DATA&
SOCIETY

European Center for
Not-for-Profit Law

Recommendations for Assessing AI Impacts to Human Rights, Democracy, and the Rule of Law

Introduction

We are at a turning point for the future of algorithmic accountability. Already, numerous jurisdictions have proposed legislation that would implement algorithmic impact assessments as a tool for bringing accountability to the algorithmic systems increasingly used across everyday life. Despite this heightened focus on impact assessments as an algorithmic governance mechanism, there is no standardised process for conducting such assessments that can be considered truly accountable.

This is especially urgent given the human rights impacts that artificial intelligence (AI) systems have on everyday life. From job applications to the delivery of public services, AI systems are being used in high-stakes settings where error, biases, and misuse of AI systems perpetuate economic and social inequalities.[1] While many governments and proponents of AI point to the potential for AI technology to stimulate economic growth, many people—particularly marginalized communities—are already subject to the worst excesses, mistakes, and harms perpetuated by discriminatory and extractive effects of these powerful algorithmic technologies.[2] Without a commitment to regulatory approaches that center accountability, transparency, and the protection and uplift of marginalized and vulnerable groups, AI will benefit the few while threatening the economic opportunity and social well-being of many.

1 For a list of AI failures, see Partnership on AI, AI Incident Database. Accessible at https://partnershiponai.org/aiincidentdatabase/

2 Democratize AI? How the proposed National AI Research Resource falls short, AI Now Institute and Data & Society, (October 5, 2021). https://medium.com/@AINowInstitute/democratize-ai-how-the-proposed-national-ai-research-resource-falls-short-96ae5f67ccfa

The Council of Europe's Ad Hoc Committee on AI (CAHAI) shares a concern over how best to govern AI systems. The **CAHAI's feasibility study (hereinafter "the study") concluded that measures must be put in place to radically reorient technology companies toward embedding human rights, democracy, and rule of law principles in their products and activities**.[3] These include the development of a "uniform model" for human rights impact assessments (HRIAs) of AI systems, which could form one tool in the toolbox for AI developers to meet their human rights due diligence obligations. The CAHAI has called this tool a Human Rights, Democracy, and Rule of Law Impact Assessment (HUDERIA).

As organizations that have published research on accountability and impact assessments for AI systems, we welcome the CAHAI's efforts to develop a legally binding framework on AI based on the Council of Europe's standards on human rights, democracy, and the rule of law.[4] During the consultation process for this framework, there was broad support for it to incorporate HUDERIAs as a governance mechanism.[5] However, as we emphasize in this report, the Council of Europe will need to undertake a significant amount of work to develop, test, and implement a workable methodology for HUDERIAs.

In this report, we explore three questions the CAHAI and Committee of Ministers must consider when developing HUDERIA methodology:

- **In which ways can HUDERIAs hold AI systems accountable? What are the limitations of HUDERIAs as an algorithmic accountability mechanism?**
- **What are the constitutive components that any HUDERIA method must account for?**
- **What are our recommendations for establishing HUDERIA methodologies and constructing mechanisms for meaningful democratic control of AI and related technologies?**

In addressing these questions and offering recommendations, our goal is to strengthen the HUDERIA proposal and expand our collective understanding of whether, and in which ways, impact assessments can work effectively to prevent further societal harm from algorithmic technology.

A key concern this letter seeks to address is the risk of a HUDERIA methodology becoming performative, weakened, or ineffective. The HUDERIA proposal must be designed and implemented in a way that elevates civil society and affected communities' concerns, instead of

3 Ad Hoc Committee on Artifiical Intellgience (CAHAI). 2020. Feasibility Study. Accessible at https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

4 After significant awareness raising efforts by the European Center for Not-for-Profit Law, civil society was the most represented sector with 31% submissions, followed by representatives of government and public administration (28%), academic and scientific community (20%) and private business sector (19%). ECNL supports civil society engagement through a written guide on how to answer CAHAI consultation survey and a workshop series on AI and human rights, (April 6, 2021), https://ecnl.org/news/have-your-say-ai-regulation-debate-deadline-may-9-2021

5 HUDERIAs are the preferred choice of governance mechanism for the majority of stakeholders: 81% of respondents selected "Human rights, rule of law and democracy impact assessments", followed by "audits and intersectional audits" (70%) and mechanisms of "certification and quality labelling" (51%). Council of Europe's public survey on how to define and regulate AI shows some unequivocal positions from civil society, (September 7, 2021). https://ecnl.org/news/ai-policy-message-delivered-will-states-listen

instrumentalizing them.[6] We encourage the CAHAI and the Committee of Ministers to work with these groups directly to develop a blueprint for the HUDERIA methodologies. Racialized persons, women and gender non-binary persons, LGBTQ+, disabled persons, persons of lower socio-economic status, and representatives from affected and marginalized communities must be included in formulating research priorities, definitions, and outcomes of the HUDERIA, including the decision on whether and when to deploy AI.

We hope the recommendations below will support the CAHAI and the Committee of Ministers in further developing a HUDERIA methodology that addresses how accountability is produced through impact assessments, and that meaningfully engages external stakeholders and affected communities in holding AI systems more accountable.

Limitations of HRIAs for Algorithmic Accountability

This section highlights important functions of HRIAs, situates HRIAs in context alongside other types of algorithmic accountability mechanisms, outlines several key limitations of existing approaches to HRIAs as an algorithmic governance and accountability mechanism, and offers recommendations for addressing these limitations for HUDERIAs.

Impact assessments

The CAHAI's proposed Human Rights, Democracy and Rule of Law Impact Assessment (HUDERIA) builds on a history of impact assessments used in a range of domains such as finance, environment, data protection, and health to consider the benefits and impacts of a business practice, technology or policy. Impact assessments broadly serve the following objectives:

- Providing an *ex ante* or *ex post* assessment of the potential or actual impacts of a technology, policy, or business practice.
- Creating a document of these impacts that can be shared with stakeholders, regulators, or members of the public.
- Providing a reflexive exercise for developers of a policy or technology to question what intended outcomes they hope to achieve, and what mitigative measures they may need to put in place to address potentially harmful outcomes.[7]
- Providing a mechanism for developers and policymakers to engage with a range of stakeholders who may be affected by that policy or technology.

6 On the instrumentalization of participation, see: Sloane, Mona, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. (2020). "Participation Is Not a Design Fix for Machine Learning." In *Proceedings of the 37th International Conference on Machine Learning*, 7. Vienna, Austria. https://arxiv.org/ftp/arxiv/papers/2007/2007.02423.pdf.

7 Reflexivity is a concept from the social sciences that refers to the process of examining one's own beliefs, practices, and judgments during the research process. This mindset is useful for developers of AI systems to adopt as a way to consider the potential wider societal impacts of their work. See Finlay, L. 1998. 'Reflexivity: an essential component for all research?', British Journal of Occupational Therapy, 61, 10: 453–456, and M. Brundage. 2019. Responsible Governance of Artificial Intelligence: An Assessment, Theoretical Framework, and Exploration. A Dissertation Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy, Arizona State University. Accessible at https://repository.asu.edu/attachments/223627/content/Brundage_asu_0010E_19562.pdf.

Algorithmic impact assessments (AIAs) have been proposed as a form of impact assessment directly applicable to AI systems, currently adopted as a requirement by the Government of Canada and being trialed elsewhere. However, they are still early in their conception and implementation, without a standardised form for general use.

Human Rights Impact Assessments

Human rights impact assessments (HRIAs) are a more widely used form of impact assessment that can be described as "a process for identifying, understanding, assessing and addressing the adverse effects of a [project, product, services, or activities] on the human rights enjoyment of impacted rightsholders."[8] HRIAs are grounded in the UNGPs, a non-binding framework that was unanimously endorsed by the United Nations Human Rights Council in 2011.[9] They have been used as a mechanism for assessing ex ante or ex post the human rights impacts of a business practice, technology, or policy. However, to our knowledge, HRIAs have not been widely applied to AI systems—existing publicly available examples within technology tend to look at a whole technology platform, and are typically *ex post* to deployment of the system.

Other kinds of algorithmic accountability practices

As the CAHAI and the Committee of Ministers develops its HUDERIA framework, it must fit within a wider constellation of emerging accountability and governance mechanisms for AI systems.[10] These include algorithmic auditing, regulatory inspection methods, voluntary principles and frameworks, legal prohibitions and moratoria on specific uses or applications,[11] public transparency initiatives such as transparency registers,[12] the establishment of external or independent oversight bodies to advise on ethical questions or procurement

8 Danish Institute for Human Rights. 2020. Guidance on Human Rights Impact Assessment of Digital Activities. Accessible at https://www.humanrights.dk/sites/humanrights.dk/files/media/document/A%20HRIA%20of%20Digital%20Activities%20-%20Introduction_ENG_accessible.pdf

9 The UNGPs prescribe that businesses should respect human rights by performing human rights due diligence (Guiding Principle 17), establishing a process for identifying and assessing adverse human rights impacts with which a business is involved (Guiding Principle 18), preventing and mitigating these impacts (Guiding Principle 19), tracking them (Guiding Principle 20), and communicating them (Guiding Principle 21). HRIAs are born out of Guiding Principle 18. See https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

10 For example, medical systems in the UK must undertake a Data Protection Impact Assessment as part of their development lifecycle, and some developers may be required to undergo an Equalities and Health Impact Assessment. See: Data protection impact assessments, *Information Commissioner's Office*, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/; Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021). Algorithmic Accountability for the Public Sector. Available at: https://www.opengovpartnership.org/documents/ algorithmic-accountability-public-sector/

11 Ex ante HRIAs for broad categories or types of AI systems (e.g. algorithmic-driven biometric surveillance) conducted by government agencies should inform legislative and regulatory measures. If the findings of the HRIA point to severe human rights impacts that should not be tolerated, there should be a robust discussion about whether the technology should be banned. See European Commission should consider risks of AI systems within a rights-based framework - as risks they pose to human rights, rule of law and democracy, European Center for Not-for-Profit Law, (March 23, 2021). https://ecnl.org/news/evaluating-risk-ai-systems-human-rights-tier-based-approach

12 See also CAHAI 9.2.2/165

decisions,[13] the establishment of rights to hearing and appeal of algorithmic decisions,[14] and legal conditions for public sector procurement of such systems.[15] These mechanisms and practices tend to focus on specific kinds of harm and risk assessment, while HRIAs tend to involve broader, human rights-focused assessments of harm.

Limitations of HRIAs

Although HRIAs, particularly as they might be applied to algorithmic technology, are an increasingly popular accountability mechanism, there are several important limitations to keep in mind in adapting them to a HUDERIA process for AI systems. Neither HRIAs nor HUDERIAs are a silver bullet for addressing potential human rights impacts or for assessing the human rights, democracy, and rule of law impacts of AI systems for several reasons:

- **HRIAs are not exhaustive.** It is difficult to assess all potential human rights impacts in an HRIA process. Some HRIAs are ex ante assessments of a business or technology's *potential* impacts, and others provide an authoritative *ex post* assessment of an AI system's *actual* impacts. Rarely does an HRIA provide both an *ex ante* and *ex post* assessment. While ongoing monitoring and follow up assessments can help address this challenge, most HRIA methodologies will only offer a means for some *ex ante* reflexive consideration captured in a document or statement, but a HUDERIA process should consider impacts before, during, and after the deployment of an AI system.
- **Currently, there is no single prescribed process or method for conducting HRIAs for AI systems.** The specific process and approach to impact assessments will depend on a range of factors, including the context in which the system is implemented and the type of AI system. Without regulatory or legislative guidance, or cultivating strong consensus-based norms for robust impact assessment, this may lead to scenarios in which AI developers, for example, conduct HRIAs when it is most convenient or safe for themselves, rather than when it's most helpful for rightsholders. A HUDERIA process that clearly delineates methods and relations of accountability can avoid such pitfalls.
- **HRIA methodologies must be adapted to best fit the needs of external stakeholders and must be responsive to the specific contexts.** Although governance mechanisms that call for impact assessments are attempting to scale these methods, HRIAs must always be adapted to specific contexts. This is particularly true for HUDERIAs. Furthermore, some components of a HUDERIA process must be left open-ended to allow for development of methods, practices, and knowledge about sector- or application-specific forms of algorithmic harm. There are currently few robust empirical methods for

13 The West Midlands Police and Crime Commissioner, for example, uses an ethics committee to assess department requests for new algorithmic policing systems, including the procurement and use of predictive policing and facial recognition systems. Private sector bodies have also turned to oversight mechanisms, such as the Facebook Oversight Board, which was established to provide independent judgments on Facebook content moderation decisions.

14 For example, Article 22 of the General Data Protection Regulation provides a right to human intervention in full automated decision making, and Article 15 provides a right to an explanation of an automated decision making system.

15 The CIty of Amsterdam, for example, published a set of "Standard Clauses for Municipalities for Fair Use of Algorithmic System" which included several requirements on providers of an AI system.

such *ex ante* assessments of algorithmic systems, although we anticipate a mandate to conduct such assessments would lead to proliferation of methods.

- **HRIAs alone do not create accountability.** HRIAs produce a document which can enable a change in practice, but it does not create an inherent binding commitment for any change to occur. The creation of statutory footing for the use of HUDERIAs in certain contexts could help address this issue by creating mandatory requirements and enforcement mechanisms for a failure to comply.
- **HRIAs can be co-opted.** Like any governance process, HRIAs run a risk of being conducted in bad faith to legitimise an existing business practice or policy. HUDERIAs face a similar risk of being co-opted in bad faith to justify the use of AI systems, without leading to any material change in the way that AI systems are designed, developed, deployed, or governed. If organizations that build and deploy algorithmic systems are allowed to define for themselves how they are evaluated, many harms to the public interest will proceed unchecked. There is a pressing need to develop a HUDERIA methodology that is rooted in the public interest and meaningfully engages stakeholders in defining what constitutes a human rights impact.

HUDERIAs are uniquely positioned for measuring the human rights impact of technology and are a critical component of algorithmic governance and accountability. However, the CAHAI and the Committee of Ministers must engage more deeply on the limitations outlined above and use them to set a research agenda. With that agenda in place, it becomes clearer what other policy mechanisms are needed to ensure a more holistic approach to algorithmic governance and accountability, and ultimately the mitigation and prevention of AI-driven human rights harms.

A framework for evaluating impact assessment processes

Regulating and accounting for the impacts of algorithmic systems present a special challenge to governments. HRIAs, as currently conducted, are an imperfect solution with important limitations. Despite these limitations, the CAHAI's goal to create a model for HUDERIAs can help create more accountable and rights-respecting AI systems.

To support the algorithmic accountability policy mechanisms and ensure that the development of HRIAs as a governance mechanism are effective, this section outlines a framework for evaluating impact assessment processes and offers HUDERIA-specific recommendations. This framework is drawn from an analysis of existing forms of impact assessment across many domains, and deepens our understanding of the mutual shaping of accountability and practices of evaluating harms through the assessment of their measurable impacts.[16]

This framework can work as a practical tool for regulators, advocates, public-interest technologists, technology companies, and critical scholars who are identifying, assessing, and acting upon algorithmic impacts.

16 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. (2021). "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." New York: Data & Society Research Institute. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

Defining Accountability

The underlying goal of HUDERIA is to help hold the developers of an AI system accountable to impacts it may have on human rights, democracy, and the rule of law. Here, we draw on public administrator scholar Professor Mark Boven's definition of accountability as a relational mechanism in which actors have a responsibility to explain and justify their actions to forums, who are empowered to make inquiries, pass judgement, and impose consequences on the actor.[17] **A HUDERIA process should create just this form of accountability relationship by asking actors to produce an account of how their systems may impact human rights, democracy, and the rule of law and by empowering appropriate government agencies and the public at large to act as a forum in evaluating those impacts.**

The various entities brought into a relationship through an accountability framework matters greatly, but in the context of algorithmic systems, we should think of that relationship as existing broadly between actors who wish to develop, operate, or use algorithmic systems and a forum of governmental agencies representing the public. Who represents the public in this capacity, and which actors ought to be obligated to make themselves accountable to that forum, must be worked out through policy.

Impact Assessment Framework components and recommendations for HUDERIA methods

This section outlines the ten constitutive components common in existing types of impact assessment practices. These components provide a framework for evaluating current and proposed algorithmic impact assessment regimes,[18] as well as for making recommendations aimed at developing an impact assessment framework that addresses the adverse impacts AI systems pose to human rights, democracy, and the rule of law.

1. Sources of Legitimacy: Impact assessments need to be legitimized either through legislation or within a set of norms that are officially recognized and publicly valued.

Human rights principles should be the root source of legitimacy for HUDERIAs, and made concrete through legislative mandates. The human rights framework should apply not only when identifying impacts, but also when assessing their severity and designing mitigation measures. As noted in the CAHAI feasibility study, relevant human rights instruments include the European Convention on Human Rights (ECHR), its Protocols, and the European Social Charter."[19] The full scope of relevant human rights as prescribed in the ECHR must be considered,[20] with an emphasis on salient human rights impacts.

To ensure the "Sources of Legitimacy" component is fully incorporated into HUDERIA methodologies, we recommend:

17 Bovens, Mark. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. West European Politics - WEST EUR POLIT. 33. 946-967. 10.1080/01402382.2010.486119.

18 Moss, et al. 2021

19 CAHAI Feasibility Study, 3.3.1 / 22.

20 The Digital Freedom Fund outlines the many human rights applicable in the digital sphere, and this analysis is also applicable in the context of AI systems. See https://digitalfreedomfund.org/digital-rights-are-human-rights/

- Reviewing and incorporating a wider range of international human rights treaties and and their monitoring bodies,[21] as well as reports by the United Nations Special Rapporteurs, the ILO Conventions and recommendations, and rights protected under the EU Charter of fundamental rights and relevant Council of Europe legal instruments,[22] including a future Convention on AI.
- Recognizing that a holistic approach must apply to the analysis of human rights impacts, and that impacts vary depending on the particular circumstances of each case.[23]
- Outlining and reviewing the challenges that the impacts of AI systems pose to democracy and the rule of law, including the concentration of power and monopolies in the AI sector, public-private partnerships, mis- and disinformation, business models and mass surveillance.[24]

2. Actors and Forum: Impact assessments are rooted in establishing an accountability relationship between actors that design or deploy a system and a forum that can allocate responsibility for potential consequences of such systems and demand changes in their design, deployment, and operation.

To achieve accountability, an actor must be accountable to an external forum that can pass judgement on the actions taken by an actor. In this case, the forum holds the actor accountable for the impacts of the AI system it builds or operates to human rights, democracy, and the rule of law.

A forum is made meaningful through its power to coordinate collective action, external to the decision-making power of the actor. Therefore, the forum should be empowered to do so by law or similarly robust social, political, and economic norms. Performative assessments, where actors have all the power and the forum plays only a cosmetic role—so-called "legal endogeneity"[25] or "human rights washing"[26]—can legitimize dangerous technologies, and do more harm than good.

21 For example, the International Bill of Human Rights (i.e., the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights), the ILO Core Labour Conventions, and other issue-specific and regional human rights instruments, as consistent with the UNPGs.

22 For example, The Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108). https://rm.coe.int/1680078b37.

23 Salient rights related to AI systems often include the following (see also CAHAI 3.3.1/22): Liberty and Security; Fair Trial; No Punishment without Law; Effective remedy (Art. 5, 6, 7, 13 ECHR); Private and Family Life; Physical, Psychological and Moral Integrity (Art. 8 ECHR); Freedom of expression, opinion and information (art. 10 ECHR); Freedom of assembly and association (Art. 11 ECHR); Equality and Non-Discrimination (Art. 14 ECHR, Protocol 12); Social and Economic Rights (Art. 2, 3, 5, 11, 12, 13 and 20 ESC); Privacy (Art. 17 ICCPR and CoE Convention 108); Human dignity (Preamble ECHR, Art. 1 EU Charter of Fundamental Rights).

24 CAHAI Feasibility Study, 3.3.2.

25 Edelman, Lauren B., and Shauhin A. Talesh. (2011). "To Comply or Not to Comply—That Isn't the Question: How Organizations Construct the Meaning of Compliance." In *Explaining Compliance*, by Christine Parker and Vibeke Nielsen. Edward Elgar Publishing. https://doi.org/10.4337/9780857938732.00011.

26 Latonero, Mark, and Aaina Agarwal. 2021. "Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar." Discussion Paper. Cambridge, MA: Carr Center for Human Rights Policy Harvard Kennedy School. https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf.

To ensure the "Actors and Forum" component is fully incorporated into the HUDERIA methodology, we recommend:

- Requiring that HUDERIAs conducted by an AI developer or deployer be submitted to a public-led body, as defined by the Council of Europe, which can act as a "forum" and approve or ask for changes to the product. It is essential to mandate external review and oversight of privately-conducted HUDERIAs, and determine which public authority is responsible for said oversight.
- Specifying what actions would fall under a compliance framework[27], aiming for high standards and building on the CAHAI feasibility study's acknowledgement of the responsibility of AI developers to consider actions they can take to increase compliance with a future legal framework.[28]
- Requiring that operators of AI systems (especially those operating in the public sector) are only able to procure AI systems that are compliant with the future Council of Europe legal framework, including the development of higher standards of transparency and accountability for AI systems used in the public sector, and to make public procurement conditional on conducting an effective and transparent HUDERIA, free from any restrictions for confidentiality or trade secret reasons.[29]
- Expanding on the proposal and guidelines for a joint certification body,[30] including clarifying what role this body will play, and outlining:
    - an operational approach that addresses questions around who will be part of the certification bodies that are establishing norms and standards for HUDERIAs;
    - the level of human rights expertise they are expected to have;
    - what role civil society will play;
    - how the lived experiences of affected communities and marginalized and vulnerable groups will be surveyed and incorporated into standards;
    - the plan for meaningfully consulting with external stakeholders.
- Ensuring that this body makes all relevant information publicly accessible and available.

27 The CAHAI feasibility study recognizes the responsibility of AI developers by stating that "[a]ctors building AI-enabled systems (both private and public sector) should consider actions they can take to increase compliance with a future legal framework." See CAHAI Feasibility Study, 9.22 / 165.

28 A non-exhaustive list of UNGPs-aligned measures would include, inter alia, creating human rights policies, implementing measures to improve transparency (e.g. consent and notification of the person(s) subjected to the AI system, public registers, etc.), conducting human rights due diligence (including HUDERIAs, audits, risk management strategies, documentation, and human rights due diligence for supply chains), and establishing internal grievance mechanisms.

29 CAHAI Feasibility Study, 118.

30 CAHAI Feasibility Study, 152.

3. Catalyzing Event: Points in the development and/or procurement process that trigger a requirement to conduct impact assessments. These can be mandated by law, or solicited voluntarily at any stage of a system's development life cycle.[31]

We welcome the recommendation in the study mandating that, where necessary and proportionate, AI developers and deployers conduct the HUDERIA "at regular intervals: prior to a new activity or relationship; prior to major decisions or changes in the operation (e.g. market entry, product launch, policy change, or wider changes to the business); in response to or anticipation of changes in the operating environment (e.g. rising social tensions); and periodically throughout the life of an activity or relationship."[32] Such instances can act as catalyzing events for triggering an impact assessment, but the UNGPs leave the decision of when and whether to conduct an HRIA at the discretion of individual businesses.

One question the Council of Europe must address is whether a catalyzing event will trigger a HUDERIA for all AI systems or only some. It may be the case that the sheer volume of AI systems in operation in Council of Europe Member Countries will require a triage process that focuses on certain systems over others. It is also the case that many possible AI systems are not trained on human data, do not make decisions about human lives, and have no human rights implications. A triage process could take the form of a human rights risk assessment that precedes the HUDERIA that flags certain "high risk" applications for review. This kind of approach raises more intrinsic challenges; how can the human rights risks of a system be known before a HUDERIA process occurs? What ought to "count" as an AI technology? And what will the process for assessing risk to human rights entail? The Council of Europe will have to ensure that the enumeration of high-risk applications is not exhaustive and can be applied to new and emerging AI technologies, or new and emerging understandings of the risk of existing AI technologies to human rights, democracy, and rule of law.

To ensure the "Catalyzing Event" component is fully incorporated into HUDERIA methodologies, we recommend:

- Mandating HUDERIAs in future Council of Europe legal instruments.
- Clearly identifying events which would require that an impact assessment be undertaken. The Council of Europe will have to decide whether all AI systems should undertake a HUDERIA, or whether only certain kinds of systems should.
- Developing a robust approach for determining what kinds of events trigger a HUDERIA, ensuring that new and emerging AI systems are included in scope.

4. Time Frame: Once triggered, the time frame is the period often mandated through law or mutual agreement between actors and the forum within which an impact assessment must be conducted. Most impact assessments are performed *ex ante*, before developing a system, but they can also be done *ex post* as an investigation of what went wrong.

HRIAs can be performed during the design and development process, or after deployment.

31 See DIHR 2.4 p.20-24 list of non-exhaustive examples. See also OHCHR B-tech. https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf

32 UNHR OHC. 2011. "Guiding Principles on Business and Human Rights." HR/PUB/11/04. New York: United Nations. https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.

However, they can also be done as an investigation into impacts revealed by the operation of the system in society, or on an ongoing basis to monitor emergent impacts on human rights. In stipulating a time frame for an impact assessment process, it should be kept in mind that HRIAs can be conducted prematurely, before important aspects of a system have been determined and/or implemented, or too late to address impacts that have been caused. The tempo and degree of scrutiny of such an ongoing assessment should balance the degree of risk against the burden and benefit of conducting such ongoing assessments.

To ensure the "Time Frame" component is fully incorporated into HUDERIA methodologies, we recommend:

- Ensuring that the time frame on which a HUDERIA is conducted covers all stages of the AI lifecycle, starting with the ideation stage and running through post-deployment.

5. Public Access: Achieving genuine transparency and accountability requires the ability of the public to scrutinize and contest an impact assessment's process and documentation. The broader the public access, the stronger is its potential to enact accountability.

While some sectors and applications have been exempted from regulatory scrutiny by claiming that they involve national security or active criminal investigations, high-risk contexts and applications such as predictive policing and sentencing demand more transparency, not less. Such domains are notoriously opaque, and are all-too-often excluded from requirements for public scrutiny that impact assessments can provide.

To ensure the "Public Access" component is fully incorporated into HUDERIA methodologies, we recommend:

- Clarifying that certain contexts, such as criminal justice, national security, or border control are not exempted from scrutiny and are subject to rigorous transparency and public consultation requirements.
- Ensuring that the output of HUDERIA processes is made available to the public by depositing it into public registers, providing public notice through press releases, social media posts, and other online bulletin boards that it is available and accessible, and depositing physical copies at libraries and other publicly accessible archives.
- Translating outputs into the language(s) of those communities most likely to be concerned with a product or technology's impacts.

6. Public Consultation: The conditions for solicitation of feedback should be from the broadest possible set of stakeholders in a system. Who constitutes this public and how they are consulted are critical questions for the success of an impact assessment.

The right to public participation is protected by article 25 of the International Covenant on Civil and Political Rights. Engaging the public in assessing impacts to their rights should be a key part of gathering evidence for HUDERIAs. The UNGPs explicitly state that the HRIA process should "[i]nvolve meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation" (Guiding Principle 18.b).

External stakeholders play a key role in ensuring that the forum leads to true accountability. Engagement with rightsholders is therefore central throughout the HUDERIA process, from scoping the assessment and identifying external stakeholders, to identifying impacts and later assessing them, to developing harm prevention measures.

Who constitutes this theoretical public and how they are consulted are both critical concerns for the success of a HUDERIA. Engaging a broad diversity of groups and stakeholders on an ongoing basis should be the minimum required standard to identify often forgotten or misunderstood harms. There is an urgent need to develop methodologies and practical guidance for engagement and addressing these challenges, as discussed below.

To ensure the "Public Consultation" component is fully incorporated into HUDERIA methodologies, we recommend:

- Ensuring that vulnerable and marginalized groups are adequately represented and, by requiring the actors carrying out the HUDERIA to facilitate such groups' participation in the impact assessment process and giving them meaningful decision-making power (e.g. right to veto, voting rights, etc.).
- Incorporating representation from stakeholders representing all Council of Europe Member-States, and ensuring that representatives from Member-States that are typically excluded, such as those from Eastern European and Balkan countries, or from rural and poor regions, are sufficiently supported.
- Requiring (or at least strongly recommending) that non-European actors be consulted where relevant, not only because the future Council of Europe convention can be implemented by Observer countries who aren't members, but also because AI systems often have global impacts.
- Providing civil society representatives with adequate resources (both financial and capacity building) to meaningfully participate in HUDERIA processes.
- Requiring that all information related to HUDERIAs that is not covered by trade secrets or confidentiality agreements (narrowly interpreted) be made publicly available and easily accessible.

7. Methods: These are standardized techniques of evaluating and foreseeing how a system would operate in the real-world. Most impact assessments have a roster of well-developed techniques that can be applied to foresee the potential consequences of deploying a system as impacts.

While some HRIA methodologies already exist which follow strict and high standards based on stakeholder engagement and transparency principles,[33] HUDERIA methods should be

33 See for example Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529; https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf p. 29; Alison Taylor, Charlotte Banchilon, Cecile Oger, & Jonathan Morris, Five-Step Approach to Stakeholder Engagement, (April 29, 2019). https://www.bsr.org/en/our-insights/report-view/stakeholder-engagement-five-step-approach-toolkit; https://www.humanrights.dk/sites/humanrights.dk/files/media/document/Cross-cutting_%20Stakeholder%20Engagement_ENG_accessible.pdf

developed further in partnership with affected communities and stakeholder groups.[34]

The methods deployed in an impact assessment process can be too disciplinarily narrow, limited to overly technical renderings of impact that do not speak to the socially embedded nature of algorithmic harms. They also can be too conceptually narrow, excluding the possibility that novel deployments of algorithmic systems might produce unforeseen categories of impact upon deployment. Furthermore, assessment methods may create too great a distance between the harm being measured and the category of impact through which that harm is measured. The success of a HUDERIA process will hinge on the ability of the process to incorporate these multiple complementary forms of expertise.

An impactful HUDERIA methodology is one that builds mechanisms for meaningful democratic oversight of AI technologies by affected communities. Community-based HRIAs typically use a bottom-up approach, which contributes to empowering affected communities in claiming their rights and ensuring accountability. Such assessments help to voice the concerns of affected individuals and local communities, putting them on a more equal footing with the public and private actors.

To ensure the "Methods" component is fully incorporated into HUDERIA methodologies, we recommend:

- Expanding beyond narrow, technocratic frames to ensure that technical assessments of the data, infrastructure, and performance characteristics of an algorithmic system be complemented by qualitative studies of the social contexts in which algorithmic systems are intended to be deployed, in order to better understand the likely context-specific impacts of a system.
- Employing ethnographic studies, participatory research methods, focus groups, and historical archival research to identify potential harms which would otherwise be undetectable from an inspection of system specifications.
- Adapting existing HRIA methods to algorithmic systems in an open-ended approach that can combine the evolving expertise of technical experts to document the operating characteristics of algorithmic systems, social scientists to document the qualitative impacts of systems, and community experts to document the contextual risks and harms posed by algorithmic systems.

8. Assessors: Impact assessments are conducted by assessors. The independence of assessors from the actor as well as the forum is crucial to how an assessment process identifies impacts, how those impacts relate to tangible harms, and how it acts as an accountability mechanism that avoids, minimizes, or mitigates such harms.

Every aspect of an impact assessment is deeply connected with who conducts the assessment. As evident in the case of HRIAs in the private sector, accountability can become severely limited when the accountable actor and the accountability forum are collapsed

34 Fundamental questions include (i) is the purpose of the technology a legitimate one?; (ii) if so, is the technology effective in achieving that purpose?; and (iii) even if it is effective, is it proportionate, i.e., is there no other less intrusive way to achieve the same result?

within the same organization.[35] To resolve this accountability concern, and because most private sector corporations do not maintain the in-house expertise needed to evaluate human rights impacts, HRIAs typically use external consultants as assessors.

Assessors' expertise ought to be expansive, including not only technical and legal expertise, but also that of social science experts, human rights experts, and other qualitative experts, as well as community members—particularly community members from marginalized and vulnerable groups—who hold expertise themselves on the experience of living with the harms of AI systems.

To ensure the "Assessors" component is fully incorporated into HUDERIA methodologies, we recommend:

- Strengthening the existing recommendation around voluntary or mandatory certification schemes by clearly identifying who should act as auditors and certifiers, mandating that they include human rights experts, and identifying what their relationship to developers, owners, operators, and subjects of AI systems ought to be.[36]
- Ensuring that the forum to which HUDERIAs are submitted remains independent of the developers, owners, and operators of AI systems, to prevent industry capture of the assessment process.
- Requiring developers who are tasked with conducting their own impact assessments to demonstrate relevant expertise in conducting the impact assessment process, partner with external organizations that hold this expertise, and submit their assessment to a publicly accountable body (a forum) who can pass judgement on and mandate changes to the scope of the overall assessment process.
- Requiring that assessors have access to the design specifications, training data, and any other artifacts they might need to conduct a robust impact assessment.
- Developing, where necessary, legal mechanisms that protect private companies' trade secrets and intellectual property, while still providing access to assessors.

9. Impacts: Impacts are proxies for the effects of the deployment of a system in the real world. They are what assessors document through their assessment methods, and are used to identify, measure, and ameliorate potential harms. When communicated to a forum, they enable the forum to mandate changes to the system being assessed, or to otherwise hold actors accountable.

Impact assessment is the task of determining what will be evaluated as a potential impact, what levels of such an impact are acceptable (and to whom), how such determination is made through gathering of necessary information, and finally, how the risk of an impact can be offset through financial compensation or other forms of redress.[37] Under the UNGPs, a company should identify and assess the actual or potential human rights impacts it caused,

35 Moss et al. 2021.

36 Certifiers and auditors can act as assessors of AI systems. As prescribed in the UNGPs, the HRIA process should "[d]raw on internal and/or independent external human rights expertise" (Guiding Principle 18.a).

37 Moss et al. 2021, p. 24.

contributed to, or were linked to through its activities, services, products, or relationships (Guiding Principle 18). **Identifying relevant impacts for assessment, and ensuring that emerging harms can be rendered assessable as impacts, is perhaps the most important component of a HUDERIA.**

In a HUDERIA, the degree of scrutiny should be linked to the AI system's level of risk to human rights. This raises an epistemological challenge for the Council of Europe; how do they determine what AI systems are worthy of more scrutiny under the HUDERIA than others? The severity of human rights impacts depends on their scope, scale and irremediability.[38] These factors should be determined by rightsholders and/or their legitimate representatives, as the goal is to determine impact to rightsholders and not to AI actors. The analysis should be adapted on a case-by-case basis mindful of the specific context at play; from geographic location and language, to demographic group, socio-political situation, and temporal scope.

Assessing the severity of an adverse human rights impact is an important element of determining preventive and remedy measures going forward, and ultimately, integrating the overall findings of a HUDERIA in the AI developer or deployer's activities, services and products.[39]

To ensure the "Impacts" component is fully incorporated into HUDERIA methodologies, we recommend:

- Prioritizing harm reduction and the adverse impacts to rights experienced by marginalized and vulnerable groups, as opposed to focusing on potential beneficial or harmful uses of AI systems for a few (privileged) groups.
- Taking a holistic and open approach to assessing the impacts of AI systems on a wide range of human rights, including collective rights, economic, social and cultural rights, and environmental rights.
- Bringing newly-identified harms within the scope of HUDERIA assessments by encouraging their inclusion as impacts.
- Ensuring that impacts to marginalized groups are elevated to a degree of concern such that the algorithm may not be used (without mitigating that impact) even if there are countervailing claims that the system will serve the state's economic and political interests.

38 Guidance on Human Rights Impact Assessment of Digital Activities (2020). *The Danish Institute for Human Rights*, pp. 33-34. https://www.humanrights.dk/sites/humanrights.dk/files/media/document/A%20HRIA%20of%20Digital%20Activities%20-%20Introduction_ENG_accessible.pdf.

39 The CAHAI study outlines examples of risk factors such as "the potential extent of the adverse effects on human rights, democracy and the rule of law; the likelihood or probability that an adverse impact occurs; the scale and ubiquity of such impact; its geographical reach; its temporal extension; and the extent to which the potential adverse effects are reversible. In addition, a number of AI-specific factors that can influence the risk level (such as the application's level of automation, the underlying AI technique, the availability of testing mechanisms, the level of opacity) can also be considered" (CAHAI 126). Other factors to consider are the AI developers and deployers' compliance and human rights due diligence mechanisms and the causal link between the AI system and the adverse impact. We encourage the Committee of Ministers to consider the factors for evaluating the risk levels of AI systems from a human rights-based approach as outlined by the European Center for Not-for-Profit Law in their 2021 paper: https://ecnl.org/news/evaluating-risk-ai-systems-human-rights-tier-based-approach.

10. Harms and Redress: Harms are lived experiences of the adverse consequences of a system's deployment and operation in the real-world. Some of these harms can be anticipated through impact assessments as impacts, others cannot be foreseen. To secure justice, redress procedures must be developed to complement any anticipated harm to secure justice.

As mentioned above, **harm prevention should be the fundamental and guiding principle of HUDERIAs**.[40] Core to this goal is the emphasis on marginalized and vulnerable groups and the precautionary principle: where it is difficult or impossible to assess the impact due to uncertainty, the impact should be determined as more severe rather than less.[41]

When thinking about harm and accountability in the context of AI systems, it's important to consider scale, scope, speed, exacerbated power inequities, technical robustness, reliability, and measures that prevent the risk of adversarial attacks or malicious uses. It is also critical that harms are adequately rendered as documentable impacts however they might manifest.

When thinking about redress in the context of AI systems, it's important to address whether the HUDERIA provides an opportunity to contest its process and outcome, and whether the HUDERIA supports future access to remedy should the adverse impact occur in practice, and the person(s) harmed by the AI system want to use information in the HUDERIA as evidence supporting their case.

To ensure the "Harms and Redress" component is fully incorporated into HUDERIA methodologies, we recommend:

- Centering the harm to individuals, communities, and the environment in the HUDERIA analysis, taking into consideration both individual and collective rights.
- Including an appeals process or a mechanism to challenge HUDERIA processes or outputs.[42]
- Requiring that HUDERIA methods include a process for reviewing the impacts in an iterative and ongoing way, with appropriate resources and capacity allocated for this purpose, to ensure adequate classification and assessment of impacts.
- Ensuring civil society organizations and external stakeholders have explicit standing to contest the findings of the HUDERIA or request additional information.

40 CAHAI Feasibility Study, 7.1.2 / 100.

41 The Danish Institute for Human Rights, 2020.

42 For example, impacts of AI systems to human rights, democracy or rule of law can be misclassified because harms cannot always be anticipated in advance. Misclassification can also occur when the risk levels of impacts are not adequately determined, or when impacts are excluded altogether. This is striking in the EU AI Act, where some technologies were determined to have a medium or low level of risk as opposed to high (e.g. emotion recognition technology and biometric categorisation for the purpose of predicting ethnicity, gender, political or sexual orientation). Similarly, some AI systems were determined to have high (but tolerable) risk, when they should have been prohibited altogether (e.g. risk assessments for criminal justice and asylum).

Recommendations for Public Accountability through HUDERIA

Our organizations hope to support the CAHAI and the Committee of Ministers in developing HUDERIA requirements, as well as deepening engagement across sectors on impact assessments as a mechanism for algorithmic governance and accountability. Establishing a human rights-based approach to impact assessments and algorithmic accountability would be a significant step forward in securing public accountability for the impacts AI technology has on society. To help translate this vision into a reality, we recommend that the CAHAI and the Committee of Ministers take the following steps.

- **Establish legal requirements regulating the HUDERIA process.** The CAHAI and the Committee of Ministers should establish legal requirements, as well as binding and non-binding guidelines for implementing these requirements, to determine what process a HUDERIA should follow, with an emphasis on meaningful participation of a range of stakeholders from diverse sectors. These requirements should clearly indicate at what stages of the life cycle of AI systems HUDERIAs are mandatory, and which events or circumstances would trigger a HUDERIA. The legal requirements should also clearly identify who should act as auditors and certifiers, mandate that they include human rights experts, and identify what their relationship to developers, owners, operators, and subjects of AI systems ought to be to ensure their independence.
- **Determine the criteria for assessing impacts to human rights, democracy, and rule of law.** The HUDERIA should prioritize harm reduction and the adverse impacts on marginalized and vulnerable groups. It should take a holistic approach and assess the impacts of AI systems on a wide range of human rights, including collective rights, economic, social and cultural rights, and environmental rights. The areas of scrutiny should be linked to the AI system's risk to human rights, democracy, and the rule of law *in situ*, as determined by rightsholders and/or their legitimate representatives. They should be assessed on a case-by-case basis, mindful of specific contexts at play, including geographical location, language, demographic group, socio-political factors, and temporal considerations.
- **Establish an oversight mechanism for HUDERIAs.** The CAHAI and the Committee of Ministers should mandate external, iterative, and ongoing review and oversight of privately-conducted HUDERIAs, and determine which public authority is responsible for said oversight. All information related to the oversight body and their assessments or review should be made publicly available and accessible. Civil society organizations and affected communities should have explicit standing to contest the findings of the HUDERIA and/or request additional information in the appeals mechanism, and receive adequate funding for this role. These mechanisms will ensure the HUDERIA methodology is not co-opted or watered down, and that developers of AI systems are not left to grade their own impacts.
- **Create sandboxes and pilot programs to spur HUDERIA development.** The CAHAI and the Committee of Ministers should provide a "regulatory sandbox" for developers, assessors, external stakeholders (including civil society, academics, and affected communities) as well as Member States to reach agreement on what a HUDERIA ought to consist of. A regulatory sandbox would indemnify and otherwise support

those who assume such risk and liabilities in developing an impact assessment process. In addition, the CAHAI and the Committee of Ministers should develop a pilot program to test HUDERIA methodologies in different sectors and fund ethnographic and other qualitative research that assesses the contents of impact assessments and whether impact assessments are having their desired effect on accountability. These pilots and studies should evaluate the effects of a HUDERIA process on a range of different actors, including developers and regulators, with an emphasis on affected communities and vulnerable and marginalized groups.

- **Develop methods for participatory inclusion of external stakeholders.** The CAHAI and the Committee of Ministers ought to develop methods for including the public, particularly marginalized and vulnerable groups, in the HUDERIA development process. Protecting the rights of marginalized and vulnerable groups must be central to the construction of HUDERIAs, along with transparency requirements for the public sector and high-risk use of AI systems. Engaging with the public and external stakeholders requires outreach and funding to ensure that industry is not the only party that influences the creation of a HUDERIA nor another approach to AI regulation. HUDERIAs also should require that technical assessments of the data, infrastructure, and performance characteristics of an algorithmic system be complemented by qualitative studies of the social contexts in which algorithmic systems are intended to be deployed.

- **Create an expanded research agenda.** Without additional investment in a research agenda to study the effectiveness of a HUDERIA framework, we risk fragmentation and confusion about what constitutes a high-quality impact assessment. We also risk furthering the dynamic in which AI developers choose the basis on which their own products are evaluated, resulting in narrow audits that do not allow for the uncovering of novel, context-specific harms or for evaluating companies' claims about the effectiveness of algorithmic products.

- **Integrating HUDERIA with other accountability mechanisms and impact assessments.** A key question that the CAHAI and the Committee of Ministers must address is how the HUDERIA and other forms of impact assessments should interact in key domains like healthcare, education, finance, and public service delivery. These domains may have existing sectoral impact assessment requirements for developers to comply with, just as developers may also have overlapping compliance requirements to conduct data protection impact assessments or other privacy impact assessments. The CAHAI and the Committee of Ministers must also consider what HUDERIAs limitations are, and how this framework complements or augments other kinds of accountability mechanisms like audits, prohibitions, and legal guidelines.

Acknowledgments

We are grateful to many experts and organizations whose work informed this paper. We would also like to extend our gratitude to the **Ada Lovelace Institute** for significant contributions to this research.

Data & Society is an independent nonprofit research organization. We believe that empirical evidence should directly inform the development and governance of new technology. We study the social implications of data and automation, producing original research to ground informed, evidence-based public debate about emerging technology.

As a resource, a catalyst and a convenor, we seek to inform and develop frames for discussion for these complex issues both through our own research and by supporting the voice of other stakeholders who face the adverse consequences of these technologies on an everyday basis.

Brittany Smith, Policy Director

Serena Oduro, Policy Research Analyst

Jake Metcalf, Program Director, AI on the Ground Initiative

Emanuel Moss, Postdoctoral Scholar

Emnet Tafesse, Research Analyst

The **European Center for Not-for-Profit Law (ECNL)** is a non-governmental organisation working to empower civil society. We aim to create legal and policy environments that enable individuals, movements and organisations to exercise and protect their civic freedoms and to put into action transformational ideas that address national and global challenges. We envision a space in which everyone can exercise their rights freely, work in solidarity, and shape their societies.

Our work on technology and human rights is a critical pillar of the organization. We aim to create safeguards for fundamental rights and civic freedoms in the context of emerging technologies, and streamline these safeguards into the development and functioning of AI systems.

Through research, advocacy, and capacity building, we advocate for protections of civic freedoms in AI-related regulation and empower civil society groups and activists to engage with technology companies, governments, as well as with regional and global institutions.

Francesca Fanucci, Senior Legal Advisor

Marlena Wisniak, Senior Legal Consultant

Vanja Skoric, Program Director