Strasbourg, 28 November 2024

CAI(2024)16rev2

COMMITTEE ON ARTIFICIAL INTELLIGENCE (CAI)

METHODOLOGY FOR THE RISK AND IMPACT ASSESSMENT OF ARTIFICIAL INTELLIGENCE SYSTEMS FROM THE POINT OF VIEW OF HUMAN RIGHTS, DEMOCRACY AND THE RULE OF LAW (HUDERIA METHODOLOGY)

www.coe.int/cai

Table of Contents

Introduction

What is the HUDERIA?

The risk and impact assessment of artificial intelligence (AI) systems from the point of view of human rights, democracy and the rule of law ("the HUDERIA") is a guidance which provides a structured approach to risk and impact assessment for AI systems specifically tailored to the protection and promotion of human rights, democracy and the rule of law. It is intended to play a unique and critical role at the intersection of international human rights standards and existing technical frameworks on risk management in the AI context.

The HUDERIA can be used by both public and private actors to aid in identifying and addressing risks and impacts to human rights, democracy and the rule of law throughout the lifecycle of AI systems.

The HUDERIA originates in the work of the *Ad Hoc* Committee on Artificial Intelligence (CAHAI) (2019-2021) and specifically its Policy Development Group, which mandated the Alan Turing Institute, the UK's national institute for data science and AI, to prepare an original proposal operationalising the outline of a Model for a Human Rights, Democracy and the Rule of Law Impact Assessment. The HUDERIA Methodology was adopted by the Committee on Artificial Intelligence (CAI) of the Council of Europe on 28 November 2024.

Relationship to the Framework Convention

The HUDERIA is a stand-alone, non-legally binding guidance that, as such, does not have legal effect. It is not mandatory, nor intended as an interpretive aid for the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, hereinafter referred to as "the Framework Convention". Many existing or future frameworks, policies, guidance, standards or tools may be used to assist in conducting AI risk and impact management, including the HUDERIA.

Parties to the Framework Convention have the flexibility to use or adapt the guidance, in whole or in part, to develop new approaches to risk assessment or to use or adapt existing approaches in keeping with their applicable laws, provided that Parties fully meet their obligations under the Framework Convention, including, in particular, the baseline for risk and impact management set out in its Chapter V.

Principal objectives of the HUDERIA

The aims of the HUDERIA are:

- to help determine the extent to which risk management activities related to human rights, democracy and the rule of law may be called for, and to offer a methodology for risk and impact identification, assessment, prevention, and mitigation that is applicable to a variety of AI technologies and application contexts and is responsive to future developments in AI technologies and applications;

- to promote compatibility and interoperability with existing and future guidance, standards and frameworks developed by relevant technical, professional and other organisations or bodies (such as ISO, IEC, ITU, CEN, CENELEC, IEEE, OECD, NIST),

including the NIST AI Risk Management Framework and risk management and fundamental rights impact assessment under the EU AI Act.

What is the approach of the HUDERIA?

The HUDERIA combines contemporary knowledge about the technical and socio-technical governance processes and mechanisms that can facilitate the responsible activities within the lifecycle of AI systems with the diligence procedures needed to protect and promote human rights, democracy and the rule of law.

The HUDERIA takes as a basis well-known variables, concepts and language for the assessment of risks to human rights (scale, scope, probability and reversibility of potential adverse impacts on human rights). It aims to facilitate their examination by providing additional guidance in view of the socio-technical complexity of the AI lifecycle.

Socio-technical approach

The HUDERIA adopts a socio-technical approach, which views all aspects of the AI system lifecycle as affected by the interconnected relationship of technology, human choices, and social structures. In this approach, risk and impact management of AI systems takes account of both their technical aspects and the legal, social, political, economic, cultural and technological contexts in which they operate. Such approach promotes the development of safe, secure and trustworthy AI that is both performant and promotes respect for human rights, democracy and the rule of law.

General and specific guidance

The HUDERIA offers structure by combining general and specific guidance and flexibility by allowing room for adaptation in the practical implementation.

At the general level, the **HUDERIA Methodology** describes high-level concepts, processes and elements guiding risk and impact assessment activities of AI systems that could have impacts on human rights, democracy and the rule of law.

At the specific level, the **HUDERIA Model**[1] will provide supporting materials and resources (such as flexible tools relevant for different elements of the HUDERIA process and scalable recommendations) that can aid in the implementation of the HUDERIA Methodology. These resources are referred to throughout the text and will provide a library of knowledge that can facilitate consideration of risks and impacts related to human rights, democracy, and the rule of law, including in other approaches to risk management.

Adaptability and flexibility

Both the HUDERIA Methodology and HUDERIA Model allow room for adaptation to different contexts, needs and capacities by setting goals, principles and objectives, while leaving a margin of appreciation to decide how to meet them and offering a range of policy and governance options that can be tailored to contexts.

Graduated and differentiated approach

The HUDERIA aims to establish a graduated and differentiated approach to measures for risk and impact identification, assessment, prevention and mitigation that takes into account the severity and probability of the occurrence of the adverse impacts on human rights, democracy and the rule of law as well as relevant contextual factors.

[1] to be elaborated and adopted by the CAI in 2025

Outline of the HUDERIA

The HUDERIA Methodology contains four elements:

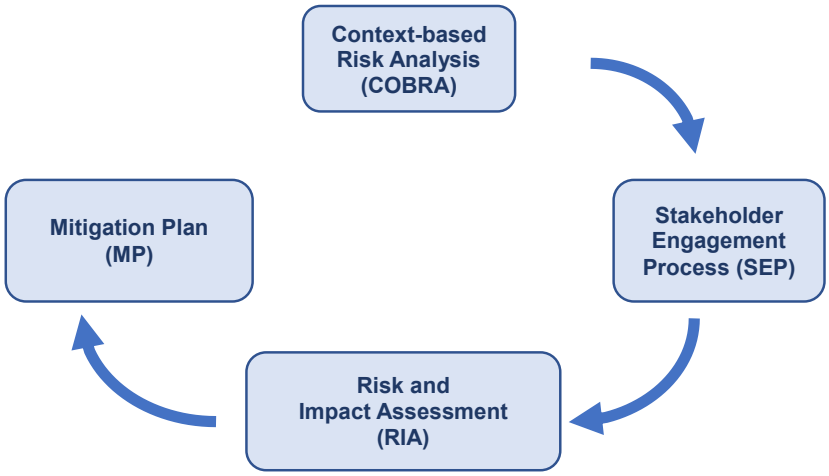
1. the **Context-Based Risk Analysis** (COBRA) provides a structured approach to collecting and mapping the information needed to identify and understand the risks the AI system could pose to human rights, democracy and the rule of law in view of its socio-technical context. It also supports an initial determination as to whether the AI system is an appropriate solution for the problem being considered;

2. the **Stakeholder Engagement Process** (SEP) proposes an approach to enabling and operationalising the engagement, as appropriate, with the relevant stakeholders in order to gain information regarding potentially affected persons and contextualize and corroborate potential harms and mitigation measures;

3. the **Risk and Impact Assessment** (RIA) provides possible steps regarding the assessment of the risks and impacts related to human rights, democracy and the rule of law;

4. the **Mitigation Plan** (MP) provides possible steps on defining mitigation and remedial measures, including access to remedies and iterative review.

While it is logical to carry out the COBRA element first, depending on the needs and approaches, one may choose to change the sequence of the elements and/or apply or otherwise use only certain parts of the methodology based on existing AI governance approaches and specific contexts, needs and capabilities.



I. The Context-Based Risk Analysis (COBRA)

Introduction

The COBRA assists in the identification of different risk factors - characteristics or properties of an AI system and its context that affect the probability of adverse impacts on human rights, democracy, and the rule of law. These factors are not necessarily to be treated as causes of adverse impacts but rather as conditions that are correlated with an increased chance of harm and therefore need to be anticipated and considered in risk management and impact mitigation efforts. The risk factors are categorised into three broad areas: the system's application context, its design and development context, and its deployment context[2].

The examination of the risk factors is intended to facilitate the mapping of potential adverse impacts on human rights, democracy, and the rule of law. The results of this risk factor and impact mapping analysis are intended to inform the extent of the approach to subsequent elements of the HUDERIA, including by establishing the proportionality of subsequent HUDERIA activities.

The results of this risk factor and impact mapping analysis may also help pinpointing the specific socio-technical contexts across the system’s lifecycle that need focused governance attention.

The COBRA element consists of four steps:

1) Preliminary scoping;
2) Analysis of risk factors;
3) Mapping of potential impacts on human rights, democracy, and the rule of law;
4) Triage.

Preliminary scoping

Objectives

The main purpose of this stage is to carry out the preliminary background research needed to inform subsequent risk factor identification and impact mapping activities.

Explanations

The COBRA process begins with preliminary scoping research that outlines the purpose of the system, key components of the system, the contexts in which it is intended to be used, the area/domain(s) in which it will operate, the degree of human intervention, and the nature and amount of data it will process and on which it will be trained, noting any checks that may have already been done to assess bias in the dataset or model, identifies persons or groups who may be affected by, or may affect, the system, focusing on the relevant contextual characteristics of identified persons and groups including protected characteristics and vulnerability factors, provides a preliminary scoping of potential adverse impacts on human

[2] See page 9 for a detailed explanation of the areas

rights, democracy and the rule of law by exploring the illustrative areas of concern[3]; and provides an initial mapping of roles and responsibilities across the AI system's lifecycle[4].

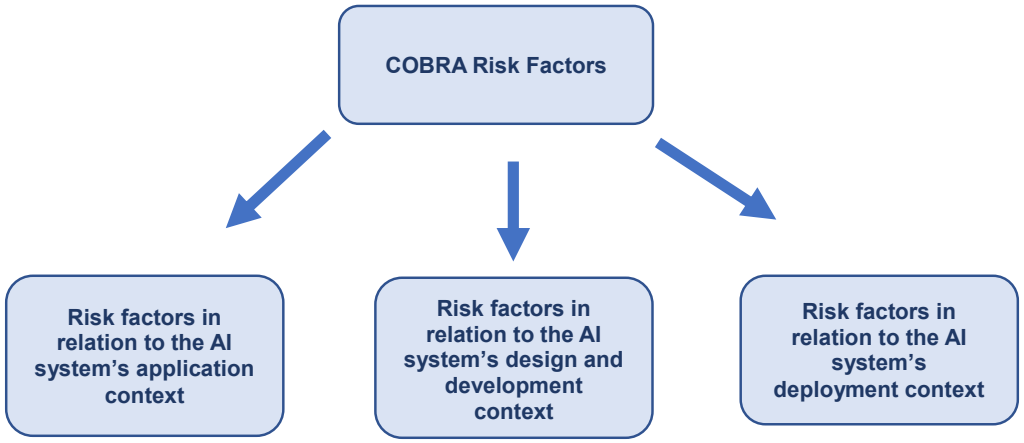
This preliminary scoping activity could draw on organisational documents (i.e. the project business case, proof of concept, or project charter), collaboration, and desk research (if necessary). This preliminary scoping activity, and subsequent elements of the HUDERIA process, should take place, as appropriate, in a multidisciplinary team, consisting of experts with a range of complementary specialisations[5] and both technical and non-technical backgrounds.

Analysis of risk factors

Objectives

The main purpose of this stage is to collect the relevant information about risk factors related to the system's intended application context, design and development context and deployment context. These risk factors will facilitate the mapping of potential adverse impacts on human rights, democracy and the rule of law and the subsequent assessment of key risk variables: severity (scale, scope and reversibility)[6] and probability.

Explanations



AI systems are designed, developed, and used in a wide variety of contexts and in numerous different ways, making it important to holistically assess various factors related to the system's application context, its design and development context and its deployment context.

[3] **COBRA Resources E (Illustrative areas of concern from the point of view of Human Rights, Democracy and the Rule of Law)** [to be elaborated and adopted by the CAI in 2025] provides a tool which could be used to perform or inform this assessment.
[4] The **Roles and Responsibilities Section** in the HUDERIA Model will provide guidance in connection with this aspect of the Methodology.
[5] Relevant domain expertise may include, as appropriate, issues of human rights, privacy and personal data protection, data science, data set management, security, AI risks, and AI testing, evaluation, verification and validation.
[6] Following the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, the UN Office of the High Commissioner for Human Rights, and UN Guiding Principles on Business and Human Rights, for the purposes of the HUDERIA the term "severity" is understood to be composed of a combination of the variables of scale, scope, and reversibility.

The AI **system's application context**[7] includes information about the system's application sector and domain, the legal and regulatory environments in which the system is being developed and used, the system's intended purpose, and other relevant details of the system's application context, such as any known legacies of bias of discrimination.

The AI **system's design and development context**[8] includes the relevant technical characteristics of the system. This may include known limitations of the system, considerations related to data collection, enrichment, storage, use, and retirement; and considerations related to the algorithm or model itself. Particularly relevant considerations include technical characteristics related to privacy and data protection, bias and discrimination, and explainability and interpretability.

Lastly, the AI **system's deployment context**[9] includes factors that govern how potential risks may manifest and be managed in practice, such as steps that will be taken to protect privacy and personal data, mitigate harmful bias, ensure proper training, guard against unintended uses, and ensure accountability and legal compliance.

Mapping of potential impacts on human rights, democracy and the rule of law

Objectives

The mapping step identifies potentially affected persons or groups and makes the initial assessment of the key risk variables - severity (scale, scope, reversibility) andprobability . The mapping helps to inform subsequent elements of the methodology, and the extent of governance intervention and mitigation measures that may be appropriate (see Triage section below). The analysis of the key risk variables is crucial for a clear, structured overview of where threats are most likely to occur and their potential impact.

Explanations

The **COBRA Resources E**[10] and **F**[11] could be used to identify potentially sensitive application areas and potentially relevant areas of concern related to human rights, democracy, and the rule of law[12].

Using the information collected in previous steps:

(a) determine whether the system will operate in proximity to activity(ies) (decision-making or actions) which may produce impacts on affected persons in the relevant sectors/domains[13];

[7] **COBRA Resources A (List of Risk Factors Arising in the System's Application Context)** [to be elaborated and adopted by the CAI in 2025] provides a tool which could be used to perform or inform this assessment.
[8] **COBRA Resources B (List of Risk Factors Arising in the System's Design and Development Context)** [to be elaborated and adopted by the CAI in 2025] provides a tool which could be used to perform or inform this assessment.
[9] **COBRA Resources C (List of Risk Factors Arising in the System's Application Context)** [to be elaborated and adopted by the CAI in 2025] provides a tool which could be used to perform or inform this assessment.
[10] [To be elaborated and adopted by the CAI in 2025]
[11] [To be elaborated and adopted by the CAI in 2025]
[12] References in the **COBRA Resources** to human rights as set forth in various international human rights instruments are included for illustrative purposes. Those references only apply to States which are Parties to those instruments. Each State is expected to apply its own applicable laws and international obligations.
[13] **COBRA Resources F** could be used to identify potentially sensitive sectors/domains.

(b) identify and enumerate the relevant areas of concern[14] and, with this in mind, answer the question of whether the system may have potential or actual impacts on specific human rights, democracy, and the rule of law;

(c) for each potential and actual impact identified, describe the nature of the potential and actual impact[15], taking account of differential impacts on affected persons and groups given relevant contextual characteristics including protected characteristics and vulnerability factors;

Analysing these points will provide information for the initial assessment of key risk variables - severity (scale, scope, reversibility) and probability - that assist in determining risk and choosing the right approach to the subsequent elements of the methodology, which will help ensure that governance interventions and mitigation measures are aligned with the needs throughout the AI system lifecycle.

The results of this analysis can also help identify opportunities for using the AI system to support positive actions that advance human rights, including promoting and ensuring non-discrimination.

Determination of Risk Level

The following variables may be employed to index the risk level of each of the potential adverse impacts to human rights, democracy and the rule of law that have been identified in as the result of the mapping exercise:

1. The **scale**[16] of the potential adverse impacts (i.e. the seriousness of the potential harm);

2. The **scope** of the potential adverse impacts (including the number of persons affected, the protected characteristics or vulnerability of individuals or groups and the timeframe of the impacts);

3. The **reversibility**[17] of the potential adverse impacts is the information on the possible reparability or restoration for affected persons to their pre-impact situation or equivalent.

4. The **probability**[18] of the potential adverse impacts.

Relevant teams should go through each of the potential impacts that have been identified and consider for each area of concern related to human rights, democracy and the rule of law, and each affected group, the scale, scope, reversibility and probability of the potential or actual adverse effect. Domestic law or policy may provide more detailed definitions that can be used to inform this determination of risk and to determine suitable and proportionate approaches to subsequent HUDERIA activities (e.g., stakeholder engagement).

Consideration may be given to establishing a method for combining these variables to enable the calibration of risk and the determination of suitable and proportionate approaches to

[14] **COBRA Resources E** could be used to identify potential areas of concern related to human rights, democracy and the rule of law.
[15] For clarity of assessment both adverse or restrictive impacts as well as beneficial, enhancing or otherwise positive impacts produced by AI systems should be accounted for, since various issues with bias and discrimination may arise in respect of systems that produce both types of impacts.
[16] The term “scale” may sometimes be referred to as “gravity” in risk assessment contexts.
[17] The term “reversibility” may sometimes be referred to as “remediability” in risk assessment contexts.
[18] The term “probability” may sometimes be referred to as “likelihood” in risk assessment contexts.

subsequent HUDERIA activities (e.g., stakeholder engagement) as well as the extent and depth of downstream governance interventions and risk management and mitigation measures. This may involve the formulation of quantitative or semi-quantitative methods of risk calculation, risk matrices, or more qualitative or rules-based procedures. Any risk calibration mechanism resulting from the combination of these variables for the purposes of the HUDERIA assessment may take into account:

- regarding human rights, low scope and high gravity effects as well as high scope and low gravity effects on each affected person;

- regarding democracy and the rule of law, the mechanism could take into account in particular high scope and long-lasting effects on persons, institutions and the society in general.

Triage

Objectives

The main purpose of this stage is to build on the information collected in previous COBRA activities, and therefore:

- to facilitate the task of identifying and triaging systems that pose significant risk, so that the HUDERIA Methodology is not onerous for minimal or low risk AI systems;

- to make an initial determination of whether the AI system should be developed or deployed, based on whether the benefits of developing or deploying the AI system outweigh its risks, particularly given its potential impacts on human rights, democracy and the rule of law, as well as whether the use of the AI system is incompatible with respect for human rights, democracy and the rule of law.

Adaptable approach to triaging

The prior activities in this stage provide a preliminary picture of the risk profile of the AI system.

The information gathered may, for example, be sufficient to determine that the system is unlikely to have any impact on human rights, democracy, or the rule of law, making the subsequent elements of HUDERIA unnecessary. A similar conclusion could be reached if the identified impacts are insignificant or unlikely. If the identified impacts lead to a decision that an AI system will not be developed or deployed because it is considered incompatible with respect for human rights, democracy and the rule of law, subsequent elements of HUDERIA are also unnecessary. Finally, in cases where serious potential impacts are identified, a range of risk management strategies and responses (including the **Stakeholder Engagement Process** described in the next section) may be justified. To address this complexity, the HUDERIA does not prescribe detailed guidance for adjusting risk management efforts, but simply sets out proposed elements that may be applied as appropriate based on the risk of adverse impacts to human rights, democracy and the rule of law. Different approaches to determining risk management steps based on the potential and actual impacts identified - or a combination of them - may be applied based on the specific domestic regulatory framework or environment, industry, system, and context (e.g., threshold-based, scenario-based, proportionality, dynamic, or context-specific approaches).

The final determination of whether to use a qualitative, quantitative, mixed or any other method is left to the discretion of the authorities or, where applicable, the AI project teams responsible for the system.

‘Zero questions’

To help determine whether the benefits of building or deploying the AI system, including additional social benefits that may result from the use of the system beyond its primary purpose, outweigh its risks given the risks factors and potential impacts identified, consider:

- whether the use of the system is appropriate considering the nature of the problem that the AI system is trying to solve;- the extent to which existing technologies and processes already in place to solve the problem under consideration are better placed to do so, considering the risk profile and potential adverse impacts of the prospective system with a particular focus, where appropriate, on any marginal risk added by introducing AI into the current context;

- the extent to which the prospective system will be able to meet the deployer’s needs and expectations;

- the extent to which the impacts of the prospective system will be equitable across affected groups;

- the extent to which the quality and representativeness of currently or potentially available data is sufficient for the prospective system to be effective, safe, and reasonably avoid potential harmful bias;

- the extent to which sufficient resources (human and material) are available and able to meet technical requirements and perform technical and governance actions to adequately mitigate identified risks; and

- the system’s potential use contexts and risks for misuse or abuse, including through deployment beyond its intended purpose.

II. The Stakeholder Engagement Process (SEP)

Introduction

The possibility to run this step can be considered in order to improve the quality of information for the next element of the HUDERIA - **the Risk and Impact Assessment** - by incorporating the views of identified potentially affected persons, including those in vulnerable situations.

Stakeholder engagement, as set out in the HUDERIA Methodology, may take various forms. The level of participation of affected persons should be informed by the risks factors and potential and actual impacts identified as part of the COBRA stage. Involving stakeholders throughout the AI system lifecycle can also offer a variety of additional benefits, such as fostering transparency, building trust and improving usability and performance of the AI system.

Explanation

The SEP involves five key steps[19]: Stakeholder Analysis, Positionality Reflection, Establishment of Engagement Objectives, Determination of Engagement Method, and Implementation.

Stakeholder analysis

The stakeholder analysis identifies stakeholder groups which may be affected by, or may affect, the activities within the lifecycle of the system. Such analysis[20] assesses the relative interests, rights, potential and existing vulnerabilities and advantages of identified stakeholders as well as the salience of identified stakeholder groups. At this step, consider meaningfully including the views of those who:

1) are disproportionately at risk from the use of the system;
2) are particularly vulnerable to potential harms; or
3) have particularly limited ability to influence how the system is designed and used (e.g., currently or historically marginalised, disadvantaged or underrepresented groups or persons in situations of vulnerability or presenting specific needs).

Positionality reflection

The next step involves reflection on the positional standpoint *vis-à-vis* affected stakeholders with a view to recognising the limitations of HUDERIA users' perspectives and identifying missing viewpoints which would strengthen the assessment of the system's potential and actual impacts.

[19] The process described in this section is illustrative in nature with the final determination on the process of stakeholder engagement being up to the discretion of the authorities or, where applicable, the AI project teams responsible for the system.

[20] **SEP Resources A (List of Questions to Assess Relative Stakeholder Salience)** [to be elaborated and adopted by the CAI in 2025] provides detailed questions and tools to guide the identification of relevant stakeholders.

Depending on the relevant risk factors and potential impacts identified at the COBRA element, this may include an assessment of HUDERIA users' self-identified demographics, education and training, socioeconomic background, and the institutional and team context.

The main questions on which HUDERIA users should reflect when undertaking this stage of the methodology are:

- To what extent do my personal characteristics, group identifications, socioeconomic status, educational, training and work background, team composition, and institutional frame represent sources of power and advantage or sources of marginalisation and disadvantage?

- How does this positionality influence my and my team's ability to identify and understand affected stakeholders and the potential impacts of the AI system?

Depending on the risk factors identified during the COBRA process, HUDERIA users should also consider engaging external stakeholders or consultants with specific expertise, such as human rights law expertise, related to the system's potential and actual human rights impacts.

Establishment of Engagement Objectives

Setting clear objectives for stakeholder engagement aims to create a clear understanding of how and why engagement activities are being conducted. These facilitate the inclusive, informed and meaningful involvement of potentially affected persons[21].

Determination of Engagement Method

Determining the appropriate stakeholder engagement method(s)[22] necessitates evaluation and accommodation of the needs of potentially affected persons, taking into consideration, as appropriate, the outcomes of the COBRA process and other relevant factors such as resource constraints, difficulties in reaching isolated or socially excluded groups, capacities constraints such as challenges resulting from digital divides or information gaps, timeframes, etc.

The following criteria may serve as guidance in the SEP element:

1) **engagement** - meaningful involvement of affected or potentially affected persons is integrated during the relevant elements of the process;

2) **equality and prohibition of discrimination** - engagement and consultation processes are inclusive, gender-sensitive, and account for the needs of persons and groups with protected characteristics or who may be at risk of vulnerability or marginalisation;

3) **empowerment** - consideration of age-appropriateness and accessibility needs, and capacity building of persons and groups with protected characteristics or who may be at risk of vulnerability or marginalisation is undertaken to ensure their meaningful involvement;

4) **transparency** - provide for the sharing of meaningful and intelligible information between stakeholders at relevant and regular intervals, make available information about the AI system to participating stakeholders that is adequate for giving a comprehensive

[21] **SEP Resources B** [to be elaborated and adopted by the CAI in 2025] provides indicative detailed questions and the description of options for stakeholder engagement.

[22] **SEP Resources C (Examples of Relevant Engagement Methods with Relevant Questions)** [to be elaborated and adopted by the CAI in 2025] provides possible examples of relevant engagement methods and a list of relevant questions that can aid determination of appropriate stakeholder groups.

understanding of potential implications and human rights impacts, where appropriate, publicly communicate HUDERIA findings and impact management plans (action plans); and

5) **accountability** - responsibility for the implementation, monitoring and follow-up of mitigation measures is assigned to particular entities, individuals or functions within the organisation.

Implementation

Having completed the prior four activities, the appropriate engagement processes can be performed. It should be consistent with the results of the stakeholder analysis, positionality reflection and established engagement objectives and methods, and be appropriately documented.

III. The Risk and Impact Assessment

Introduction

The purpose of the Risk and Impact Assessment is to provide detailed evaluations of the potential and actual impacts which the activities within the lifecycle of an AI system could have on human rights, democracy and the rule of law.

In accordance with the triage made in the COBRA step, carrying out the Risk and Impact Assessment is particularly important for AI systems that may pose significant risks to human rights, democracy and the rule of law. Following the triage of the COBRA analysis, this step of the processes may be needed only for certain AI systems, in particular those assessed as posing significant risks to human rights, democracy and the rule of law.

The Risk and Impact Assessment aims at:

- re-examining, contextualising and corroborating the potential and actual harms identified in the COBRA;

- identifying and analysing further potential and actual harms by engaging in extended reflection to pinpoint gaps in the completeness and comprehensiveness of the previously enumerated harms;

- evaluating the risk variables of scale, scope, reversibility and probability of the potential adverse impacts, so that their risks can be better assessed to be subsequently prioritised, managed and mitigated;

The Risk and Impact Assessment builds upon the initial identification of the context-based risk factors to human rights, democracy, and the rule of law and the mapping of potential impact on human rights, democracy and the rule of law carried out in the COBRA and the potential insights from the SEP to address the potential and actual impacts of the AI system.

This is done meaningfully through a two-step process that enables the formation of a Mitigation Plan and establishment of Access to Remedies at the next step of the HUDERIA Methodology.

Explanations regarding the Risk and Impact Assessment questions and prompts

Introduction

The Risk and Impact Assessment in the context of the HUDERIA is organised in two steps.

At the first step, the focus is identifying potential impacts and, more specifically, “how” the potential and actual impacts identified at the COBRA and eventually SEP steps could occur, enabling a more open-ended and exploratory approach that allows for deeper analysis of the specific contexts, scope, scale and reversibility of impacts, particularly concerning individuals in vulnerable situations or vulnerable groups.

At the second step, the assessment of the risk variables of scale, scope, reversibility and probability of potential or actual impacts identified takes place. A thorough context-responsive consideration of these variables helps prioritise mitigation actions by differentiating the severity of AI system impacts.

Scale

The scale of a potential and actual adverse impact refers to the seriousness of the potential harm's expected consequence.

Consideration of the gravity of any potential harm should include reflection on the different ways and different extents to which persons or groups (in particular, those who possess characteristics that could make them more vulnerable to the adverse impact) could suffer that harm.

Deliberations on scale should consider the following additional questions:

1) For each potential and actual adverse impact identified, are there persons or groups who possess characteristics that could make them more vulnerable to the impact? If so, what are these characteristics and could those who possess them suffer the harm more acutely or seriously than others?

2) For each potential and actual adverse impact identified, which persons or groups could encounter the gravest impact from the harm under consideration?

Responses to these questions will subsequently serve an important function during the mitigation planning stage when the redress and prioritisation of potential harms are under consideration.

Scope

The scope of a potential and actual adverse impact refers to the estimation of both the number of affected persons and of the timeframe of the impacts.

The estimations of scope for identified potential and actual adverse impacts are analysed one by one with special consideration given to the exposure levels of particular groups of affected persons to harm and to cumulative or aggregate impacts of the system on present and future potentially affected persons and groups of persons.

Deliberations on scope may include consideration of these questions:

1) For the potential and actual adverse impact identified, are there groups who possess characteristics which could make them vulnerable to higher levels of exposure[23][24] to the impact? If so, how much exposure could these groups face?
2) For the potential and actual adverse impact identified, consider the overall timescale of the AI system's impacts on the right or area of concern (in the case of democracy or the rule of law) under consideration. Are there cumulative or aggregate impacts of the system on affected persons and future affected persons that could expand the impacts of the system beyond the scope of impact already identified?

[23] **SEP Resources A** and **B** [to be elaborated and adopted by the CAI in 2025] provide detailed questions **(List of Questions to Assess Relative Stakeholder Salience)** and the description of formats of stakeholder engagement **(List of Factors Determining the Objectives and Levels of Stakeholder Engagement)** that can assist project teams in determining particularly relevant stakeholder groups and objectives for engagement.

[24] The term "level of exposure" here is understood as the proportion of a group that is adversely impacted by an AI system, where, in the case that a small fraction of the group is impacted, members have low levels of exposure and in the case that a very large fraction of the group is impacted, members have high levels of exposure. As an example, members of a group that is characterised by low socioeconomic status may have a high level of exposure to the potential adverse impacts of an AI model that is used to allocate public benefits.

Some “big picture” questions to reflect on when assessing cumulative or aggregate impacts may include:

- could the provision and use of the system contribute to wider adverse human rights, democracy or the rule of law impacts when its deployment is coordinated with (or occurs in tandem with) other systems that serve similar functions or purposes?

- could the provision and use of the system replicate, reinforce or augment socio-historically entrenched legacy harms or inherent characteristics in ways that could create knock-on effects for impacted persons and groups?

- could the provision and use of the system be understood to contribute to wider aggregate adverse impacts (e.g. on the environment and public health) when its deployment is considered in combination with other systems that may have similar impacts?

Reversibility

As explained previously, reversibility refers to the information about the degree of reparability or restoration that is possible for potentially affected persons as the result of efforts to overcome the adverse impact under consideration and to restore those affected to a situation similar or equivalent to their situation before the impact. Much as with considerations surrounding the scale of a potential impact, gaining an understanding of how reversable a harm is will depend on knowledge both about the specific context of the harm and about the affected persons who are subjected to it. Establishing the degree of reversibility for a potential adverse impact involves considerations about the effort needed to overcome and (potentially) reverse the harm.

Members of different groups may require different levels of effort to overcome adverse impacts, depending on their age, their positions in society and the circumstances of the harm (with vulnerable and marginalised groups often possessing less resilience than other dominant, privileged or majority groups).

Probability

Assessing the probability of a risk involves estimating the likelihood that a given adverse impact will occur, based as appropriate on qualitative judgment, quantitative analysis, and contextual understanding.

Determining a risk’s level of probability involves a broad analysis of contextual and operational conditions and generally determined by the level (kind, quantity and quality) of information that the risk is likely to materialise. This ensures that risk assessments are grounded in both data and expert insights, making it easier to prioritise and mitigate potential risks.

Outcome of the Risk and Impact Assessment

After questions and prompts on identifying and assessing potential and actual adverse impacts have been completed, impact prevention as well as mitigation prioritisation and planning can be launched. The process of impact mitigation planning and setting up access to remedies is covered in the next step.

IV. Mitigation Plan

Introduction

Once potential and actual adverse impacts have been identified and assessed, a Mitigation Plan should be drawn up and a reflection regarding the provision of access to remedies to potentially affected persons should take place, as appropriate.

This part of the HUDERIA process specifies the actions and processes aiming at addressing potential and actual adverse impacts through:

- formulating mitigation measures;
- drawing up a Mitigation Plan based upon the severity and probability of the identified harms;
- where appropriate, setting up access to remedy for potentially affected persons and other relevant parties.

Explanations

Scoping and prioritisation

Diligent risk and impact prevention and mitigation planning begins with a scoping and prioritisation stage. With input from engaged affected persons if and as appropriate, one should go through each identified potential and actual adverse impacts and map out the interrelations and interdependencies between them as well as surrounding social risk factors identified at the COBRA stage (such as, for instance, contextually specific vulnerabilities and precariousness).

Where prioritisation of prevention and mitigation actions is necessary (for instance, where delays in addressing a potential harm or the specific vulnerability of an affected individual or group could reduce its reversibility), decision-making should be steered by considerations of the relative probability and severity of the impacts under consideration.

Legal obligations

An important consideration in the elaboration of a Mitigation Plan is that legal obligations in regard to the respect for human rights, democracy and the rule of law, as set forth in applicable international and domestic law, should be taken into account at this stage of the HUDERIA process in considering whether and, if so how, potential adverse impacts can be mitigated and actual adverse impacts can be addressed.

The availability and effectiveness of legal remedies, including restoration or compensation as legal remedies, are determined by applicable international and domestic law.

Mitigation Hierarchy

When deciding upon the range of available actions that can be taken to prevent or mitigate potential adverse impacts, a structured approach called the “mitigation hierarchy” (avoid, mitigate, restore, compensate) may be used.

During the early stages of the AI system lifecycle, the impacts under consideration will not yet have occurred, so mitigation options of “avoid” and “mitigate” will be more relevant. In later iterations of the monitoring, review and re-evaluation (i.e. during the deployment stage) adverse impacts may have already occurred, making mitigation options of “restore” and “compensate” relevant alongside “avoid” and “mitigate”.

Descriptions of the options of the mitigation hierarchy are as follows:

AVOID	MITIGATE	RESTORE	COMPENSATE
Making changes to the design, development or deployment processes behind the production and use of the AI system, or to the AI system itself, at the outset, to avoid adverse impact. It is important to note that avoid does not equate to ignoring potential negative impacts.	Implementing actions in the design, development or deployment processes behind the production and use of the AI system, or making changes to the AI system itself, to minimise adverse impact.	Making changes to restore or rehabilitate affected persons to a situation similar to, or at least equivalent to, their situation before the adverse impact.	Compensation in kind or by other means, where feasible and other mitigation approaches are neither possible nor effective.
Level 1	Level 2	Level 3	Level 4
Most preferred	…		Least Preferred

The use of the term “mitigation hierarchy” suggests giving precedence to avoiding potential and actual adverse impacts altogether, in the first instance, and then to reducing and remediating them. It is also notable that, at later stages of the AI system lifecycle, where options of restoration and compensation become more relevant, more than one of these mitigation options may be relevant (as, for instance, where an affected individual needs to be rehabilitated simultaneously as immediate actions to minimise further harms are also taken).

In all situations, decisions about which prevention and/or mitigation action(s) to take should be guided by considerations prioritising the protection of human rights, democracy and the rule of law, and choices made to avoid and mitigate adverse impacts should be preferred to choices to compensate or remunerate potentially impacted persons for suffered harms.

In view of the entirety of the information obtained at this stage of the HUDERIA process, there is an opportunity to revisit the zero questions. This information may also be useful in informing the discussion of the question whether the lifecycle activities of the AI system at issue (under development or currently already in use in case of iterative review) align with human rights, democracy and the rule of law.

Access to remedies

Measures to address adverse impacts are not limited to legal remedies. Such impacts can be addressed using other mitigation measures such as those set out in policy, guidance, or other instruments.

When putting in place such measures the following points could be considered:

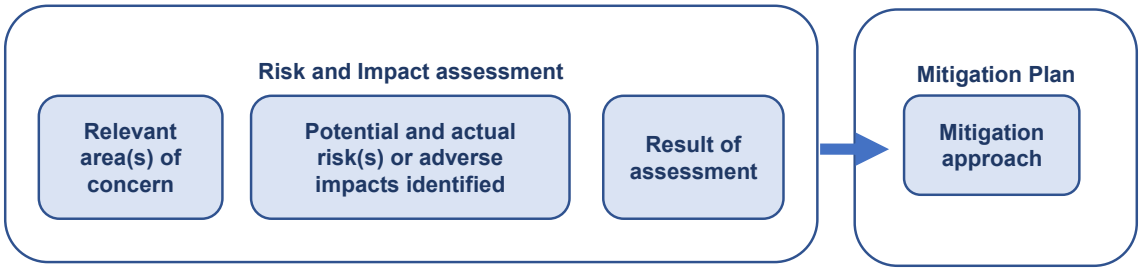
a) whether there are in place existing accountability measures and mechanisms in relation to human rights, democracy and the rule of law. It is essential that these existing frameworks are applied to the context of artificial intelligence systems;

b) the technical complexity, opacity, and data-driven nature of some AI systems can limit their transparency. This can create a significant imbalance in access to, understanding of, or control over information between the various parties involved in the AI system's lifecycle. Steps to document and provide information about the AI system and its impacts to affected persons can facilitate the provision and accessibility of effective remedies for adverse impacts on them;
c) the information provided in these measures should be context-appropriate, clear, and meaningful, ensuring that persons can effectively use it to exercise their rights in proceedings related to decisions impacting them;
d) if and as appropriate, the provision of further effective procedural guarantees and safeguards to the affected persons, in line with applicable international and domestic law, may be required.

Outcome of this element of the HUDERIA process

This element should produce a clear description of the measures and actions to address the risk and impacts identified, along with a clarification of the roles and responsibilities of the various actors involved in mitigation, management, and monitoring. Where appropriate, this element should also produce an accessible outline of the remedial mechanisms and measures available to impacted persons.

Additionally (see the Iterative Review section below), a plan is established for monitoring mitigation efforts, and for iteratively re-assessing and re-evaluating these efforts throughout subsequent phases of the AI system lifecycle.



Iterative Review

Introduction

Carrying out the HUDERIA at the beginning of the AI system's lifecycle is the first - albeit critical - step in a longer, iterative process of responsible monitoring and re-assessment. The process of Iterative Review ensures that risk and impact assessment remains effective throughout the whole AI system lifecycle. It is an ongoing process, offering regular opportunities to identify new impacts and update the Mitigation Plan.

Over time, the impacts of the AI system are likely to evolve, either due to decisions made during its development and implementation, contextual applications, or because of external changes in the real-world environment. These changes, which may include those regarding the data lifecycle, AI system development and design, procurement processes, changes in AI techniques, system integration or operationalisation, security vulnerabilities, as well as significant events or occurrences leading to harmful or unintended consequences, can influence the AI system's performance and/or its impact on affected persons and groups.

Such changes necessitate review to ensure that human rights, democracy, and the rule of law are continuously upheld throughout the AI system lifecycle. Particular attention should be paid to how these changes affect the system's performance and its impact on persons and communities.

Production, implementation and deployment factors

Choices made at any point during the lifecycle of the system as well as events occurring during the system's deployment may require review of prior decisions and assessments, particularly those made as a result of the HUDERIA process, creating the need for re-assessment, reconsideration, and amendment.

These changes, specifically those regarding the data lifecycle, AI system development and design, changes in AI techniques, system integration or operationalisation, security vulnerabilities, as well as significant events or occurrences leading to harmful or unintended consequences, can influence the AI system's performance and/or its impact on affected persons and groups. The processes of an AI system lifecycle are iterative and often non-linear, frequently requiring revision and updates, as appropriate.

Real-world Environment Factors

Changes in social, regulatory, policy or legal environments in which the system is in production or use may have effects on how well the AI system works and on how it impacts the rights of affected persons or groups.

Likewise, regulatory and policy changes or changes in data recording methods may take place in the population of concern in ways that affect whether the data used to train the model accurately portrays phenomena, populations or related factors in an accurate manner.

In the same vein, cultural or behavioural shifts may occur within affected populations which alter the underlying data distribution and hamper the performance of a model, which has been trained on data collected prior to such shifts. All of these alterations of contextual conditions

can have a significant effect on how an AI system performs and on the way it impacts affected persons, groups, communities and society in general.

Implementing the iterative review

While the HUDERIA Methodology provides flexibility regarding the exact modalities, thresholds, triggers, monitoring and governance mechanisms for the Iterative Review process, the following principles could be considered:

a) continued review of the HUDERIA plays a pivotal role in its continued efficacy and reliability;

b) a plan is established for monitoring impacts and for re-assessing and re-evaluating the HUDERIA during each phase of the project lifecycle up to system retirement or decommissioning;

c) processes used for iterative review should remain as responsive as possible to the way the AI system interacts with its operating environments and with impacted persons (e.g. possible application areas of the AI system, the emergence of new forms of system misuse etc.);

d) in rapidly evolving or changing contexts, there may be a need for more frequent re-assessment and re-evaluation interventions.