COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 21 May 2021

CAHAI-PDG(2021)05
Provisional

AD HOC COMMITTEE ON ARTIFICIAL INTELLIGENCE (CAHAI)

POLICY DEVELOPMENT GROUP (CAHAI-PDG)

Human Rights, Democracy and Rule of Law Impact Assessment of AI systems[1]

www.coe.int/cahai

[1] This draft document is going to be reviewed by the CAHAI-PDG and should by no means be considered as final.

Human Rights, Democracy and Rule of Law Impact Assessment of AI systems

Introduction and Scope

Following the adoption of the CAHAI feasibility study in December 2020, which included a specific chapter (9) on practical and follow-up mechanisms needed to ensure compliance with a legal framework on Artificial Intelligence (AI) based on Council of Europe's standards on human rights, rule of law and democracy, the CAHAI, through its Policy Development Group (CAHAI-PDG) has decided to examine more closely one of such mechanisms, namely human rights impact assessments, and to:

1. define a methodology to carry out impact assessments of Artificial Intelligence (AI) applications from the perspective of human rights, democracy, and the rule of law, based on relevant Council of Europe (CoE) standards and the work already undertaken in this field at the international[2] and national level.

2. develop an impact assessment model.

3. examine the complementarity of such an assessment with other compliance mechanisms outlined in chapter 9 of the feasibility study.

This document follows a similar structure.

In section I, the methodological considerations relevant to a Model for a Human Rights (HR), Democracy (D), and Rule of Law (R) Impact Assessment of AI applications (a HRDRIA) are outlined. In this regard, the document presents existing impact assessment tools and frameworks which either relate in particular to AI or apply to the impact assessment of human rights in general. Furthermore, it explores the most relevant features of these tools and frameworks, in particular main traits and assessment variables of AI systems and how to integrate rule of law and democracy in an integrated assessment model for AI systems.

In section II, the document proposes a possible methodology for a HRDRIA, with a view to elaborating a concrete model of assessment at a later stage.

In section III, the complementarity of a future HRDRIA along the lines of the existing impact assessments mentioned in section II, with the compliance mechanisms outlined in chapter 9 of the feasibility study is examined, with a view to strengthening added value and complementarity.

[2] See for instance CM/Rec (2020)1 on the Human Rights Impact of Algorithmic Systems; the ongoing OECD work on classification of AI systems; and the Human Rights Impact Assessment Toolbox developed by the Danish Institute for Human Rights).

Section I. Methodological considerations for a Human Rights, Democracy and Rule of Law assessment model

The methodological considerations for any possible model of impact assessment of AI systems based on Human Rights, Democracy, and the Rule of Law (coined above as an HRDRIA for AI) should build upon the already established practices and experience with Human Rights Impact Assessments (HRIAs) **(A)**. It should also acknowledge the origin of HRIA which can be situated in human rights due diligence as included in the United Nations Guiding Principles on Business Human Rights and the OECD Guidelines for Multinational Enterprises[3]. Stressing this origin is important as human rights due diligence is an ongoing process and not a snapshot of a moment in time. Furthermore, such due diligence is not limited to the operations of a single company but covers the entire value chain. The same should count for any HRDRIA for AI.

However, the peculiarities of AI systems present challenges to a simple import of the human rights due diligence acquis to the AI domain. Additionally, there is the issue of including Democracy and the Rule of Law as dimensions to any comprehensive AI system assessment – these dimensions typically being absent in existing impact assessments. Hence, this section also aims to review and analyse the elements of the HRIA framework which can be transposed to the AI context and how to extend the HRIA framework to include Democracy and the Rule of Law (towards a HRDRIA) **(B)**.

In terms of scope, it is important to note that general HRIA frameworks tend to focus on adverse impacts of the operations of a company on human rights. This is the case also for most of the current impact assessment models of AI systems being used in the private or in the public sector. Accordingly, a HRDRIA should be developed in line with this approach, focusing in particular on adverse impacts on Human Rights, Democracy, and the Rule of Law.

Obviously, this does not imply that the use of AI generates adverse impacts only. AI has many advantages and can create a huge beneficial impact for mankind. It may even assist in the enjoyment, protection and strengthening of human rights, and this positive contribution should not be neglected. However, the specific function of HRIA is to detect possible risks of infringement for human rights arising from a given AI system, and not to balance them against possible beneficial impacts arising from such an application. Balancing benefits against risks is not part of the assessment methodology but would rather be performed later as part of a judgement of opportunity as to whether deploy such application. For instance, in certain cases public authorities could conclude that the beneficial impacts offset adverse impact and hence decide using such application for a given purpose. If in this case one or more human rights are curbed (which the HRDRIA can help assess) it is essential that this occurs in a manner that is justified through an approach that is both proportionate and necessary in a democratic society, for instance in the interest of national security or another legitimate public interest.

Another relevant issue is whether HRDRIA should apply to private actors, public actors or both. Even though there are no doubt differences between private and public actors, in terms of roles

[3] Which can be accessed at https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf and http://www.oecd.org/daf/inv/mne/48004323.pdf.

and scope of obligations when it comes to ensuring respect for human rights, performing a HRDRIA is equally important for private and public actors when developing, deploying or procuring AI systems, as already pointed out in several Council of Europe documents[4]. Indeed, both type of actors can deploy AI in a manner that negatively impacts human rights, democracy and the rule of law. Moreover, it can be noted that the distinction between the public and the private sphere is often blurred when it comes to AI systems: private actors developing AI solutions play an increased role in the public sphere[5], and private and public actors tend to increasingly work together, as AI applications used in the public sector are very often developed by private actors. The aim of the HRDRIA is to identify risks caused by an AI application and hence help respect human rights, which is a relevant duty for both private and public actors.

We will first analyse the sources and content of the traditional human rights impact assessments and provide a list of existing general and AI specific impact assessments with their main features relevant to this context, to be taken into consideration in the effort to develop a comprehensive model for HRDRIA.

A. Human Rights Impact Assessment, sources, materials and experiences

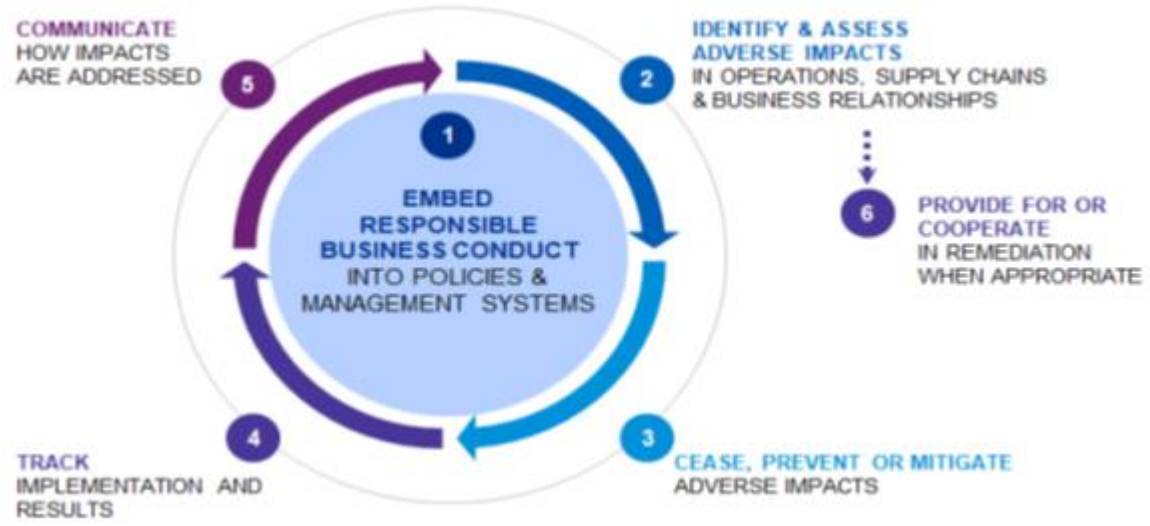
General, not AI specific, Human Rights Impact Assessments (HRIA) draw from the international frameworks referred to earlier which require human rights due diligence[6]. Human rights due diligence is an ongoing and iterative process that includes the following steps[7]:

[4] Public actors and those private actors working with them are also expected to undertake a fundamental rights impact assessment. See Recommendation from the Commissioner for Human Rights, *Unboxing Artificial Intelligence*, p. 7 and 8, which can be accessed at https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64; Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, under 5.2, which can be accessed at https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154.

[5] For example, some AI tools regulate entire global communities (social networks) with their own rules and dispute resolution and enforcement mechanisms. The regulators in these communities may often resemble public actors.

[6] As defined in Principle 17 of the United Nations Guiding Principles on Business Human Rights.

[7] OECD Due Diligence Guidance for Responsible Business Conduct, p. 21, which can be accessed at http://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf



That said, HRIA is not the same as human rights due diligence, but part of it and especially connected to steps 2 (identifying risk) and 6 (provide for or cooperate in remedy, if the impact assessment has revealed human rights abuse which an actor caused or contributes to).

It is also important to note that HRIA does not include a normative element: its purpose is to identify and assess adverse impact, without entering further into the implications of such an assessment – whether, for instance, an AI application should not be further developed or used, that its type of use should be limited, or whether the AI model should be changed or the data quality improved. Thus, while a HRIA does not indicate per se whether a public supervisor or certification body should act or take measures, it might be a relevant source of information for them, as well as for developers or users in the value chain of the AI application who want to further develop or use it. It is also worth noting that conducting a human rights impact assessment is mandatory in some countries.

General HRIA is elaborated in specialized toolkits such as for instance, the one developed by the Danish Institute for Human Rights. The framework of HRIA practices and standards provides relevant baselines for developing a model for HRDRIA, which is important to consider.

Recently, a specialized methodology for performing a Data Protection Impact Assessment (DPIA) has emerged. Although the proposal aims to develop a practical model to assess Automated Decision-Making (ADM) impact on fundamental rights, it is mostly centred on the EU General Data Protection Regulation (GDPR). Since it is a novel approach to ADM assessment, this model is relevant for a possible generalization of the proposal for a broader model to HRDRIA, beyond the GDPR. Several other AI specific impact assessments have been developed too, such as the Trustworthy AI Assessment List developed by the High-Level Expert Group on AI of the EU as part of their broader Ethics Guidelines for Trustworthy AI.

Firstly, we will be looking to the existing human rights impact assessments both general (i) and AI related (ii).

1. Examining existing Human Rights Impact Assessments

(i) General documents and frameworks referring to human rights impact assessments in general

The Guiding Principles on Business and Human Rights

The Guiding Principles on Business and Human Rights (UNGPs[8]) were adopted by the United Nations in 2011 and underline: a) the obligation of States to respect, protect and comply with human rights, b) the important role that companies fulfil as regards ensuring compliance with laws and human rights, and c) the need to implement legal protection and judicial remedies in cases of negative impact on human rights.

Identifying and Assessing Human Rights Risks related to End-Use

This document[9] elaborated by the Office of the High Commissioner for Human Rights is addressed to leaders within technology companies who seek to understand the basic expectations of the UNGPs when it comes to identifying and assessing human rights risks related to products and services. In this regard, the UNGPs expect companies to: 1) provide a broad overview of possible impacts; 2) focus on the most serious harms; and 3) engage and communicate meaningfully with stakeholders.

OECD Guidelines for Multinational Enterprises

The OECD document[10] (2011) aims to address recommendations for multinational enterprises operating in or from adhering countries. It provides principles and standards for responsible business conduct in a global context regulated by international law. The document requires that governments promote values around the respect of legal norms by private agents. In specific, it requires that companies comply with legal norms on: 1) human rights,
2) employment and industrial regulation, 3) environment, 4) antitrust, 5) consumer interests, and 6) taxation.

(ii) Specific Documents and Frameworks related to AI

Unboxing Artificial Intelligence: 10 steps to protect Human Rights

The Recommendation on "Unboxing Artificial Intelligence: 10 steps to protect Human Rights[11]" was issued by the Council of Europe's Commissioner on Human Rights in 2019[12]. It establishes recommendations to prevent and mitigate the negative impacts of Artificial Intelligence on human rights. It focuses on 10 areas of action: 1) Human Rights impact assessment, 2) Public consultations, 3) Obligation of States of facilitate the implementation of human rights standards in the private sector,
4) Information and transparency, 5) Independent oversight, 6) Non-discrimination and equality, 7) Data protection and privacy, 8) Freedoms of expression, assembly and association, and the right of work, 9) remedies, 10) Promotion of Artificial Intelligence literacy.

[8] https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.
[9] https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf
[10] http://www.oecd.org/daf/inv/mne/48004323.pdf.
[11] Suggested by Access Now
[12] https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64

Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems

This Recommendation[13] aims to guide States and private actors in their actions related to the design and development of algorithm systems, and to ensure that human rights and individual freedoms of the European Convention of Human Rights are protected against technological development. Among other recommendations, it states that: 1) States should review their legal framework to adapt it to the technological context, and 2) private actors should comply with the laws and respect human rights in accordance with the provisions of the UNGPs.

Ethics Guidelines for Trustworthy AI

The Ethics Guidelines for Trustworthy Artificial Intelligence were drafted by the European Commission's High-Level Expert Group on Artificial Intelligence (2019)[14].This document aims to set the parameters of trustworthy Artificial Intelligence. According to these parameters, trustworthy Artificial Intelligence should be: 1) lawful, 2) ethical, and 3) robust. In addition, it establishes some requirements that Artificial Intelligence systems should meet to achieve trustworthiness, drawn from human rights, among them: the protection of personal data, the guarantee of transparency and accountability. To assess whether AI systems fulfil these requirements, an Assessment List for Trustworthy AI was included in the Guidelines, meant to help AI developers and deployers to evaluate and improve alignment of their systems with the requirements.

Examining the Black Box

This report drafted by the Ada Lovelace Institute[15]clarifies terms of algorithm audits and algorithmic impact assessments and describes the current state of research and practice. Regarding the algorithm audits, it identifies two key approaches: 1) Bias audit: a targeted approach, focused on assessing algorithmic systems for bias, 2) Regulatory inspection: a broad approach, focused on an algorithmic system's compliance with regulation or norms, necessitating a number of different tools and methods; typically performed by regulators or auditing professionals. Regarding algorithmic impact assessments, it also identifies two approaches: 1) Algorithmic risk assessment: assessing possible societal impacts of an algorithmic system before the system is in use, 2) Algorithmic impact evaluation: assessing possible societal impacts of an algorithmic system on the users or population it affects after it is in use.

Algorithmic Equity Toolkit

The American Civil Liberties Union has produced an Algorithmic Equity Toolkit also known as AEKit[16], this toolkit is a collection of four components designed to identify surveillance and decision-making technologies used by governments; make sense of how those technologies work; and pose questions about their impacts, effectiveness, and oversight. Those components are: 1) Flowchart; 2) System map; 3) Fill-in-the-blank; and 4) Questionnaire. They are intended to be used in the aforementioned order but can be applied as the user sees fit.

[13] https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154
[14] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[15]https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf
[16] https://www.aclu-wa.org/AEKit

Ethics and Algorithms Toolkit

GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC have developed the Ethics & Algorithms Toolkit[17]. This toolkit focuses on anyone who is building or acquiring algorithms in the government sector. It walks the user through a series of questions to help them 1) understand the ethical risks posed by their use of the algorithm, and 2) identify what they can do to minimize those ethical risks. It should be used whenever an algorithm is being used to inform a decision in the public sector.

Canadian Algorithmic Impact Assessment Tool

This is a mandatory risk assessment tool intended to support the Canadian Treasury Board's *Directive on Automated Decision-Making*[18]. Composed of 48 risks and 33 mitigation questions, it is a questionnaire that determines the impact level of an automated decision-making system. The assessment is organized according to the government's policy, ethical, and administrative law considerations of automated decision system risk areas as established through the Treasury Board of Canada Secretariat's consultations with academia, civil society, and other public institutions. It is designed to help departments and agencies better understand and manage the risks associated with automated decision systems. It should be completed at the beginning of the design phase of a project, and the results should be released in an accessible format in both of Canada's official languages.

Algorithmic Impact Assessments: A practical framework for public agency accountability

AI Now has developed a model for impact assessments, entitled Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability[19]. This article recommends that: 1) public agencies should conduct a self-assessment of existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias, or other concerns across affected communities; 2) agencies should develop meaningful external researcher review processes to discover, measure, or track impacts over time; 3) agencies should provide notice to the public disclosing their definition of "automated decision system," existing and proposed systems, and any related self-assessments and researcher review processes before the system has been acquired; 4) agencies should solicit public comments to clarify concerns and answer outstanding questions; and 5) governments should provide enhanced due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased, or otherwise harmful system uses that agencies have failed to mitigate or correct.

In addition to the guidance which has been provided by the above-mentioned institutions, it is worth recalling that different scientific articles have also addressed the issue of human rights impact assessment of AI systems[20].

[17] https://ethicstoolkit.ai/

[18] https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[19] https://ainowinstitute.org/aiareport2018.pdf

[20] See for instance, Alessandro Mantelero, "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment", Computer Law and Security Review, Volume 34, Issue 4, August 2018, Pages 754-772, as well as Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In ACM Conference on Fairness, Accountability,and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.ACM,

B. Searching for the relevant features for a model of HRDRIA

Today's AI systems present specific characteristics that should be considered when developing a model for HRDRIA, especially when taking the general HRIA frameworks as a starting point. These will be discussed in the first part of this section **(1)**. In addition, the use of AI can also impact the values of Democracy and the Rule of Law, making them equally important dimensions of an assessment model for AI systems. However, the inclusion of these two dimensions in an impact assessment model presents conceptual challenges that we must analyse in detail in a separate section **(2)**.

1. AI systems, Main Traits as Assessment Variables

As explained in the CAHAI Feasibility Study, AI systems can be beneficial to individuals and societies, but can also risk undermining individual Human Rights, Democracy and the Rule of Law (HRDR). In almost all technical decisions there is always a trade-off between the value that the technical system can provide and the damage that a concrete system may produce. A more complete, full-fledged approach could balance and consider both the value that a given AI system brings about and the likelihood and extent of harming HRDR[21].

When assessing and grading the likelihood and extent of risks associated with an AI System, the following elements could, as a minimum, be considered:

i) the context in which the AI system is used;

ii) its underlying technology, covering dimensions such as scope, reliability, traceability, explainability, data used, level of automation, security and accessibility of the AI system;

iii) the actors involved and the stage of development of the AI system;

iv) the stakeholders to be involved in the assessment.

Within the grading scale of value versus potential harm, any HRDRIA should be guided by the no-harm pre-emptive principle: if the assessment identifies a high HRDR risk that cannot be mitigated immediately, the AI system should neither be developed, implemented or used in that form[22] by any private or public authorities at least until effective measures are adopted to prevent potential or further HRDR risks (bans/moratoria of high-risk AI[23]). Pre-emptive measures should be implemented regarding those AI applications that pose the biggest risk in terms of scale, severity and irremediableness and the minimum benefit-value added[24].

In addition to the no-harm pre-emptive principle, any HRDRIA should also apply the proportionality principle: in particular, it should be assessed whether a particular AI system merits

[21] Suggested by ES
[22] Cfr. Ethics Guidelines for Trustworthy AI, p20: “In situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form.” Reference suggested by NS.
[23] Taken from CINGO Guidelines for impact assessment.
[24] In response to FRA clarification request on “first measures.”

a full and comprehensive HRDRIA or not. This is important to ensure the proportionality of any legal framework[25].

(i) The Context of Application as a Variable of Assessing Impact

An AI system application could pose a high risk to Human Rights, rule of law and democracy, but the same AI system configured for a different application could yield a lower (or no) risk[26]. Therefore, consideration needs to be provided to the geopolitical, social or economic context in which the AI system will operate. Likewise, in certain contexts the AI system creates value and benefits that could be the opposite in a different context.

When looking at the context of an application, it is important to consider the system's declared purpose by the designer, the developer or as per request of the client[27]. Consider the following example: an AI system which is intended to trace financial transaction patterns to signal out potential money laundering operations. The designer or the developer of the system developed and created the system to perform a specific task, which is to detect money laundering operations. However, the system's operator can change the system's purpose and thus change the context that is relevant for risk-assessment. That will be the case if the system operator in a treasury department changes the data flow from suspicious money transactions to regular transactions that will yield a business-financial strategy of corporations and individuals when allocating and moving their money assets. The new purpose of the AI system is now to reveal business transactions and cash flow strategies.

It is equally important to assess who are the users[28] of AI systems. Consider once more a money laundering prevention AI system. The same AI system intended for Treasury Officials and Financial Intelligence Units presents a different context of risks if the system is destined for training data scientists in a public policy lab.

(ii) The underlying technologies of AI systems as a variable of assessing impact

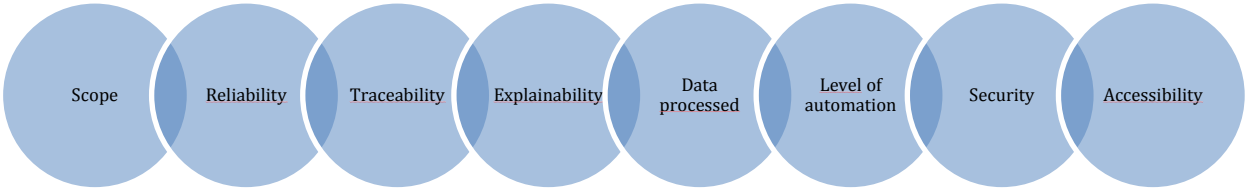
While the context of application helps estimate the level of risk of an AI system, it is essential for any risk-assessment to consider the type of underlying technology of the particular AI system. The field of AI has different approaches - from thought processes and reasoning to behaviour, and in both cases with different measures of success, be it the fidelity to human performance, or an ideal or rational performance[29]- or techniques related to those approaches[30]. The list of AI

[25] Suggested by EE
[26] See also OECD, OECD Framework for the Classification of AI Systems – Preliminary Findings, p. 9-15, not yet published, as well as the Trustworthy AI Assessment List developed by the High Level Expert Group on AI of the EU as part of their broader Ethics Guidelines for Trustworthy AI, in particular as regards considering the outcome generated by an AI application: in this respect, it should be clarified whether the system generates an advice, decision or other type of outcome (page 7)
[27] Suggested by EE.
[28] In fact, following the CAHAI Feasibility Study, every relevant subject in the compliance mechanism is a source of context. Cfr. Sec. 9.2 of the CAHAI Feasibility Study.
[29] Russell & Norvig, 2015. (Reference suggested by ES)
[30] Chowdhary, 2020.

techniques has evolved and progresses continuously[31]. It is not uncommon to see that some techniques can exist in combination to produce more complex AI systems (i.e., a generative adversarial neural network has two neural networks, one with a discriminative model and the second one with a probabilistic model, when used to produce a system to detect counterfeit currency). To maintain the HRDRIA framework relevant in view of the ever-evolving AI techniques the assessing methodology should remain as 'algorithm-neutral' as possible[32] but not without a deep awareness of the differences and implications of each AI technique.

The type of technology behind an AI system is relevant because it gives us eight, preliminary, dimensions that could signal out potential risks.

The eight dimensions behind an AI technology pointing to potential risks



The first dimension is scope. For example, an AI system that uses a supervised learning algorithm trained with actual data has a margin of error, of over or under classifying its target prediction[33]. The technology's capability is relative to the training set, the specific algorithm used to classify the training examples, the control data set employed to adjust the learning cycles –be it that the updating is based on newly collected data or with data generated through the algorithm in real deployment[34]– and the real-world scenario of the system it's applied to: some processes are more chaotic and less predictable than others[35]. In our example, to know the inner workings of the training technique and the type of algorithm is relevant to interpret the outcome of the AI system, but also to scrutinise the data procurement, pruning, and potential biases. There are two worrisome cases regarding the dimension of scope. The first case is the use of an AI system beyond the scope of its underlying technology[36]. For example, an AI system that presents the statistical chances of aggregated forensic evidence-tests produced during a trial as the basis for the chances that a particular individual is guilty (the so-called prosecutor's fallacy[37]). The second case is a system with a low error rate that makes a fundamental error[38]. For example, an AI

[31] See The Feasibility Study Ch. 8.
[32] Suggested by UK and Access Now.
[33] Classification algorithms for modelling classification predictive modelling problems. Classification predictive modelling algorithms are evaluated based on their results. Classification accuracy is a popular metric used to evaluate the performance of a model based on the predicted class labels. Or , alternatively, instead of class labels, some tasks may require the prediction of a probability of class membership for each example. (Clarification requested by ES)
[34] Suggested by FRA
[35] Suggested by UK
[36] Flynn et al, 2020: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156005/
[37] Sesardic, 2008: https://philpapers.org/archive/DEMGBS.pdf
[38] Suggested by UK

system that performs raw DNA sequencing based with a low error rate[39] but that could make a fundamental error concerning an experimental drug target producing an undesirable effect.

The second dimension is reliability[40]. The level of consistent output that is expected behaviour of the AI system can be determined following the technology in which it is based[41]. The use of adaptive algorithms for an AI system organizing the public utilities in a smart city could evolve in time because that is what the algorithm is expected to do. A determinative algorithm based on rules is reliable if the infrastructure that operationalizes the AI systems functions accordingly and does not experience data corruption. And while technologies do not have universal dependability –because that can be different from one group of people to the next– a measure of reliability, a measure of confidence, is useful to assess risk. "Apart from high-accuracy [Deep Neural Network] algorithms, there is a significant need for robust machine learning systems and hardware architectures that can generate reliable and trustworthy results in the presence of hardware-level faults while also preserving security and privacy[42]". We can tolerate slight inaccuracy in an AI language translator system (NLP) but not so much in an autonomous driving vehicle[43].

The third dimension is traceability. The output of an AI system can be traceable in terms of the architecture of the system. To be more precise, traceability requires establishing not only how a system worked but how it was created and for what purpose, in a way that explains why a system has particular dynamics or behaviours[44]. Traceability is a predicate that accepts granularity. If an AI system is built on a technology that can be fine-tuned to individual output (i.e. why and to what steps of processing did the AI system produce a specific output-decision) then the level of traceability enables the explainability of the output. However, the use of multiple neural networks combined with large amounts of training data, produce AI systems that become highly costly to trace step-by-step. There is then a lower level of traceability and thus of explainability[45]. Furthermore, AI is often developed building on earlier versions of other software. This may increase complexity in terms of the cause of specific impact[46].

The fourth dimension is explainability[47]. That an AI system can be, at some degree, traceable in terms of its process-outputs is an enabling condition to explain the system's behaviour. Explainability, however, entails a communicative aspect. An AI system can be explainable to an AI knowledgeable audience but not to the general public and civil society at large[48].

---

[39] Such as the program PHRED: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC310698/

[40] A close concept of reliability is dependability. The former is the probability that the AI system will correctly (expected behavior) deliver services as expected by designers/developer/operator/user. The latter is a measure of the designer/developer/operator/user trust into the system. One way to achieve trust is when a system is reliable. See O'Regan, 2017: https://link.springer.com/chapter/10.1007/978-3-319-57750-0_11

[41] Hong et al, 2021: https://arxiv.org/abs/2102.01740

[42] Hanif et al, 2018: https://ieeexplore.ieee.org/document/8474192

[43] There are proposals, for example to introduce confidence measures for critical systems, such as military AI, see Jah et al, 2019: https://papers.nips.cc/paper/9355-attribution-based-confidence-metric-for-deep-neural-networks

[44] Kroll, 2021: https://arxiv.org/abs/2101.09385

[45] Felzmann et al, 2020: https://link.springer.com/article/10.1007/s11948-020-00276-4

[46] Generally, newer versions shouldn't affect this problem, it's when one system loops into another and then back that you have problems. Generally, because there are logistical (not mathematical) reasons why differing versions could make life difficult (Observation by UK)

[47] Proposed by CINGO.

[48] Umang Bhatt et al's research shows that most explainable AI systems are intended for "debugging", thus targeting AI developers as opposed to AI users/civil society: https://arxiv.org/abs/1909.06342) (Reference by CINGO)

The fifth dimension is which type of data is extracted and processed by AI systems. Obviously, privacy sensitive data may pose larger challenges than for example data on greenhouse gas emissions. The same goes for the use of content or just formatted data or data with a level of personal information[49]. Very large data sets[50] have different statistical properties (and thus potential impacts) to smaller ones[51]. This dimension entails an analysis on the importance of having diverse data sets as well as acknowledging the difficulty of creating adequately diverse and non-discriminatory data sets, because biases and discrimination are embedded in society[52].

The sixth dimension is the level of automation[53] of AI systems. The automation of a system should be examined in close relation to the technology dimension. For example, a completely autonomous AI system used for machine calibration can be highly predictable and reliable and may exceed operator-level precision[54]. If that system is not fully automated, but operator-human dependent, then there is a greater chance for human error. However, this is not always true. Even if a human remains in the loop, they may trust the outcomes the AI system generates and decide accordingly without making their own assessment[55]. There are documented cases where human operators did not act to prevent harm or a catastrophic consequence because of the presence of an automated system[56]. Humans in the loop have been made accountable even if they were not able to see the relevant information from the interface of the system they operated or because they could not advert the failure of the automated system. In those cases, humans function as a moral crumple zone for accountability purposes[57].

The seventh dimension concerns the security of AI systems. The higher the risk of hacks, adversarial attacks or other security incidents (such as negative side effects of reinforcement learning[58]), the higher the risks of non-desired impact may become.

The eight dimension is the technology's accessibility[59]. The type of technology in a social context

[49] Observation by FRA
[50] Although the expression "BigData" is used informally to refer to vast amounts of data; or flows of data at high speed, the expression is still vague even among practitioners and data-scientist See Favaretto M et al 2020: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0228987. Of course, there has been efforts to conceptualize the term in a more precise way: M. Al-Mekhlal and A. Ali Khwaja (2019): https://ieeexplore.ieee.org/abstract/document/8919591
[51] Suggested by UK
[52] Suggested by CINGO
[53] Suggested by FRA
[54] Observation by ES
[55] Cfr. Elish, 2019: https://estsjournal.org/index.php/ests/article/view/260
[56] Cfr. Elish, 2019: https://estsjournal.org/index.php/ests/article/view/260
[57] Very relevant in this respect is the work of the High-Level Expert Group on Artificial Intelligence, namely the Assessment List for Trustworthy Artificial Intelligence (ALTAI) which includes a section at page 8 on Human Oversight. Underlying the importance of assessing whether the system operates autonomously or whether it allows a human in the loop (capability for human intervention in every decision cycle of the system), human on the loop (capability for human intervention during the design cycle of the system and monitoring the system's operation) or human in command (capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the AI system in any particular situation). In connection with this it is relevant to observe whether a detection and intervention mechanism exists in case undesired outcomes or functioning of the AI system emerge as well as whether the system can be stopped and, in case of autonomous learning, whether oversight over this process exists.
[58] Suggested by CINGO
[59] Proposed by CINGO

of broad digital gaps could only increase the technological exclusion of the disenfranchised groups. Marginalized groups are often locked-out of tech because of lack of accessibility/digital inclusion.

Convergences with the OECD recent work on a Framework for the Classification of AI Systems

It is interesting to note that many of the elements identified above converge with those being considered by the OECD in its ongoing Framework for the Classification of AI Systems.

The OECD has developed several guiding questions for transparency, explainability and robustness[60]. For example, if explainability and transparency are not implemented, it is not possible to engage with stakeholders and discuss the human rights impact of an AI application. It is also important to assess whether capabilities and limitations of an AI system have been communicated to (end)users[61]. Furthermore, the way in which data quality is assured is relevant too. For example, the way in which data are collected, by automated systems or by humans, the scale of collection and the dynamic nature of data may influence their quality for use by AI systems[62]. Next to this the format and structure of data are important as well as the rights and identifiability of individuals to which data refer[63]. Also the type of model used to develop AI and the way in which it is built are relevant[64]. It is important to note often composite systems are used to develop an AI application which may hamper assessing how AI reaches certain results, especially if the model deploys unsupervised learning. Beyond this it is important how a model is built[65]. Finally, the tasks and objectives of an AI system are relevant. For example, it is used for event detecting, forecasting or goal-optimisation[66]. The OECD has also pointed out several tools exist for this technical analysis[67].

(*iii) Actors involved and stage of development of the AI system*

In addition to the context and technology dimensions, there are also two parameters to consider when building a model of HRDRIA: (a) the actors involved and their role in relation to the AI system and (b) the stage of development of the system within its life-cycle.

60 OECD, OECD Framework for the Classification of AI Systems – Preliminary Findings, p. 24, not yet published.
61 p. 15.
62 OECD, OECD Framework for the Classification of AI Systems – Preliminary Findings, p. 16-18.
63 OECD, OECD Framework for the Classification of AI Systems – Preliminary Findings, p. 18-21.
64 See in this regard Section I, 1, B *supra*. Also: OECD, OECD Framework for the Classification of AI Systems – Preliminary Findings, p. 23-28.
65 OECD, OECD Framework for the Classification of AI Systems – Preliminary Findings, p. 26 and 27.
66 OECD, OECD Framework for the Classification of AI Systems – Preliminary Findings, p. 29 and 30.
67 OECD, Tools for Trustworthy AI, p. 8, not yet published.

(a) The role in relation to the AI system

An HRDRIA should be an ongoing assessment tool to be used throughout[68] the life-cycle of the AI system. In the life-cycle of an AI system, various agents are playing different roles in relation to the AI system: designers, developers, distributors, operators, users. Additionally, some agents can play a role of control or display a control function on the AI system's performance (i.e., human in charge, human on the loop, human in the loop). But humans are not the most frequent agents in control of an AI system. Other AI systems can be in control of another AI system; or an AI system can be in relation to a non-automated system[69].

(b) The stage of the system within its life cycle

The second parameter is the stage of the AI system in its life cycle. In turn, the life cycle of the AI system is relative to the type of AI technology. There are many different engineering processes to build AI systems[70]. Just to provide an example: it is an overwhelmingly crude simplification to state that to use an AI model, a predictive model of machine learning, you only need to feed data to a predictor. When implementing AI, there is a complex pipeline: configuration, automation, data collection, data verification, feature engineering, testing and debugging, resource management, model analysis, process management, metadata management, serving infrastructure, and monitoring[71]. The risk-assessment of an AI system at the "AI implementation" stage will be quite different from the previous stages of the AI (Machine Learning) life cycle: scoping-understanding[72].

(iv) Stakeholders

Community engagement is crucial for successful HRDRIA[73]. There should be effective mechanisms to identify the stakeholders, within the relevant communities and to produce active participation regarding the assessment process. Proper identification of relevant stakeholders and the processes to bring them on board in an open, sufficiently informed and layman-oriented manner is a crucial factor that will determine the potential impact on HRDR[74].

2. Democracy and the Rule of Law as dimensions of an integrated assessment model for AI systems[75]

Any HRIA on AI systems assumes a common reference to human rights (HR). In that regard, the inclusion of the assessment of AI's impact on Democracy and the Rule of Law present stark

[68] Suggested by EE
[69] Observation by UK
[70] Observation by ES
[71]{$NOTE_LABEL} MLOps: Continuous delivery and automation pipelines in machine learning
[72] At least lifecycle management by Microsoft, Google and DataRobot all acknowledge the scope-understanding stage. https://cloud.google.com/blog/products/ai-machine-learning/making-the-machine-the-machine-learning-lifecycle; https://azure.microsoft.com/en-ca/blog/how-to-accelerate-devops-with-machine-learning-lifecycle-management/; https://www.datarobot.com/wiki/machine-learning-life-cycle/
[73] Proposed by HomoDigitalis
[74] Emphasis on these aspect suggested by NS.
[75] The entire section has benefited greatly from clarifications, nuances and recommendations from NS.

challenges from a methodological point of view that should be considered in some detail **(i)** before proceeding with a possible solution along the lines of rights-grounded benchmarks **(ii)**.

(i) Methodological challenges for assessing the impact of AI applications on Democracy and the Rule of Law

Human rights typically express more concrete values and common goods that should be considered, such as the freedom of expression (Article 10 ECHR) or the right to an effective remedy (Article 13 ECHR). In contrast, 'democracy' is not reducible (although it is connected) to a set of rights or to a set of rules.

The mere promulgation of rules does not entail a democracy. At the same time, the common set of rules that comprises democracy is a necessary condition for its emergence. Several norms exist to safeguard the democratic process, such as rules for accessing and exercising power and public decisions, but democracy has also a cultural, institutional and a social dimension. The report of the Secretary General of the Council of Europe on the state of democracy, human rights and the rule of law (2021) notes in this respect that "Democracy is more than a matter of laws and institutions, which are necessary but not sufficient: functioning democracies depend on what is often called a culture of democracy"[76]. The report also underlines the close interrelation which exists among democracy, rule of law and human rights, already acknowledged in several Council of Europe's key texts[77].

Far from being limited to free and fair elections, democracy includes open, diverse and accessible participation, integrity, an active civil society, and a fair distribution of powers amongst the organs of the state and citizens, living in equality and dignity[78]. Good governance at all levels (local, regional and national) is also a key component of democracy[79], and the organic link between the quality of democracy and the quality of governance cannot be overstated: a degradation of democracy will lead to lesser accountability and a degradation of governance; a degradation of governance will in its turn lead to dissatisfied citizens and hence to a debasement of democracy.

The Rule of Law[80] is also an essential component of democracy, as also highlighted by the Venice Commission in its Rule of Law Checklist.[81]

The Commission recalls that, in the Preamble of the Council of Europe's Statute, the rule of law is mentioned as one of the "principles which form the basis of all genuine democracy" together

[76] State of Democracy, Human Rights and the Rule of Law: A democratic renewal for Europe, page 137
[77] See for instance the Preamble of the European Convention on Human Rights.
[78] Page 6 of the report of the Secretary General.
[79] See in this respect page 68 of the above-mentioned report of the Secretary General, which includes as measurement criteria for Good Governance efficiency and effectiveness, competence, efficiency and accountability of public institutions, to name but a few.
[80] Rule of Law is not merely a set of rules but the conjunction of a complex social situation that involves rules. On the one hand it requires that officials internalize the Law so that every official's act is according to Law. On the other hand, citizens are expected to obey the outcome of official-authoritative mandates that are produced according to a previously established Law. Acceptance and obedience of the law are social, not formal conditions for the Rule of Law.
[81] Cfr. https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-AD(2016)007-e

with individual freedom and political liberty. According to the Commission, "the notion of rule of law requires a system of certain and foreseeable law, where everyone has to be treated by all decision-makers with dignity, equality and rationality and in accordance with the laws, and to have the opportunity to challenge decisions before independent and impartial courts through fair procedures". Its core elements[82] are: 1) Legality, including a transparent, accountable and democratic process for enacting law; 2) Legal certainty; 3) Prohibition of arbitrariness; 4) Access to justice before independent and impartial courts, including judicial review of administrative acts; 5) Respect for human rights; and 6) Non-discrimination and equality before the law. However, the broad scope of some of these tenets creates challenges as regards their use to assess, in practice, if a particular AI system will impact Democracy and the Rule of Law[83].

This challenge may explain why the current general and AI specific HRIA frameworks by and large do not include impact assessment on Democracy and the Rule of Law. Only the self-assessment list of the Commission's High-Level Expert Group on AI includes a general question on Democracy[84].

However, we need to address this challenge to propose a complete, practical, and coherent assessment model. For example, the way AI impacts individual[85] human rights may differ considerably from impacts on democracy as a whole[86]. There could be cases for instance, where to small-scale individual human rights violations could correspond major threats to the good functioning of democratic institutions and processes. Hence, it is important to examine possible criteria and benchmarks for a HRDRIA.

(ii) Rights-grounded benchmarks for Democracy and the Rule of Law

Given the multitude of elements that make up democracy and the rule of law, and the difficulty of relating some of the underlying components of rule of law and democracy to clearly measurable parameters, it is proposed herewith a practical methodology to address this challenge, which consists, in first place, to use human rights as proxies to Democracy and the Rule of Law. In other words, to explore if there are human rights violations that could be used as proxies to Democracy and the Rule of Law. The idea is to explore if the magnitude of certain individual human rights violations could undermine the institutions and social practices that constitute Democracy and the Rule of Law. Or, alternatively, if there is a systemic connection between the assurance and

[82] The EU Rule of Law Report 2020 provides a consistent approach when indicating that « Under the rule of law, all public powers always act within the constraints set out by law, in accordance with the values of democracy and fundamental rights, and under the control of independent and impartial courts. The rule of law includes principles such as legality, implying a transparent, accountable, democratic and pluralistic process for enacting laws; legal certainty; prohibiting the arbitrary exercise of executive power; effective judicial protection by independent and impartial courts, effective judicial review including respect for fundamental rights; separation of powers; and equality before the law. These principles have been recognised by the European Court of Justice and the European Court of Human Rights. In addition, the Council of Europe has developed standards and issued opinions and recommendations which provide well-established guidance to promote and uphold the rule of law ». See https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1602583951529&uri=CELEX%3A52020DC0580

[83] In response to ES's request for clarification on "conceptual challenges" of extending HRIA to Democracy and Rule of Law.

[84] See page 20, which only considers whether it has an impact on society at large.

[85] Suggested by FRA

[86] Suggested by FRA

efficacy of certain Human Rights and democratic-rule-of-law institutions.

For instance, without the right of assembly, citizens would be deprived of an opportunity to come together publicly, to jointly deliberate and form a public opinion, or to collectively express, promote and defend their interests through peaceful action (for instance through the organisation of manifestations or protests). While democracy is not reducible to this right and the elements it enables, it nevertheless hinges upon the protection of this right, which can hence be seen as one of multiple potential 'democratic proxy' in an impact assessment model.

Human rights which can be considered as proxies to democracy are the right to respect for private and family life (article 8 ECHR), right to freedom of thought (article 9 ECHR), freedom of expression (article 10 ECHR), freedom of assembly and association (article 11 ECHR), right to non-discrimination (article 14 ECHR), the right to vote and to be elected to free and fair elections (Article 3, Protocol 1 to the ECHR).

When translating this into the context of AI, it is worth noting that some AI applications can censor specific groups excluding them from expressing ideas. This would be an extreme case where the lack of respect for the freedom of expression is a proxy for the strength of public discourse, a socio-institutional factor that supports democratic systems. But another more subtle example is the use of facial recognition targeting or biometrical footprint tracking systems deployed on public manifestations, political rallies and town halls. The mere use of those AI systems can inhibit the spontaneous gathering of people in fear of being watched, recorded and tracked. Some AI systems tracking political assembly manifestations imposes a cost, a psychological burden, on the mind of people willing otherwise to exercise their civil rights. The erosion of the freedom of assembly reduces the strength of a social factor that favours democracy: the active, public, discussion and support of political platforms, causes and ideas[87].

In yet another example of how AI applications can reduce the strength of social-institutional factors that favour democracy we find the actual formation process of deliberation prior to voting. AI applications can impact the set-up of representative institutions such is the case of certain voting applications. The voting apps aim to make that choice clearer by using a quiz to line up users' opinions with parties' policy positions[88]. But the design of the questions (in affirmative or negative form), the issues associated to a political party and even the interface of the apps can nudge voters to favour certain political party's candidate[89]. The issue is, from a sociological point of view, that when people rely on these applications their agency will become affected and thus the collective legitimacy of elected institutions will be diminished. Another example is found in AI applications for nudging voters, what is also termed political micro-targeting. When AI applications are deployed to nudge voters, they focus on extremely narrow messages assuring the maximization of voting intention in favour of a politician, party or cause. The collective consequence for democracy is the loss of plurality and diversity of information in the political

[87] In response to a comment from UK.
[88] Cf. Cedroni and Garzia, 2010 at https://www.researchgate.net/publication/285054035_Voting_Advice_Applications_in_Europe_The_State_of_the_Art
[89] Cf. Garzia and Marschall, 2019 at https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-620

debate[90].

By the same token, the Rule of Law can be connected to specific human rights, such that the violations or the diminishment of the safeguards of such rights are proxies for an impact on the Rule of Law. A rights-grounded benchmarking can be found in the Rule of Law Checklist adopted by the Venice Commission, which refers to the rights of equality before the law and non-discrimination, access to justice before independent and impartial courts, including judicial review of administrative acts, the right to a fair trial and the right to respect for private and family life, including protection of personal data.

The way in which these proxies may be operationalized when assessing an AI application is as follows[91].

The first stage is to identify the specific technology and/or the potential applications of the AI application in question.

The second stage is to identify the human rights involved either with Democracy or the Rule of Law and break it down to the component parts of it. Let's look at some examples.

Freedom of expression could be broken into components. A relevant component is the freedom to seek, receive and impart information and ideas of all kinds. This sets out the right to freely receive and impart information or ideas. This may be impaired by an AI application producing downranking or de-monetising content based on rules that haven't been collectively agreed[92]. Another, blunter example would be a AI system that assists in censoring independent media outlets, targeting, harassing or tracking journalists or media, restricting access to government information for the public and restricting internet access by blocking of sites. The components of the right can become benchmarks against which to measure the impact of an AI application.

Furthermore, freedom of thought can be broken down in the right not to reveal one's thoughts, not to be penalized for one's thoughts, and not to have one's thoughts manipulated. This risk of manipulation can be found in AI applications micro-targeting voters or generating deep fakes. If an application uses affect recognition too, the risks also arises that individuals can be penalized for their (perceived) thoughts or emotions.

The right of freedom of assembly and association can be, amongst other things, broken down in the right to organize, plan and have (online) interaction and to organize (spontaneous) assemblies. Here, for example, AI used to (temporarily) block websites or fora for this interaction may jeopardize this right. The same may be true for AI chatbots or trolls who are used to manufacture false assent or false discourse, both positive (around, say, national energy policy) or negative (around bodies critical of those policies)[93].

[90] Examples suggested by CDDG.
[91] Suggested by IBA.
[92] Example suggested by UK.
[93] Example suggested by UK.

The third stage is to consider the potential impact of the AI application on the right in question. HRIA regarding AI can assist in this.

The fourth stage is to map out indicators for assessing such impact. However, the indicators should be adapted to the impact of the specific AI application. Stages 2 and 3 may be undertaken concurrently, identifying the AI application, human rights and potential impacts until all eventualities have been assessed. After this the right may be broken down to its component parts and the indicators can be mapped.

This methodology is coherent with the analysis previously elaborated of the close interlink which exists among human rights, democracy and rule of law: the broad spectrum of human rights which would have to be examined as part of the HRDRIA assessment would also allow to cover many important substantive elements of democracy and the rule of law. Creating a checklist of questions to support the HRDRIA could be considered, addressing all “proxy rights” which have been indicated earlier as backbones of democracy and the rule of law.

Section II. Towards a model for performing Human Rights, Democracy and Rule of Law Impact Assessment of AI systems, building on the current experience of AI Human Rights Impact Assessment Systems

It has been analysed earlier which models for general or AI specific impact assessment exist and which issues have to be addressed in HRDRIA. All these frameworks apply a risk-based approach when analysing impact on rights. Thus, it makes sense to apply this approach for the human rights, democracy and rule of law impact assessment of AI as well. In this section, we provide some preliminary remarks and challenges to take into account and set out the steps that should be taken in the context of a HRDRIA.

(i) Preliminary remark on when a HRDRIA should be undertaken

It is questionable whether an HRDRIA should be undertaken for the development or deployment of all AI applications. Given the time and resources necessary to undertake such an assessment, we believe that a full-fledged HRDRIA should only be undertaken if an initial assessment shows heightened human rights risk.

Drawing on the experience of data protection impact assessment, in particular the Guidelines on Data Protection Impact assessments (DPIA) [94] it would be advisable to develop such initial criteria for the necessity to undertake HRDRIA, too. These criteria could be connected to the scale of an application, the way it systematically involves or interacts with humans and/or affects humans, the type and purpose of application or specific use cases (such as facial recognition, deep fake technologies, social networks) which could lead to human rights violation.

Should the initial assessment point to the need to carry out a fully-fledged HRDRIA, some further considerations should be kept in mind. Firstly, experience with conducting DPIAs has shown how difficult it can be for companies to undertake such an assessment as well as the difficulty arising from the fact that many different methodologies exist to do so, which make the outcomes of such DPIA's less comparable and thus, less easy to use and interpret. These difficulties may also hamper the development of a level playing field between actors in the member states. Thus, it is important that guidance is developed on when and how a fully-fledged HRDRIA should be undertaken as well as to work towards more comparable outcomes. Such guidelines have been developed for carrying out DPIA[95]. Therefore, it is advisable to do the same in the context of the HRDRIA.

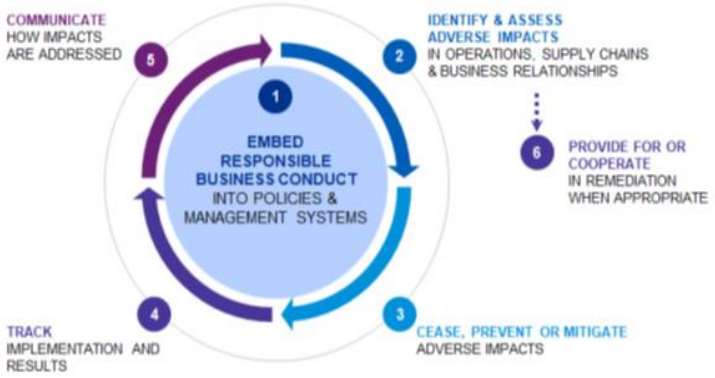
---

[94] Guidelines on Data Protection Impact Assessment (DPIA) provide guidance on determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, p. 8, which can be accessed at https://ec.europa.eu/newsroom/document.cfm?doc_id=47711. In particular a DPIA should take place when either of the following circumstances apply: "(a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;
(b) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offenses referred to in Article 10; or (c) a systematic monitoring of a publicly accessible area on a large scale".

[95] Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, which can be accessed at https://ec.europa.eu/newsroom/document.cfm?doc_id=47711.

ii) General challenges to undertake an AI-specific HRDRIA

An analysis of AI specific impact assessments reveals that, unlike general HRIA, most AI specific impact assessments do not focus on human rights as such but implement a broader approach and often include other, especially ethical, considerations. Furthermore, AI impact assessments generally do not include several features embedded in general HRIA, such as impact assessments throughout value chains, stakeholder engagement and implementation of grievance mechanisms as part of human rights due diligence[96]. General HRIA models typically include the six steps of human rights due diligence and these are, except for steps 1 and sometimes 5, by and large not embedded in AI-specific impact assessments[97]. However, it is important HRDRIA implements and documents these six steps.



That said, the digital footprint and complexity of AI systems makes the HRDRIA as a whole fundamentally different from the general HRIA. For example, an AI application may have global impact, whereas the traditional HRIA tends to focus on specific projects or production locations in supply chains. Engaging with stakeholders is, thus, more challenging in connection with these globally deployed AI applications, for example because of the different languages involved and public regulation and supervision in place as well as (technical) knowledge of stakeholders regarding the functioning of AI.

Specific AI assessment tools often do not clarify whether they entail a one-off assessment or an ongoing continuous learning and improvement process. Beyond this, the broader, not human rights centred approach implemented in AI-specific HRIA could also be a challenge to address, as mentioned earlier. In addition, an AI-specific HRIA may not be undertaken by the same department in charge of performing a general HRIA, which may even worsen this knowledge gap. Furthermore, more broad impact assessments may miss specific fundamental rights risks posed by AI. Therefore, collaboration or even embedding both types of assessments in one department may be helpful. Beyond this, it may be costly and create unnecessary administrative burdens if a private actor has to undertake two types of non-aligned impact assessments on fundamental

[96] However, the self-assessment list of the High Level Expert Group of the EU has included stakeholder consultation and participation (see p. 18) and an option for third parties (amongst which actors in the value chain) to report vulnerabilities (p. 22) but not an obligation to assess risk throughout the value chain and life cycle of AI.
[97] Compare this to the proposed methodology by Turkey.

rights including one general assessment and one AI specific. Furthermore, AI specific assessment may miss risks the general HRIA may identify and the other way round.

(iii) Proposed Methodology for the HRDRIA

The model for HRDRIA should provide a coherent and integrated approach for assessing adverse impact on human rights, democracy and the rule of law generated by AI systems, addressing simultaneously the risks arising from the specific and inherent characteristics of AI systems and the impact of such systems on human rights, rule of law and democracy.

Step 1: identifying relevant rights

As a first step, the HRDRIA model should allow for the identification of relevant human rights – including rights-proxies for democracy and the rule of law – that could be potentially adversely impacted by the AI application.

Step 2: assessing the impact on those rights

In light of the methodology presented earlier under Section B.2, the impact assessment should build on and integrate the currently existing general and AI-specific impact assessments[98]. This assessment should encompass both technical and non-technical aspects.

First, a more technical analysis should allow to assess the human rights challenges of an AI application as such, focusing on the underlying technology used and the specific technical features to be embedded in the AI application to prevent possible negative impact, such as explainability, transparency, cyber security and protection against usage beyond the intended application. This part could build on existing AI-specific impact assessments.

Second, the analysis should also contain a non-technical part, analysing the broader socio-technical environment in which the system is operating, competencies and skills required for the deployment and use of a given AI application. In addition, identifying, addressing and tracing risks of deploying AI in the overall value chain and life cycle (i.e., data labelling and other enrichment services, human content moderation[99]) should also be considered[100].

The distinguishing feature of a HRDRIA, which would deviate in part from the existing AI specific or general HRIA, would be that it includes specific analysis of impact on fundamental rights proxies which are directed towards the Rule of Law and Democracy, based on the broad spectrum of rights and the guidance provided in section I), B). This would prevent the duplication of existing models and unnecessary cost and administrative burden to private actors. It would also

[98] Beyond this it may build on underlying frameworks such as Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems: https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154.
[99] Suggested by CINGO
[100] Cf. OHCHR B-Tech project, Identifying and assessing human rights risks related to end-use, which can be accessed through https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf.

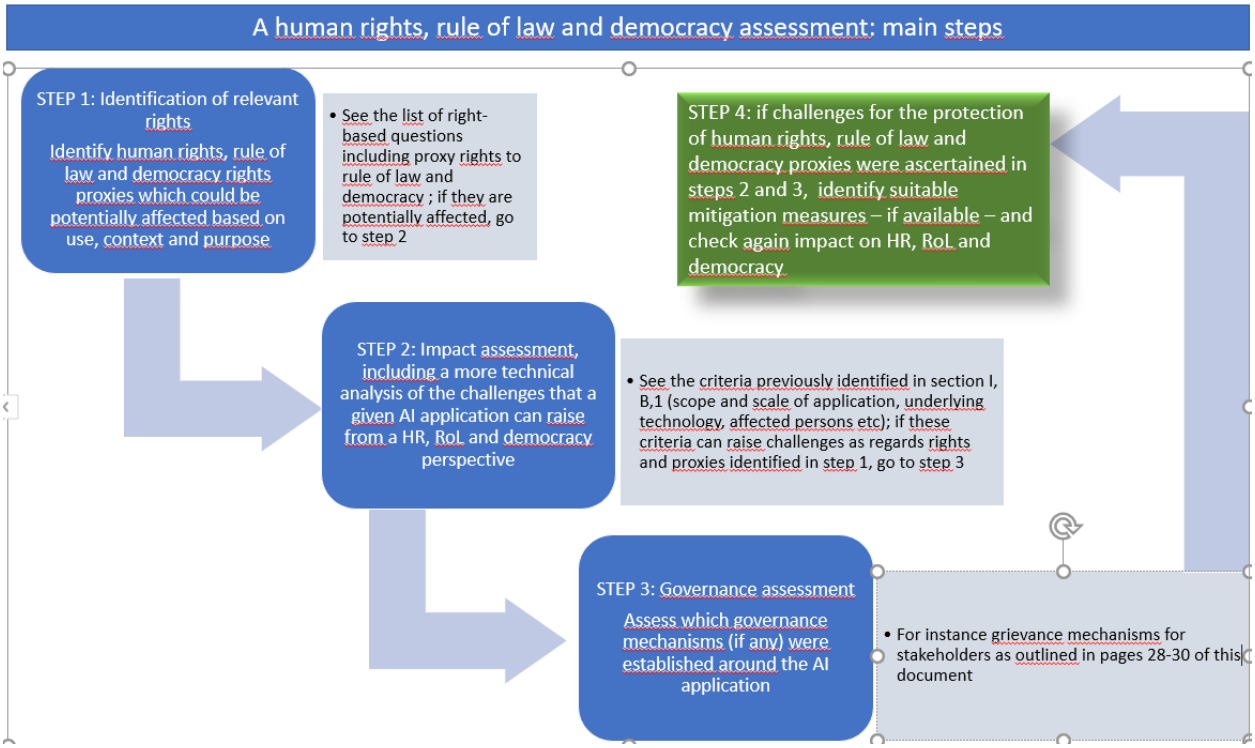
strengthen application of current AI specific and general HRIA frameworks in the AI arena.

Step 3: assessing governance mechanisms

In addition to the analysis of the impact on rights in step two, it is also useful to consider whether there are existing or established governance mechanisms that can help mitigate potential risks. In this regard, one must for instance consider stakeholder engagement and the establishment of, or participation in, grievance mechanisms to address possible complaints. To date, such grievance mechanisms in connection with AI are scarce. The opacity surrounding the use of AI in certain contexts contributes to this. For instance, a private or public entity deploying surveillance cameras may be identifiable, but the developer of the AI technology may be harder to identify. In this respect, lessons can be learned from general HRIA where this issue has been addressed. The outcomes of these grievance processes are an indispensable element of the continuous learning process of HRDRIA.

Step 4: continuous evaluation

It should finally be assessed whether identified governance mechanisms – or other mitigation measures – can provide a solution to mitigate the impact that was identified on human rights – including rights-based proxies of democracy and the rule of law. The impact assessment process should be continuously evaluated considering changes in the AI system, the environment of the system as well as the governance mechanisms surrounding the system. After all, an HRDRIA should be seen as an iterative process rather than a one-time step.



iv) Additional considerations

When applying general and AI-specific HRIA in an integrated and coherent manner in the AI arena it is important to note that inclusivity and meaningful participatory processes should be ensured. Moreover, unlike fundamental rights challenges in the traditional environments, collision of fundamental rights, for example, privacy vis-a-vis freedom of speech, may be more frequent in the AI arena. Thus, HRDRIA should include guidance regarding dealing with these collisions and how to balance conflicting fundamental rights.

It would have also to be considered whether additional features or safeguards have to be added to HRDRIA if AI is applied in the public sector, considering the adverse impact on human rights, democracy and the rule of law that the use of some AI applications in this sector might produce

Furthermore, capacity building on fundamental rights challenges may be required especially where the more technical part of HRDRIA is concerned. Thus, practical guidance as to how to apply the methodology described earlier, including a checklist for the operators to address all these aspects, is provided (see Appendix I) and could be finetuned in the coming months.

Continuous learning with a view to mitigating impact. Access to remedy should also shape HRDRIA in such a way that it enhances accountability towards relevant stakeholders. It is important to note that HRDRIA will provide an assessment of a certain moment in time, whereas human rights due diligence is an ongoing process of continuous learning. For example, HRDRIA may reveal that changes to a given AI application would be strongly recommended. When these changes are made, the HRDRIA may be repeated to check whether the impact has diminished.

Repeating HRDRIA may also become relevant if new human rights risks, or an increase of risk, is identified in the human rights' due diligence process. This could for instance happen if the AI application is used beyond its intended scope or when it becomes part of a larger network of systems. Thus, HRDRIA is not a one-off exercise but may have to be repeated. In this regard, it is important to review results and conclusions of consecutive HRDRIA's in a coherent manner and not as stand-alone exercises. Obviously, it is important to implement the results of HRDRIA, for example by revising the AI application or the dataset it makes use of and not stop after undertaking HRDRIA. HRDRIA should lead to action plans as to how address and mitigate the identified human rights risks.

Stakeholders' involvement. When undertaking HRDRIA the Guidance on Human Rights Impact Assessment of Digital Activities is a first framework to look at[101]. Although it is designed for digital technologies at large, it also has relevance for AI applications and emphasizes that HRDRIA should not only be conducted by developers of AI but also by those who sell, procure and deploy AI[102]. It adapts the steps of human rights due diligence mentioned in paragraph IA to digital technologies in particular[103]. It rightly emphasizes stakeholder engagement, also on the local level

[101] https://www.humanrights.dk/sites/humanrights.dk/files/media/document/A%20HRIA%20of%20Digital%20Activities%20-%20Introduction_ENG_accessible.pdf
[102] p. 7.
[103] p. 13.

where the impacts occur, is pivotal for HRDRIA[104]. Thus, HRDRIA should not only focus on technical aspects of AI or the organization developing, selling, procuring or deploying AI but also include engagement with relevant in- and external stakeholders. However, as explained hereinabove, this is required when heightened human rights risk is identified. The extent to which this type of engagement has to be conducted should be commensurate with the severity, scale and irremediability of the human rights impact of AI. The more severe and irremediable the impact is, or the larger its scale, the more extensive this stakeholder engagement has to be, especially with external stakeholders.

However, it may not be easy to identify the relevant external stakeholders and to engage with them. In connection with traditional HRIA some guidance has been developed on this[105]. In connection with AI, the scale of its application may make this identification and engagement more challenging as the scale of its application may be for example global, whereas traditional HRIA usually focuses on specific locations. If human rights risks are limited to specific countries or locations, stakeholder identification and engagement resembles the traditional HRIA stakeholder identification and engagement and more traditional approaches may be implemented.

Stakeholders should be involved in the entire HRDRIA process from planning and scoping, designing, stakeholder consultation until the final evaluation of it[106]. It is also important that it enables an ongoing dialogue and is accessible for stakeholders in terms of the technology and language used, as well as on potential human rights impact expected by an AI application. As mentioned earlier, the transparency of the HRDRIA is important and concerns the need to provide comprehensive and accessible information on the AI application itself to enable stakeholders to understand its potential impact on their human rights, as well as an appropriate time for the discussion of the findings of the HRDRIA.[107]. Such HRDRIA may build and enhance partnerships between business and stakeholders to identify human rights risk[108].

Access to remedy. Access to remedy is an important aspect to be considered should negative impacts be detected.[109]. In this regard it is important to assess whether an operational level grievance mechanism is in place which is compliant with the requirement of UNGP 31 and whether a company cooperates in legal proceedings. In terms of establishing and operating these grievance mechanisms, the company engaging with stakeholders on an AI application should also participate or develop a grievance mechanism for these stakeholders in case issues arise after the stakeholder engagement or are missed in the stakeholder engagement.

Beyond this HRIA's in the digital arena, the data protection impact assessment (DPIA) may provide guidance too. DPIA is required by article 25 GDPR in case of high risk (AI) applications. One important feature deployed in DPIA are questionnaires relating to specific risks in connection

[104] p. 15.
[105] See for example the Guidance Document for Social Accountability 8000 (SA8000:2014), p. 124 which can be accessed at https://sa-intl.org/wp-content/uploads/2020/02/SA8000-2014-Guidance-Document.pdf.
[106] p. 28. This for example also means consultation takes place when stakeholders are expected to be able to provide input, for example when women are not at work or younger people are not at school.
[107] p. 31.
[108] p. 20, where other advantages are identified as well.
[109] p. 36.

with data protection. The format used is the implementation of guiding questions which relate to relevant issues in connection with data protection, for example on fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation and security[110]. As mentioned earlier, a comparable approach may be advisable for HRDRIA and draw on the experience of the Trustworthy AI Assessment List developed by the High Level Expert Group on AI of the EU as part of their broader Ethics Guidelines for Trustworthy AI[111], which dwells in particular upon the risks which are typically connected to AI systems.

In conclusion, issues belonging to a more traditional HRIA and more technical aspects have to be combined in the HRDRIA, thus combining two dimensions, the human rights aspects and the technical dimension. A third dimension including governance should be addressed as well. The various elements and dimensions are in a simplified manner depicted in the following table appearing as Appendix I to the document. It should be noted that HRDRIA should be more elaborate where the answers to the guiding questions indicate AI risks are severe, have a large scale and/or are irremediable. Furthermore, the intensity of the HRDRIA may vary depending on whether a company causes or contributes to a human rights impact or is linked to it. Finally, HRDRIA is an ongoing process. Thus, if the answers to the guiding questions do not indicate heightened human rights risk, this may change if an AI system is used beyond its intended use case, to build another AI system or in case of development of other applications. In all those cases the guiding questions have to be answered again and these answers may then show heightened risk (see Table 1).

If HRDRIA is undertaken in the way just described it may reveal negative and also positive impacts on human rights, rule of law and democracy. HRDRIA is not designed to subsequently balance negative and positive human rights impacts, as this is a task of national authorities which may also depend on the specific features of legal systems. As mentioned earlier, HRDRIA does not implement an approach to balance human rights impact. It just identifies human rights risk.

[110] European Data Protection Supervisor, Accountability on the ground Part II: Data Protection Impact Assessments & Prior Consultation, p. 11-15, which can be accessed at https://edps.europa.eu/sites/edp/files/publication/18-02-06_accountability_on_the_ground_part_2_en.pdf.

[111] The self-assessment list can be accessed at https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, especially p. 5 and 6 are relevant in connection with human rights. The broader guidelines can be accessed at https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Section III. Synergies between HRDRIA and Compliance Mechanisms

A human rights, democracy and rule of law impact assessment as proposed above does not stand separate from other compliance mechanisms, such as certification and quality labelling, audits, regulatory sandboxes and continuous automated monitoring, which were mentioned the CAHAI Feasibility Study.

On the one hand, the above-mentioned compliance mechanisms may build on HRDRIA in the sense that they may require it for developing or deploying AI; even the level of public supervision may vary depending on the extent to which HRDRIA is deployed. It is important that HRDRIA and other compliance mechanisms are aligned as it would be unjustifiably costly and burdensome to require HRDRIA that diverge from public supervisory or regulatory approaches **(A)**.

On the other hand, a risk-based approach means that the risks posed by AI systems should be assessed and reviewed on a systematic and regular basis. Any mitigating measures should be specifically tailored to these risks, particularly risks affecting vulnerable and marginalized groups (e.g., BIPOC). In addition to the risk-based approach, where relevant, a precautionary approach, including potential prohibitions, should be considered[112] **(B)**

A. Aligning HRDRIA and Compliance Mechanisms

Ideally, HRDRIA could[113] become the foundation for compliance mechanisms and, thus, contribute to creating a level playing field between member states. In that regard, HRDRIA in conjunction with other compliance mechanisms could create an ecosystem of tools, a common framework for promoting trust and increasing transparency around the use of AI[114].

The compliance ecosystem is the result of the alignment of states' policy aimed to ensure national regulatory compliance with any future legal framework, on one hand, and the roles of different actors, on the other hand. Alignment includes, specifically for HRDRIA, two dimensions. The first dimension is common grounding; the model for performing HRDRIA should be common to assurers, developers, operators and users of AI applications. The second dimension is mutual reinforcement; the common use of HRDRIA should provide compliance and oversight incentives to the different actors and collectively, those actors should contribute in a complementary way to produce a new culture of AI applications respectful for Human Rights, Democracy and the Rule of Law.

An example of mutual reinforcement between HRDRIA and other compliance mechanisms is how the information available after performing a HRDRIA should be used by other actors. Member States should share information about the HRDRIA performed in their jurisdiction to promote the emergence of best practices, but also monitor and oversee AI applications that have a multiple jurisdictional scope. HRDRIA has a dynamic nature, meaning that an assessment should be

[112] Suggested by CINGO
[113] With heavy editing suggestions by EE.
[114] Cfr. Paragraph 155 of the Feasibility Study.

performed throughout all the AI application lifecycle[115]. At each point of the life cycle-assessment, information is generated that should be available to those organizations and private actors developing evidence-based standards and certifications.

All actors, assurers, developers, operators and users are key to bring about a new culture of compliance of AI systems with the new legal framework. When a HRDRIA reveals a significant risk of human rights impact, then a proper identification and involvement of the relevant stakeholders should follow so as to assess the level of impact on their human rights, as well as on democracy and the rule of law. However, HRDRIA should pay special attention to operators and users for two independent reasons. The first reason is that operators and users of AI applications are naturally interested in being informed about the potential harm and risks of a given AI application, whichever the source (individuals, organizations or governmental agencies) [116]. The second reason is that –as a necessary precondition of the new culture of human rights compliant AI– the existence, process, rationale, reasoning and possible outcome of algorithmic systems at individual and collective levels should be explained and clarified in a timely, impartial, easily-readable and accessible manner to individuals whose rights or legitimate interests may be affected, as well as to relevant public authorities.

Confidentiality considerations or trade secrets should not inhibit the implementation of effective human rights impact assessments and remedial routes[117].

Finally, compliance mechanisms can lead to prohibiting or strictly regulating the development and/or deployment of AI applications which HRDRIA has flagged as being problematic.

B. Alignment between Remedy Mechanisms and HRDRIA

Assessing risk of AI applications from the perspective of human rights, rule of law and democracy is related to remedy in three key dimensions that is useful to distinguish here. One dimension is the use of the HRDRIA as an information-empowering tool for users of AI applications and other key players of the compliance mechanisms ecosystem (i). A second dimension is the HRDRIA as a component of a broader human rights due diligence cycle (ii). A third dimension that relates HRDRIA with remedy is the common framework of HRDRIA to design remediation systems (iii).

(i) HRDRIA as an information-empowering asset

A fundamental challenge which is connected to (public) compliance mechanisms and human rights is the cost (and the entitlement) to challenge the design, development, deployment, operation and use of AI applications. In this regard, it is important that the information produced by a HRDRIA is understandable to experts and non-experts, and, equally important, that can be used to support any potential appeals and redress[118].

[115] Cfr. Paragraph 168 of the Feasibility Study.
[116] This is what the experience on Data Protection Impact Assessment reflects. Perhaps the experience in the data protection domain could serve as a practical advice for conducting HRDRIA. Opinion from the EDPS.
[117] Suggested by HomoDigitalis in refence to CM/Rec(2020)1 about contestability (p.9).
[118] Cfr. Paragraph 168 of the CAHAI Feasibility Study.

(ii) HRDRIA as a component of a broader human rights due diligence cycle

A human rights due diligence includes an impact assessment. In that regard, once a risk is identified, it should be addressed: implementing adequate actions to prevent the harm or to mitigate it. In the same manner, the result of a HRDRIA is the necessary input to make remedy available to stakeholders in case of human rights abuse in the AI arena.

Access to remedy, especially to abuses caused or contributed by private actors pose huge challenges, even outside the AI arena. Compliance should include this access to remedy aspect.

This is also compliant with the United Nations Guiding Principles on Business Human Rights and the OECD Guidelines, which include access to remedy. Private actors are expected to provide remedy if they have caused or contributed to a human rights abuse. Causing means they have created the impact themselves. Contributing means that they have facilitated or were involved in an abuse by another actor. But frameworks include a third category, linkage to abuse, where a private actor has not caused or contributed to an adverse impact. In such cases it does not have to provide access to remedy but may play a role in enabling access to remedy, for example by exercising leverage over third parties who caused or contributed to the impact. Thus, access to remedy has to be provided by developers and users of AI if an unwanted impact occurs, especially where a developer or user causes or contributes to an impact.

(iii) HRDRIA as a common ground to design remediation systems

In order to provide access to remedy judicial and non-judicial remediation systems have to be established, for example at the operational level of the developer and (commercial) user of AI[119]. The outcomes of these systems should also be used to feed into ongoing HRDRIA. In connection with the opaque nature of some AI applications it is important to incentivize creativity, innovation and collaboration between state-based and non-state-based remedy mechanisms in order to provide better access to remedy[120]. Thus, compliance mechanisms should also incentivize this type of access to remedy in the AI arena.

[119] OHCHR B-Tech Project, Access to remedy and the technology sector: a 'remedy ecosystem' approach, p. 2, which can be accessed at https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-ecosystem-approach.pdf. Cf. Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, under 4.4, which can be accessed at https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154.

[120] OHCHR B-Tech Project, Access to remedy and the technology sector: a 'remedy ecosystem' approach, p. 2, which can be accessed at https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-ecosystem-approach.pdf.

TABLE 1

Human rights aspect/ technical aspect	**High level of autonomy (human on the loop etc)**	**Usage outside intended use**	**Part of multiple deep neural networks/building on other systems/dual use**	**Explainability**	**Transparency/ reproducibility**	**Data quality**	**Robustness /security**	**Clarity whether AI system is used**	**Red flag/specific use cases (e.g. deep-fake/facial recognition/ weapons systems)**
Discrimination	If guiding questions show heightened risk: Stakeholder engagement at all stages/access to remedy in deployment phase	If questionnaire shows heightened risk: Stakeholder engagement when deploying for new use and access to remedy	If guiding questions show heightened risk: Stakeholder engagement at all stages as well as access to remedy	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	If guiding questions show heightened risk: Stakeholder engagement in building data set/access to remedy	Assessment by developer and user	Less important for this aspect	If allowed at all robust stakeholder engagement in all phases and access to remedy
Freedom of expression	If guiding questions show heightened risk: Stakeholder engagement at all stages/access to remedy in deployment phase	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use and access to remedy	If guiding questions show heightened risk: Stakeholder engagement at all stages as well as access to remedy	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	If guiding questions show heightened risk: Stakeholder engagement in building data set/access to remedy	Assessment by developer and user	Access to remedy	If allowed at all robust stakeholder engagement in all phases and access to remedy

Freedom of assembly and association	If guiding questions show heightened risk: Stakeholder engagement at all stages/access to remedy in deployment phase	Less important for this aspect	Less important for this aspect	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	Less important for this aspect	Assessment by developer and user	Less important for this aspect	If allowed at all robust stakeholder engagement in all phases and access to remedy
Privacy	Less important for this aspect	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use and access to remedy	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use and access to remedy	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	Less important for this aspect	Assessment by developer and user	Less important for this aspect	If allowed at all robust stakeholder engagement in all phases and access to remedy
Freedom of thought	If guiding questions show heightened risk: Stakeholder engagement at all stages/access to remedy in deployment phase	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use and access to remedy	If guiding questions show heightened risk: Stakeholder engagement at all stages as well as access to remedy	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	Less important aspect	Assessment by developer and user	If guiding questions show heightened risk: Stakeholder engagement at all stages as well as access to remedy	If allowed at all robust stakeholder engagement in all phases and access to remedy

Private and family life (beyond privacy)	If guiding questions show heightened risk: Stakeholder engagement at all stages/access to remedy in deployment phase	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use and access to remedy	If guiding questions show heightened risk: Stakeholder engagement at all stages as well as access to remedy	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	If guiding questions show heightened risk: Stakeholder engagement in building data set/access to remedy	Assessment by developer and user	If guiding questions show heightened risk: Stakeholder engagement at all stages as well as access to remedy	If allowed at all robust stakeholder engagement in all phases and access to remedy
Right to life	If guiding questions show heightened risk: Stakeholder engagement at all stages/access to remedy in deployment phase	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use and access to remedy	If guiding questions show heightened risk: Stakeholder engagement at all stages as well as access to remedy	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	If guiding questions show heightened risk: Stakeholder engagement in building data set/access to remedy	Assessment by developer and user	Less important aspect	Preferably not allowed at all, but if so robust stakeholder engagement in all phases and access to remedy
Right to fair trial	If guiding questions show heightened risk: Stakeholder engagement at all stages	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use	If guiding questions show heightened risk: Stakeholder engagement at all stages	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	If guiding questions show heightened risk: Stakeholder engagement in building data set	Assessment by developer and user	If guiding questions show heightened risk: Stakeholder engagement at all stages	Probably not applicable

Right to property	If guiding questions show heightened risk: Stakeholder engagement at all stages	If guiding questions show heightened risk: Stakeholder engagement when deploying for new use	If guiding questions show heightened risk: Stakeholder engagement at all stages	Pivotal for stakeholder engagement	Pivotal for access to remedy (especially with highly autonomous systems or multiple deep neural networks)	If guiding questions show heightened risk: Stakeholder engagement in building data set/access to remedy	Assessment by developer and user	Less relevant aspect	If allowed at all robust stakeholder engagement in all phases and access to remedy