COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 11 March 2021

CAHAI-PDG(2021)02
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

Draft prepared by Sub-Working Group 1[1]

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

Table of Contents

Introduction and Scope 3

Section I. Methodological considerations for a Human Rights, Democracy and Rule of Law assessment model 3

A. Human Rights Impact Assessment, sources, materials and experiences 5

1. The framework of HRA 6

(i) Documents (amongst others of Council of Europe as well as UN, OECD and EU) referring to such impact assessment 6

(ii) AI Impact assessments as part of broader human rights due diligence requirements 6

2. Existing human rights impact assessments 7

(i) General (e.g. Danish Institute and data from the FRA report) 7

(ii) Impact Assessments of different dimensions of AI, a review of existing proposals 7

B. Extending HRA with Democracy and the Rule of Law to the AI Domain 8

1. AI systems, Main Traits as Assessment Variables 9

(i) The Context of Application as a Variable of Assessing Impact 10

(ii) The Technologies of AI systems as a variable of assessing impact 11

(iii) Stakeholders as obligated input in assessing benefits and impacts 15

2. Democracy and the Rule of Law as dimensions of an integrated assessment model for AI systems 15

Section II. Towards a model for performing Human Rights, Democracy and Rule of Law Impact Assessment of AI systems 15

Section III. Synergies between HRDRA and Compliance Mechanisms 18

Introduction and Scope

The CHAI-PDF Sub-working group 1 has the following tasks:

1. Defining a methodology to carry out impact assessments of AI applications from the perspective of human rights, democracy, and the rule of law. The methodology should be based on relevant CoE standards and the work already undertaken in this field at the international and national level (see for instance in this regard CM/Rec (2020)1 on the Human Rights Impact of Algorithmic System, a publication from our colleague Heleen Janssen from the Netherlands, as well as the Human Rights Impact Assessment Toolbox developed by the Danish Institute for Human Rights).
2. Developing an impact assessment model.
3. Examining the complementarity of such an assessment with other compliance mechanisms outlined in chapter 9 of the feasibility study.

Following the task mandated to the sub-working group, this document follows a similar structure.

In part I, we outline the methodological considerations relevant to a Model of Human Rights, Democracy, and Rule of Law Assessment of Artificial Intelligence Systems (HRDRA) (I). In the second section, the document presents existing impact assessment tools and guidance which either relates to AI in particular or applies to human rights impact in general. The most relevant features of these frameworks are explored based on the methodological considerations displayed in section I (II). In the last part of the document, we correlate and analyze the connecting points between future HRDRA along the lines of the existing impact assessments mentioned in section II with the compliance mechanisms outlined in chapter 9 of the feasibility study (III).

As the time for delivering this paper was rather limited, it was not feasible to develop a complete model for HRDRA. Thus, the actual model should still be developed if necessary at all. This paper sets forward general guidance and suggestions for such a model and ways to further develop it in alignment with existing and future compliance mechanisms.

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 on Human Rights, Democracy, and the Rule of Law should build upon the already established practices of Human Rights Impact Assessment (HRA) experience (A). It should also acknowledge the origin of HRA which is human rights due diligence as included in the United Nations Guiding Principles on Business Human Rights and the OECD Guidelines for Multinational Enterprises.[2] This is important as human rights due diligence is an ongoing process and not a snapshot of a moment in time. Furthermore, it is not limited to the operations of a single company but covers the entire value chain. However, the peculiarities of AI systems present challenges to a simple import 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. There is, then, a need to discuss and analyze the elements of the HRA framework that travel well to the AI domain and how to extend the HRA framework to include Democracy and the Rule of Law (B).

However, it is important to note that the general HRA frameworks focus on adverse impacts of the operations of a company on human rights. In connection with AI the majority of the current impact assessment models implement the same approach. Therefore, it may be most logical to design a model for HRDRA which implements the same approach and focuses on adverse impact. Obviously, this does not imply AI generates adverse impacts only. AI has many advantages and creates a huge beneficial impact for mankind. It may even assist in the enjoyment and protection of human rights. This should not be neglected at all. Furthermore, the question may arise whether HRDRA should allow adverse impact of AI on human rights to be offset against beneficial impacts. This question will also be addressed.

Another relevant issue is whether HRDRA should apply to private actors, public actors or to both. Although both types of actors play a role in the enjoyment of fundamental rights, the role of public actors, who actively have to protect and not only respect fundamental rights, is considered to be different from that of private actors. Thus, it is not self-evidentiary that HRDRA can equally be undertaken by private and public actors. Beyond this, the general HRA frameworks we build on are designed for private actors as several of the AI specific frameworks also appear to be. Thus and initially, we explore a model for private actors. This does not imply HRDRA cannot be relevant for public actors, for example in public procurement which is addressed in sub-group 2, but it may be assessed in a later stage in which ways and to which extent HRDRA may also be applied to public actors.[3]

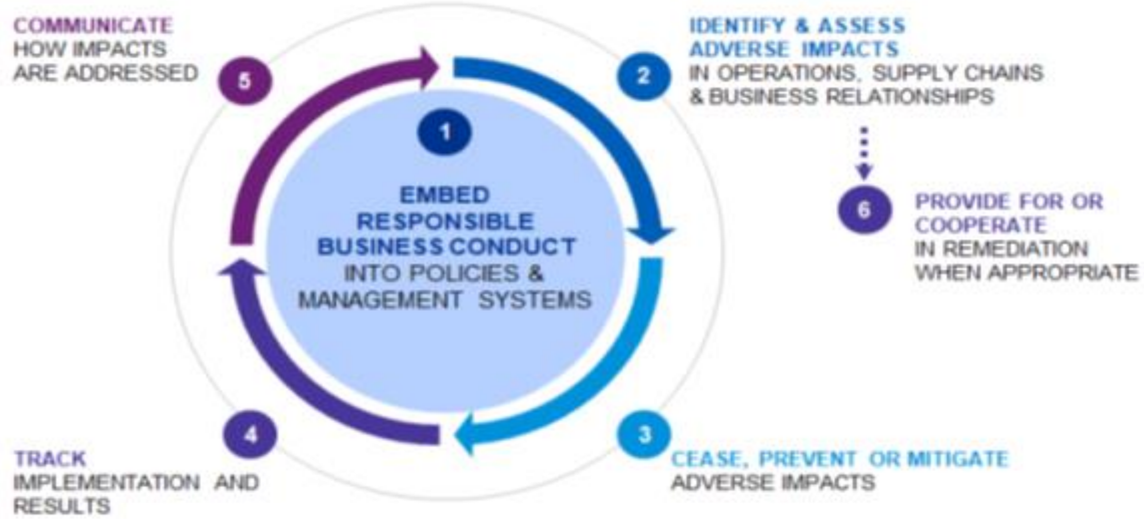
[2] Which can be accessed at https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf and http://www.oecd.org/daf/inv/mne/48004323.pdf.

[3] 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.

Hereinafter we will first analyze the sources and content of the traditional human rights impact assessments and provide a list of existing general and AI specific impact assessment and their mean features relevant in this context. These existing impact assessments raise the obvious question whether an additional model for HRDRA would duplicate existing frameworks and, thus, be superfluous. This will be analyzed after the analysis of these frameworks.

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

General not AI specific Human Rights Impact Assessments (HRA) draw from the international frameworks referred to in the foregoing and which require human rights due diligence. Human rights due diligence is ongoing and iterative process includes the following steps:[4]



General HRA is elaborated in specialized toolkits such as that of the Danish Institute for Human Rights. Conducting a human rights impact assessment is even mandatory in some countries. The framework of HRA practices and normativity provides relevant baselines for developing a model for HRDRA, and in this fashion, it is important to consider these existing general HRA frameworks. Recently, a specialized methodology for performing out a Data Protection Impact Assessment (DPIA) has emerged.[5] Although the proposal aims to develop a practical model to assess Automated decision-making (ADM) impact on fundamental rights, it is mostly centered 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

[4] 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

[5] Janssen, 2020: https://academic.oup.com/idpl/article/10/1/76/5788543

broader model to HRDRA 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.[6] In this fashion we will be looking to the existing human rights impact assessments both general and AI related. Then we will assess whether these frameworks also relate to the rule of law and democracy. Thereafter, we will analyse some basic features which most of these general or AI specific frameworks set forward and which may be relevant for a potential model for HRDRA.

1. The framework of HRA

(i) Documents (amongst others of Council of Europe as well as UN, OECD and EU[7]) referring to such impact assessment

EU High-Level Working Group (Ethics Guidelines for Trustworthy AI)[8]:

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

The Report of the EU Agency for Fundamental Rights[9]:

https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-fundamental-rights-report-2020_en.pdf

(ii) AI Impact assessments as part of broader human rights due diligence requirements

UN Guidelines on Business and Human Rights[10]

OECD Guidelines for Multinational Enterprises[11]

[6] 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.

[7] **Suggested by FRA**

[8] **Suggested by FRA**

[9] **Suggested by FRA**

[10] https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.

[11] http://www.oecd.org/daf/inv/mne/48004323.pdf.

2. Existing human rights impact assessments

(i) General (e.g. Danish Institute and data from the FRA report)[12]

Recommendation from the Commissioner for Human Rights, Unboxing Artificial Intelligence[13]: 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[14]: https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154

UN Guiding Principles of Business and Human Rights[15]
https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

Ada Lovelace Institute's Report "Examining the Black Box: Tools for assessing algorithmic systems":

https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf

Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts:

https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3736261_code2910023.pdf?abstractid=3736261&mirid=1

UN OHCHR B-Tech's foundation paper on "Identifying and Assessing Human Rights Risks Related to End-Use":

https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf

(ii) Impact Assessments of different dimensions of AI, a review of existing proposals

The American Civil Liberties Union has produced an Algorithmic Equity Toolkit:

https://www.aclu-wa.org/AEKit

GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC have developed the Ethics & Algorithms Toolkit:

https://ethicstoolkit.ai/

[12] **Suggested by FRA**
[13] **Suggested by AccessNow**
[14] **Suggested by AccessNow**
[15] **Suggested by FRA**

AI Now have developed a model for impact assessments, entitled Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability:

https://ainowinstitute.org/aiareport2018.pdf

Canadian Algorithmic Impact Assessment

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

Impact assessment methodology prepared by the AI Transparency Institute[16].

https://aitransparencyinstitute.shinyapps.io/ResponsibleAIIndex/?_ga=2.59765934.2034875787.1611733473-317365565.1611733473
https://aitransparencyinstitute.shinyapps.io/ResponIndex/?_ga=2.63239596.2034875787.1611733473-317365565.1611733473

B. Extending HRA with Democracy and the Rule of Law to the AI Domain

Today's AI systems present different characteristics that should be considered when developing a model for HRDRA, especially when applying general HRA frameworks. On the other hand, Democracy and the Rule of Law are equally essential dimensions of the assessment model for AI systems. However, the inclusion of Democracy and the Rule of Law as dimensions in an evaluation present conceptual challenges for the framework developed for HRA. For example, fundamental rights express values and common goods such as freedom of speech. In contrast, Democracy is not reducible (although it is connected) to a set of norms. Democracy has rules for accessing and exercising power and public decisions, but at the same time it has a cultural, institutional and a social dimension. 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. The same goes for the Rule of Law. It is not merely a set of rules but the conjunction of a complex social situation that involves rules. On the one hand that officials internalize the Law so that every official 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. Finally, not all laws are worthy of acceptance and obedience but those that are produced inside a democratic system in accordance with Human Rights.

[16] **Proposed by AI Transparency Institute**

However, these tenets are too vague to be of use, in practice, to assess if a particular AI system will impact Democracy and the Rule of Law[17]..

This challenge may explain why the current general and AI specific HRA frameworks do by and large not include impact assessment on Democracy and the Rule of Law. Only the self-assessment list of the High-Level Expert Group includes a general question on Democracy.[18] However, we need to address said challenges to propose a complete, practical, and coherent assessment model. For example, the way AI impacts on individual[19] human rights may differ in scale and ways through impact is caused from impacts on democracy as a whole.[20] Impact on the rule of law may be in between. As it may be hard to overcome the above mentioned challenges and develop a coherent model for HRDRA, we feel a practical method to address this issue and to prevent developing a completely new HRDRA chapter next to the one assessing impact on fundamental rights, is the use of a proxy which is connected to fundamental rights. Especially the fundamental rights of Freedom of Speech, Freedom of Association, Freedom to Receive Information, Prevention of Discrimination and Human Autonomy are the most relevant proxies. Thus, HRDRA including these fundamental rights may also have relevance to assess impact on the Rule of Law and Democracy.

1. AI systems, Main Traits as Assessment Variables

AI systems can be beneficial to individuals and societies as well as to pose risks of affecting individual Human Rights, Democracy and the Rule of Law (HRDL). In almost all technical decisions there is always a trade-off between the value that the technical system gives and the damage that a concrete system may produce. A more complete, full-fledged approach should balance and consider both aspects, the value that the AI system brings about and the magnitude and the chances of harming HRDL[21].

When performing a balancing approach between benefits and potential harm the analysis should establish clear and rich criteria for grading the risk and benefits associated with an AI System depending on, at least, the context (i), technology (ii), and stakeholders (iii). Within the grading scale of value versus potential harm, any HRDRA should guiding by the no-harm preemptive principle: if the assessment identifies a high HRDL risk that cannot be mitigated immediately, the AI system should neither be developed, implemented or used by any private or public authorities at least until effective measures are adopted to prevent potential or further HRDL risks (bans/moratoria of high-risk AI)[22]. Preemptive measures should be implemented

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

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

[19] **Suggested by FRA**

[20] **Suggested by FRA**

[21] **Suggested by ES**

[22] **Taken from CINGO Guidelines for impact assessment.**

regarding those AI applications which pose the biggest risk in terms of scale, severity and irremediability and the minimum benefit-value added[23].

In addition to the no-harm preemptive principle any HRDRA should be respectful of the proportionality principle: that is there may also be the need to have some type of assessment to determine whether a particular AI system merits a full HRDRA. This is important to ensure the proportionality of any legal framework[24].

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

An AI system application could pose a high risk to Human Rights, but the same AI system configured for a separate application could yield a lower (or no) risk. The geopolitical, social or economic context of the application of AI systems will help to calibrate[25] the level of risk and impact on human rights, rule of law and democracy[26]. To reduce the vagueness behind "context-of-application," we could think of context as a relative variable. The same applies for the value an AI system brings about. In certain contexts the AI system creates value and benefits that could be the opposite in a different context.

The context of an application can be relative to various sources. One source of context is 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, to attain certain actions/goals. In our example, the task of the AI system is to detect money laundering operations. However, the system's operator can change the system's use and thus change the context that is relevant for risk-assessment. The operator can repurpose the AI system from the previous example. 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. Users of AI systems are also a source of context for an application. 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.

[23] **In response to FRA clarification request on "first measures."**
[24] **Suggested by EE**
[25] **Observation by EE**
[26] **Suggested by CINGO**
[27] **Suggested by EE**

Thus, the adaptability of a particular AI system to different types of application is an important factor.

(ii) The 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 quintessential for any risk-assessment to consider the type of technology of the particular AI system. The field of AI has different approaches –from thought processes and reasoning to behavior, and in both cases with different measures of success, be it the fidelity to human performance or an ideal or rational performance[28]– and techniques related to those approaches[29]. The list of AI techniques has evolved and progresses continuously. And 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, when used to produce a system to detect counterfeit currency). To maintain HRDRA relevant in view of the ever-evolving AI techniques the assessing methodology should remain as 'algorithm-neutral' as possible[30] but not with a deep awareness of the differences and implications of each AI technique-approach.

The type of technology behind an AI system is relevant because it gives us eight, preliminary, dimensions that could signal out 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[31]. 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[32]– and the real-world scenario the system it's applied to. Some processes are just more chaotic and less predictable than others[33]. In our example, to know the inner works of the training technique and the type of algorithm is relevant to interpret the outcome of the AI system, but also to open a scrutiny on 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

[28] Rusel & Norvig, 2015. (Reference suggested by ES)

[29] Chowdhary, 2020.

[30] **Suggested by UK and AccessNow.**

[31] Classification algorithms for modeling classification predictive modeling problems. Classification predictive modeling 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)

[32] **Suggested by FRA**

[33] **Suggested by UK**

technology[34]. 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)[35]. The second case is a system with a low error rate that makes a fundamental error[36]. For example, an AI system that performs raw DNA sequencing based with a low error rate[37] but that could make a fundamental error that yields a fundamental error on an experimental drug target producing an undesirable effect.

The second dimension is reliability[38]. The level of consistent outputs that is expected behaviour of the AI system can be determined following the technology in which it is based[39]. 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"[40]. We can tolerate slight inaccuracy in an AI language translator system (NLP) but not so much in an autonomous driving vehicle[41].

The third dimension is traceability. A result, 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[42]. 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 neural

[34] Flynn et al, 2020: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156005/

[35] Sesardic, 2008: https://philpapers.org/archive/DEMGBS.pdf

[36] **Suggested by UK**

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

[38] 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

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

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

[41] 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

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

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[43]. Furthermore, AI is often developed building on earlier versions of other software. This may increase complexity in terms of the cause of specific impact[44].

The fourth dimension is explainability[45]. 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 behavior. 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[46].

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[47]. Data extraction adds another layer of potential impact, as does the creation of an ontological model[48] that represents raw information, as does any potential training scheme (from centralized to federated learning[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, an AI system used for machine calibration, completely autonomous, is highly

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

[44] 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)**

[45] **Proposed by CINGO.**

[46] 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)**

[47] **Observation by FRA**

[48] Braga et al, 2020: A MACHINE LEARNING ONTOLOGY - Frenxivfrenxiv.org › download

[49] Yang et al, 2018: https://arxiv.org/abs/1812.02903

[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**

predictable and reliable and exceeding operator-level precision[54]. If that systema 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.

The seventh dimension is security of AI systems. The higher the risk of hacks, adversarial or other security incidents (such as negative side effects of reinforcement learning[55]), the higher the risks of non-desired impact may become.

The eight dimension is Accessibility of Technology[56]. The type of technology in a social context of broad digital gaps could only increase the technological exclusion of the disenfranchise groups. Marginalized groups are often locked-out of tech because of lack of accessibility/digital inclusion.

In addition to the context and technology dimensions, there are also two parameters to consider when building a model of HRDRA, the role-in-relation-to the AI system and stage of the system within its life-cycle.

An HRDRA should be an ongoing assessment tool to be used throughout[57] 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[58].

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[59]. Just to provide an example: it is an overwhelming 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 pipe-line: configuration, automation, data collection, data verification, feature engineering, testing and debugging, resource management, model analysis, process management, metadata management, serving infrastructure, and monitoring[60]. The risk-assessment of an AI system at the "AI implementation" stage

[54] **Observation by ES**
[55] **Suggested by CINGO**
[56] **Proposed by CINGO**
[57] **Suggested by EE**
[58] **Observation by UK**
[59] **Observation by ES**
[60] MLOps: Continuous delivery and automation pipelines in machine learning

will be quite different from the previous stages of the AI (Machine Learning) life cycle: scoping-understanding[61].

(iii) Stakeholders as obligated input in assessing benefits and impacts

Community engagement is crucial for successful HRDRA[62]. There should be effective mechanisms to identify the stakeholders, within the relevant communities and to produce active participation regarding the assessment process.

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

The HRA on AI systems assumes a common reference to HR and in that regard, the inclusion of Democracy and the Rule of Law present stark challenges from the methodological point of view.
One idea open to discussion is to explore if there are 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 efficacy of certain Human Rights and democratic-rule-of-law institutions.

Section II. Towards a model for performing Human Rights, Democracy and Rule of Law Impact Assessment of AI systems

In the foregoing it has been analyzed which issues have to be addressed in HRDRA and which models for general or AI specific impacts assessments exist. All these frameworks apply a risk-based approach. Thus, it makes sense to apply this approach for the fundamental rights impact assessment of AI.

Analysing the AI specific impact assessments it is striking these by and large, and unlike general HRA, do not focus on fundamental rights as such but implement a broader approach and often include other, especially ethical, considerations.

[61] 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/

[62] **Proposed by HomoDigitalis**

Furthermore, they generally do not include several features embedded in general HRA, such as impact assessments throughout value chains, stakeholder engagement and implementation of grievance mechanisms as part of human rights due diligence.[63] General HRA models 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 specific AI HRA[64]. However, it is important HRDRA implements these six steps including reporting on it. Specific AI assessment tools often do not clarify whether these assessments entail a one off assessment or an ongoing continuous learning and improvement process. Beyond this, the broader not fundamental rights centered approach implemented in AI specific HRA may also hamper fundamental rights impact assessment as this concept may be less familiar to those who do not have expertise with general HRA. Thus, understanding what fundamental rights are and even more so assessing the impact of AI on these rights may be challenging for those not familiar with general HRA. In addition AI specific HRA may not be undertaken by the same department as general HRA, 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 rights including one general assessment and one AI specific. Furthermore, AI specific assessment may miss risks the general HRA may identify and the other way round.

Thus, HRDRA should provide a coherent and integrated approach to AI, which is also sufficiently adapted to identify risks to fundamental rights of future AI techniques, including autonomous systems. Thus, it should build on and integrate the currently existing general and AI specific impact assessments.[65] It includes a more technical part analyzing the fundamental rights challenges of an AI application as such, also in its technical environment (such as being part of networks) and its stage, which also requires specific technical features to be embedded in the AI application, such as explainability, transparency, cyber security and protection against usage beyond the intended application. This part could build on existing AI specific impact assessments. Beyond this, HRDRA should include a non-technical part analyzing the governance developed around AI development and identifying, addressing and tracing risks of deploying AI in value chains (i.e., data labelling and other enrichment services, human

[63] 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.

[64] **Compare this to the proposed methodology by Turkey.**

[65] 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.

content moderation for social media platforms, etc.)[66] and throughout its lifecycle.[67] This includes stakeholder engagement and the establishment of or participation in grievance mechanisms in which issues in connection with the deployment of AI can be raised. To date such grievance mechanisms in connection with AI are scarce and generally address specific fundamental rights such as privacy and freedom of speech. More general complaint mechanisms by and large lack. This may be caused by the opaque nature of the application of AI in other systems such as surveillance cameras. A private or public entity deploying these cameras may be identifiable, but the developer of the AI for facial recognition used by this entity may be harder to identify. Thus, a specific challenge regarding AI is addressing this issue. Here lessons can be learned from general HRA where this issue has been addressed for example regarding workers deeper in the value chain who may not be aware of the retailer selling the goods they produce and, thus, of the grievance mechanism this retailer has established. The outcomes of these grievance processes are an indispensable element of the continuous learning process of HRDRA.

A feature of HRDRA which may deviate in part from the existing AI specific or general HRA would be that it includes specific analysis of impact on fundamental rights proxies which are directed towards the Rule of Law and Democracy, notably Freedom of Speech, Freedom of Association, Freedom to Receive Information, Prevention of Discrimination and Human Autonomy. Especially in this context it is important HRDRA safeguards inclusivity and meaningful participatory processes

When applying general and AI specific HRA in an integrated and coherent manner in the AI arena it is important to note that, 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, HRDRA should include guidance regarding dealing with these collisions and how to balance conflicting fundamental rights.

It may be considered whether additional features or safeguards have to be added to HRDRA if AI is applied in the public sector.

The foregoing implies it may not be necessary to develop a (completely) new model for HRDRA. The model proposed by CAHAI may implement an approach which integrates and builds on existing AI specific and general HRA and explains how these fit together and can be integrated and applied to fundamental rights challenges posed by AI, also taking into consideration fundamental rights proxies regarding the Rule of Law and Democracy. This prevents duplication of existing models and unnecessary cost and administrative burden to private actors. It also strengthens application of

[66] Suggested by CINGO

[67] 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.

current AI specific and general HRA frameworks in the AI arena. Furthermore, capacity building on fundamental rights challenges may be required especially where the more technical part of HRDRA is concerned. Thus, in our view the HRDRA model would not include a completely new CAHAI HRDRA framework, but guidance on integration and coherent use of existing AI specific and general HRA frameworks, explaining how AI impacts (which) fundamental rights and what is needed in technical terms as well as in connection with general human rights due diligence to assess these impacts. The coming period may be used to elaborate this guidance.

Section III. Synergies between HRDRA and Compliance Mechanisms

The pursuit of a benefit-risk-based approach targeting the specific application context of AI systems has significant consequences on the compliance mechanisms. On the one hand, a risk-based approach means that the risks posed by AI systems should be assessed and reviewed on a systematic and regular basis. On the other hand 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.[68] Compliance mechanisms may build on HRDRA in the sense that they may require this for developing or deploying AI or may adapt public supervision to the extent to which HRDRA is deployed. It would seem to be important that compliance mechanisms are aligned with any HRDRA as it would be unjustifiably costly and burdensome to require HRDRA that diverged from public supervisory or regulatory approaches. Ideally, HRDRA could[69] become the foundation for compliance mechanisms and, thus, create a level playing field between member states.

As a necessary precondition, 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[70].

A fundamental challenge which is connected to (public) compliance mechanisms and fundamental rights is access to remedy in case of fundamental rights abuse in the AI

[68] **Suggested by CINGO**

[69] **With heavy editing suggestions by EE.**

[70] **Suggested by HomoDigitalis in refence to CM/Rec(2020)1 about contestability (p.9).**

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 fundamental 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. In connection with AI assessing causation or contribution may be even more challenging to assess than regarding regular human rights impact. This may be caused by the opaque nature of the application of AI in other systems such as surveillance cameras. A private or public entity deploying these cameras may be identifiable, but the developer of the AI for facial recognition used by this entity may be harder to identify. Technical features may be implemented to address this issue, but lessons may be learned from general HRA where this issue has been addressed for example regarding workers deeper in the value chain who may not be aware of the retailer selling the goods they produce and, thus, of the grievance mechanism this retailer has established.

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.[71] The outcomes of these systems should also be used to feed into ongoing HRDRA. 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.[72] Thus, compliance mechanisms should also incentivize this type of access to remedy in the AI arena.

Finally, compliance mechanisms may include prohibition or regulation of development and/or deployment of AI applications of which HRDRA has revealed these pose such

[71] 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.

[72] 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.

challenges to fundamental rights (red flags) that these are not allowed or under specific conditions and with government permits only.