;[24] its level of automation;[25]

ii. the sector and area of use / further context of AI system / Purpose of the AI systems use

iii. the level of compliance with other legislation;

iv. the quality, type and nature of data used;

v. any measures that are deployed to mitigate the potential harm to human rights, democracy and the rule of law;

---

[22] Risk assessments should not be considered in any way as a substitute for human right due diligence.

[23] Included in a Guide on human rights impact assessments, issued by the Danish institute for human rights, available at https://www.humanrights.dk/sites/humanrights.dk/files/media/document/DIHR%20HRIA%20Toolbox_Welcome_and_Introduction_ENG_2020.pdf, also with examples.

[24] See e.g. deep learning systems; When AI systems operate without significant and effective human control, a higher level of protection for human rights, democracy and the rule of law must be provided.

[25] See e.g. AI systems with no human intervention.

vi. the dependence of potentially affected parties on the decision of the AI system, including their ability to change this system for another or not to be exposed to its effects.[26]

vii.

d) In assessing the impact of an AI system, its **positive impact**[27] in strengthening human rights, democracy and the rule of law must be considered.

**The following risk levels could be established**

*a) No risk applications*

AI Applications with zero or negligible risks for human rights, democracy and the rule of law.

*b) Low risk application*

AI Applications with some potential for harm with regard to human rights, democracy and the rule of law.

*c) High risk applications*

AI Applications with significant risks for human rights, democracy and the rule of law.

*d) Untenable risk applications*

AI Applications with extreme (untenable) risks and therefore inherently incompatible with human rights, democracy and the rule of law.

*[Commentary: In addition to the above gradation, a special risk category may be suggested:*

*e) Uncertain risk applications*

Where the negative impact of AI applications on human rights, democracy and the rule of law is likely, but its extent is unclear precautionary measures can be adopted to reduce exposure to risk.]

26 Switchability: The ability to change the AI system for another (e.g. by switching the operator) or avoid being exposed to an AI decision altogether. In the worst case, the people effected do not have the ability to opt-out of using specific services without facing societal repercussions (e.g. health care, financial market)

27 When assessing risks and possible measures to mitigate those risks, we should also assess the risks caused by these measures ("cost-effectiveness analysis") as well we their strict necessity and proportionality. An example would be anti-COVID measures, that may negatively affect some human rights, but are still necessary for the protection of human rights on a broader scale (so they might represent a "risk" in this sense, but their absence would create an even greater "risk").

## Sub-Working Group 1 (version with comments)