T20 Policy Brief

G20 BRASIL 2024

T20 Brasil 2024 Let's rethink the world

Task Force 05

INCLUSIVE DIGITAL TRANSFORMATION

Governing AI: From Principles to Implementation in a Global World

Alessandro Mantelero, Professor of Law & Technology, Polytechnic University of Turin (Italy)

Francesca Fanucci, Senior Legal Advisor, European Center for Non-Profit Law (Netherlands)

TF05

Abstract

While there is global agreement on the human-centred approach to AI, there are several challenges that could impede this objective in the absence of proper governance with meaningful multi-stakeholder participation. This is against a backdrop where asymmetries in the distribution and use of AI technology are generating a map that replaces the usual Global North/Global South dichotomy with one between AI provider/AI adopter countries.

To ensure responsible development of AI, it is therefore necessary to look beyond the principles in ethical charters and framework legal instruments and focus on the implementation of their underlying values. Governments should consider impact assessment methodologies and by-design approaches for this purpose and for the full achievement of the SDGs, and ensure relevant rights-holders participation.

This makes an analysis of possible implementation methodologies and best practices crucial. The resulting tools for human-centred design will consist of (i) questionnaires to help AI developers identify values, rights and freedoms at risk, as well as relevant groups of rights-holders involved, (ii) matrix models to assess the impact of AI and mitigate potential negative effects, and (iii) key guidelines for meaningful participatory assessment.

Building on the ongoing debate at international and national levels and with a view to responsible innovation, the G20 should (i) call for a multi-layered impact assessment of AI covering both legal and socio-ethical issues; (ii) define best practices for this exercise, including the role of stakeholder participation in co-design of AI systems; (iii) promote transparency and accountability in risk management; (iv) elaborate on the legal and socio-ethical component of assessment, relying on universal operational solutions and quantification for HRIA, while promoting meaningful multi-stakeholder participation to bridge the gap between ethical principles and concrete practices of developers.

Diagnosis of the Issue

At the G20 Summit in New Delhi (G20 2023, para 61), G20 leaders discussed the positive effects and critical issues associated with the development of AI, and emphasised the key role of the protection of human rights in ensuring responsible AI. In order to unlock the full potential of AI and mitigate negative consequences, the leaders believe that the pro-innovation AI government must pay attention to risk mitigation.

The same vision, combining human rights and the risk-based approach, is common to several AI regulation proposals, such as the Brazilian bills and the AI Act recently approved by the EU. In addition, international organisations, such as the Council of Europe (Council of Europe 2024) and the UN (United Nations, General Assembly 2024), have followed the same path in addressing the challenges posed by various AI applications. Finally, the approaches based on soft-law instruments rely largely on lists of principles that reflect core human rights principles and the protection of human rights (Mantelero 2022, Chs 2 and 3).

Given the potential pervasive use of AI, from health care to social services, decision making processes based on this technology are increasingly shaping large sectors of our societies. In order to rely on these systems, it is therefore crucial to ensure that they do not have any adverse impact on human rights, just as the different levels of access to this technology and the key role of the main actors in AI do not exacerbate existing limitations and disparities in the enjoyment of human rights or create new barrier to exercising them.

Against this background, the potential impact on human rights is a crucial policy issue in the governance of AI, regardless of the hard or soft law instruments used, and needs to be properly addressed through a risk-based approach, as suggested at the last G20 summit.

In this regard, a purely principles-based approach to AI governance cannot fully address the risks of AI. AI guiding principles, most of which are based on human rights, need to be implemented in the concrete design of AI systems to ensure their effectiveness in protecting individual and collective rights. This aspect of AI governance on-the-ground is the least explored area, as confirmed by the lack of specific methodological guidance in the regulatory proposals mentioned above.

Given the role of the risk-based approach (G7 2023) in this context, and the difficulties of simply replicating the existing Human Rights Impact Assessment (HRIA) experience in the AI context, due to the different nature of this ex post tool, it is therefore crucial to contribute to the ongoing debate on AI governance by providing methodological guidelines on this key element of local, national and international AI strategies.

The aim of this Policy Brief is to outline the main components of a methodological approach to AI governance grounded in a prior assessment of potential risks, not only to prevent human rights violations, but also to actively promote such rights and the full achievement of the SDGs.

With a view to inclusive and human-centric AI, as outlined in the last G20 summit, a meaningful participatory approach to risk assessment is crucial to better framing the risks, engage marginalised groups, and increase trust in AI. The guiding principles that governments should consider when designing both the impact assessment tools and related participatory tools are therefore discussed in the following sections, with a focus on the potential impacts of AI on human rights.

Recommendations

In line with the two complementary focuses of this Policy Brief, namely the risk-based and the meaningful participatory approach, the following recommendations addresses them separately.

1. Recommendations on the risk-based approach and human rights impact assessment in AI

Although human rights impact assessment is not a new approach to risk management in rights protection (United Nations - Human Rights Council 2011), its application to the specific field of digital technologies and AI, with its peculiarities, is a recent development.[1]

It is only in recent years that more attention has been paid to risk-based methodologies, partly as a result of the shift from a purely ethical approach to the regulation of AI. From a policy perspective, it is worth noting that the risk assessment cannot be reduced to a mere descriptive exercise, in which potentially affected rights are outlined in general terms and some measures are proposed, without any evidence of the link between the two in terms of appropriateness and effectiveness of the measures in reducing the estimated level of risk.

[1] A first contribution to this was made by The Danish Institute for Human Rights in 2020 (The Danish Institute for Human Rights, *Guidance on HRIA of Digital Activities*, 2020). Although not specifically focused on AI, this guidance opened the debate on the use of HRIA outside its traditional domain and in relation to digital technologies.

1.1 Outline a clear procedure

In line with HRIA and risk management in general, the HRIA in AI must include at least (i) a planning and scoping phase, focusing on the main characteristics of the product/service and the context in which it will be placed; (ii) a data collection and risk analysis phase, identifying potential risks and estimating their potential impact on fundamental rights; (iii) a risk management phase, adopting appropriate measures to prevent or mitigate these risks and testing their effectiveness.

1.2. Define a sound methodological approach

Taking into account the procedural steps outlined above, best practices in the development of the first phase (planning and scoping) should emphasise the relevance of the fundamental question of alternatives to AI, which could also be addressed through a SWOT (strengths, weaknesses, opportunities, and threats) analysis, and needs to consider the limitations that affect generic checklists, which should be combined with appropriate contextualisation provided by the experts carrying out the HRIA. In designing a contextualised checklist, inspiration can be drawn from existing models, while considering their limitations and not excluding other methods of analysis.

In the subsequent data collection and risk analysis phase, impact assessment needs to consider all relevant risk dimensions, carefully selecting the key variables and their combination, and avoiding unjustified overweighting. Attention must therefore be paid to the creation of risk indices and indicators, and to the definition of the variables used to quantify the impact on the human rights potentially affected. The result must be consistent with the legal framework and the theory of fundamental rights, identify the rightsholders affected, exclude cumulative assessment of the impacts on different rights and the

fragmentation of impacted rights into different components, thus avoiding double counting, both as an individual element and as a component of the general right.

All of these aspects of operationalising the assessment of impacts on fundamental rights underline its necessarily expert-based nature.

Finally, the use of variables demonstrates the dependence of the results on contextual factors that may change over time during the life cycle of AI systems. Risk assessment should therefore take into account any changes in both the AI systems and the context where they operate.

1.3 Appropriate methods for estimating the level of risk

In order to implement the risk assessment effectively and to adopt an approach that make it possible to compare different AI design options and different AI solutions, the use of matrices to construct risk indices is recommended, also because of their relative ease of use and explainability.

As a risk matrix is a graph that combines two dimensions using colors to reflect different levels of risk, they are useful for assessing indices generated by different variables. When using matrices to assess impacts, the relationship between the relevant risk components should be carefully considered in line with the fundamental rights framework. This suggests the need to avoid purely mathematical approaches to scaling, to be transparent in the scaling criteria, and to clearly define the relationship between the risk components.

1.4. Defining an effective and robust accountability model and a life-cycle approach to AI

The resulting tools for human-centred design will consist of (i) questionnaires to help AI developers identify the values, rights and freedoms at risk and the relevant groups of people affected, (ii) matrix models to assess the impact of AI and mitigate potential negative effects, and (iii) key guidelines for participatory assessment.

Building on the ongoing debate at international and national levels, and with a view to responsible innovation, it is also possible to broaden the scope of impact assessment in AI beyond human rights, through a multi-layered AI impact assessment covering both legal and socio-ethical issues, including the SGGs.

Finally, HRIA in AI requires an expert-based approach to carry out the various stages of contextual assessment. This is even more the case for the broader assessment just mentioned, where the role of expert and participation are crucial to understand contextual societal issues and values. Furthermore, experts can actively facilitate participatory assessment and co-design best practices in AI.

2. Recommendations on multi-stakeholders' meaningful engagement in impact assessment

The UN Guiding Principles on Business and Human Rights explicitly acknowledge that human rights impact assessments should "[i]nvolve meaningful consultation with potentially affected groups and other relevant stakeholders" (United Nations - Human Rights Council 2011, 18.b). Indeed, engaging external stakeholders in assessing the risks and impacts that the design, development and deployment of AI may have on their rights should be a key part of HRIA, in order to ensure true accountability and trust in the

process. Therefore, proper methodologies and practical guidance for their engagement should be developed as a priority.

2.1 Ensuring inclusive representation

When identifying the relevant stakeholders for meaningful participation in the HRIA, particular focus should be given to ensuring that that vulnerable and marginalized groups are adequately represented, in line with the 2030 SDG Agenda targets on decision-making (United Nations 2015, SDG Goals 5.5 and 16.7).

2.2. Meaningful contribution to HRIA

Consultation of rights holders in the risk and impact assessment process should not be reduced to a perfunctory box-ticking exercise but should provide meaningful access to the discussion with the experts and policy makers, especially when AI systems are procured for design, development and use in the public sector.

2.3. Facilitating participation with resources and capacity building

Engaging with the public and external stakeholders requires outreach and funding to ensure that industry and institutions are not the only parties that influence the creation of a HRIA nor another approach to AI regulation. Providing civil society representatives with adequate resources (both financial and capacity building) to meaningfully participate in HRIA processes.

2.4. Transparency and access to relevant information

An impactful HRIA 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. In addition, protecting the rights of marginalized and vulnerable groups must be central to the construction of HRIAs, along with transparency requirements for the public sector and high-risk use of AI systems.

HRIAs should also 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.

It is important to promote transparency by making all information related to HRIA that is not covered by trade secrets or confidentiality agreements publicly available and easily accessible through public transparency initiatives such as transparency registers. In addition, public procurement should be conditional on conducting an effective and transparent HRIA and free from any restrictions based on confidentiality or trade secrets.

Achieving genuine transparency and accountability requires the ability of the public to scrutinize and contest an impact assessment's process and documentation, thereby enhancing accountability through public access. While some sectors and applications have been exempted from regulatory scrutiny, such as national security and active criminal investigations, high-risk contexts and applications such as predictive policing and sentencing demand greater transparency and should not be excluded from appropriate forms of public scrutiny, which impact assessments can provide.

To ensure that the public access component is fully incorporated into HRIA methodologies, we recommend: (i) clarifying that certain contexts, such as criminal justice, national security, or border control are not fully exempted from scrutiny and are subject to specific rigorous transparency and public consultation requirements; (ii)

ensuring that the output of HRIA processes is made available to the public by depositing it into public registers and using other possible complementary forms of communication to enhance outreach;[2] (iii) translating outputs into the language(s) of those communities most likely to be concerned with a product or technology's impacts.

[2] E.g., providing public notice through press releases, social media posts, and other online bulletin boards that it is available and accessible, and depositing physical copies of the HRIA output at libraries and other publicly accessible archives.

Scenario of Outcomes

The recommendations discussed in the previous section can make a significant contribution to ensuring the responsible development, deployment and use of AI, focusing on the protection of human rights, transparency and explainability, fairness, accountability, and socio-ethical issues (G20, *New Delhi Leaders' Declaration*). However, poor impact assessment, ineffective accountability for its implementation in AI design, limited expertise in the protection of human rights in the context of AI, and lack of effective participation of potentially affected people may significantly hamper the positive results that these recommendations can bring.

Therefore, if governments decide to actively support a risk-based approach, they should (i) invest in training to create a global community of experts and stakeholders capable of addressing AI risks with a focus on human rights and the SDGs; (ii) promote the public availably of the results of impact assessments, or at least their accessibility to independent supervisory authorities; (iii) provide methodological guidelines and risk assessment models to adequately frame the protection of human rights and the achievement of the SDGs in the design and development of AI; (iv) extend this assessment to new AI technologies to investigates the potential and actual effects of such technologies, on the basis of their known or potential applications.

References

Council of Europe. 2024. *Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Strasbourg. https://rm.coe.int/1680afae3c.

Data & Society, European Center for Not-for-Profit Law (ECNL). 2021. *Mandating Human Rights Impact Assessment in the AI Act*. https://ecnl.org/sites/default/files/2021-11/HRIA%20paper%20ECNL%20and%20Data%20Society.pdf

Data & Society, European Center for Not-for-Profit Law (ECNL). 2021. *Recommendations for Assessing AI Impacts to Human Rights, Rule of Law and Democracy*. https://ecnl.org/sites/default/files/2021-11/HUDERIA%20paper%20ECNL%20and%20DataSociety.pdf.

G20. 2023. *G20 New Delhi Leaders' Declaration. New Delhi, India, 9-10 September 2023*. New Delhi. https://www.consilium.europa.eu/media/66739/g20-new-delhi-leaders-declaration.pdf.

G7. 2023. *Hiroshima Process International Guiding Principles for All AI Actors*. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system.

Mantelero, Alessandro. 2022. *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*. The Hague: T.M.C. Asser Press-Springer. https://doi.org/10.1007/978-94-6265-531-7 (open access).

The Danish Institute for Human Rights. 2020. *Guidance on HRIA of Digital Activities*. https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities.

United Nations - Human Rights Council. 2011. *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*, Resolution 17/4 of 16 June 2011.

United Nations, General Assembly. 2024. *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*, draft resolution, A/78/L.49, 11 March 2024.

United Nations. 2015. *Transforming Our World: the 2030 Sustainable Development Goals Agenda*. https://sdgs.un.org/2030agenda.

G20 BRASIL 2024
BUILDING A JUST WORLD AND A SUSTAINABLE PLANET

T20 Brasil 2024
Let's rethink the world

Let's rethink the world

CEBRI
Rethink Tank

FUNDAÇÃO ALEXANDRE DE GUSMÃO

ipea
Instituto de Pesquisa Econômica Aplicada