Human rights-based principles for AI governance and business

Pacharapan Roehrl

Abstract

The design and use of AI systems have raised concerns among governments, civil society, academics, technologists, investors, and businesses as to how best to maximize their benefits and protect against harms, given unique AI challenging dynamics and business models.

To address a broad spectrum of concerns, this policy brief draws upon lessons from the UN B-Tech Project and the OECD Working Parties (on Responsible Business Conduct and on AI Governance). Grounded in established international standards of business conduct, these initiatives adapt and apply the due diligence frameworks outlined in the UN Guiding Principles on Business and Human Rights and the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct to all actors in the AI value chain.

To further enhance transparency, public participation, and accountability in AI development and deployment, this policy brief complements these ongoing initiatives with a proposal for process-oriented rights, inspired by the UN Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters (Aarhus Convention). Lastly, it recommends incorporating rights elements across all massive education and training programs on AI.

Background

The revolution in deep neural networks (DNN) and most recently generic artificial intelligence (AI) has led to an avalanche of proposals and initiatives on global, regional and national AI governance, such as the first ever AI Act of the EU, regulatory requirements in the USA and China, as well as a plethora of action at the level of the UN, including a new UN Advisory Body on AI, discussions in the Internet Governance Forum, UNESCO's recommendation on the ethics of AI, pioneering work in the UN Technology Facilitation Mechanism on the wider impacts of rapid tech change, recommendations on AI governance and support to developing countries by the Secretary General's 10-Member-Group of High-level Representatives.

Most notably pathbreaking are the two UN General Assembly resolutions: the Third Committee resolution 78/213, entitled "*Promotion and protection of human rights in the context of digital technologies*" adopted by consensus on 22 December 2023; and the UN GA plenary resolution A/78/L.49, entitled "*Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*," adopted on 11 March 2024.

However, the UN B-Tech Project (UN, 2023b) pointed out that, "these initiatives have tended not to incorporate the due diligence expectations laid out by the international standards of business conduct: specifically, the UN Guiding Principles on Business and Human Rights [UNGPs] (UN, 2011) and the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct [OECD Guidelines] (OECD, 2023)." The above-mentioned resolutions refer to these principles, but do not endorse them nor propose adaptations for AI.

With appropriate adaptation, this policy brief proposes to build future AI governance by incorporating the following components:

- The guiding principles as contained in the UNGPs/OECD Guidelines along a typical AI value chain (OECD, 2023c), prioritizing the ten substantive rights proposed by B-Tech Project's Taxonomy report (UN, 2023c);
- The guiding principles on process-oriented rights as contained in the UNECE Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters (Aarhus Convention) (UNECE, 1998); and
- Rights-based principles throughout ongoing education and training programs, including programs for all actors in the AI value chain, workforce development, and the public.

By building on the existing internationally agreed accountability and remedy models with nuances appropriate for the AI context, we would be able to tackle the unique challenges and dynamics presented by AI.

An adapted framework for AI governance applying the UNGPs/OECD Guidelines

The UN Human Rights B-Tech Project[i] was launched in 2019 as a platform of a multistakeholder consultation

process. Its Foundational Paper (UN, 2023b) provided a value proposition for leveraging and building on existing initiatives, good practices, and expertise based on the well-accepted UNGPs, to the development, use, and governance of digital technologies.

The B-Tech Project has developed an additional Taxonomy Report (UN, 2023c) detailing ten categories of real-world "risk examples". These examples show the potential adverse impacts of AI on substantive rights protected under the Universal Declaration of Human Rights and other relevant human rights instruments, including:

- Freedom from Physical and Psychological Harm
- Right to Equality Before the Law and to Protection against Discrimination
- Right to Privacy
- Right to Own Property
- Freedom of Thought, Religion, Conscience and Opinion
- Freedom of Expression and Access to Information
- Right to Take Part in Public Affairs
- Right to Work and to Gain a Living
- Rights of the Child
- Rights to Culture, Art and Science

Corporate responsibility

The B-Tech Project further developed practical tools and strategies to assist technology companies, investor community, civil society, and stakeholders. A series of foundational papers (UN, 2020-2021) was published to provide a global standard of expected conduct for all technology companies wherever they operate, including their operations, products, services, and their business relationships. By following the UNGPs' "know and show" principle, a technology company is expected to have in place:

1) An explicit and coherent governance structure and policy commitment approved at the most senior level and aligned with internationally recognized human rights standards and the eight ILO core conventions throughout all of its operations, relationships, and value chain.
2) A human rights due diligence and impact assessment process, embedding in its risk management framework to identify, prevent, and mitigate those "risk examples" as priority. In this process, a technology company should pay particular attention to impacts on individuals with different risks, such as women, men, children, migrant workers and their families, indigenous peoples, minorities, and persons with disabilities, namely.
3) Accountability and remediation processes, including company-based grievance mechanisms, when actual harm has occurred due to their operation or contribution, including harm done to human rights defenders.

Two effective means through which a technology company meets its responsibility to respect human rights will be based on ongoing consultation with stakeholders and transparent reporting. A meaningful engagement should be undertaken at regular intervals with affected communities, users, and civil society and as early as possible prior to a new or anticipated activity, relationship, or major decisions or changes. Transparent reporting regarding its human rights practices should be made publicly available to ensure accountability and foster trust among users and stakeholders.

State duty

The UNGPs/OECD Guidelines and OECD Recommendation of the Council on Artificial Intelligence (OECD, 2023b) provide governance frameworks to which the state duty to protect human rights can be adapted in the digital realm. In this context, Beduschi and Ebert (2021) offered recommendations, which incorporated accountability for different actors in the value chain. The OECD report on Advancing Accountability in AI identified the typical value chain in digital technologies involving various actors—from developers and service providers to end-users and intermediaries (OECD, 2023c). The following key components are essential for states' obligations to respect, protect, and fulfil human rights:

1) State commitment to implement its international obligations using the UNGPs and OECD Guidelines to guide and develop accountability mechanisms.
2) Clarity of obligations of each actor with a "smart mix of measures – national and international, mandatory and voluntary" (Beduschi and Ebert, 2021) – tailored to different actors throughout the value chain with regards to measures on privacy, security, data handling, and user protection.
3) Strengthening policy coherence across the digital technology value chain to ensure that all actors are held to the same standards of compliance and enforcement, whether they are involved in development, deployment, or management of digital technologies.
4) Implementing monitoring and compliance Mechanisms, including audits, reporting requirements, and/or the establishment of an

independent oversight body to ensure that all actors adhere to the set standards. This includes effective remedy mechanisms through judicial, administrative, or other means when harm has occurred.

5) Providing ongoing education and training for all actors in the digital value chain to ensure they understand their responsibilities, upskilling workforce and providing support for the transformation of the world of work and of society, and empowering the general public, as pioneered in Finland.[ii]

Two effective means through which States meet their duty to protect human rights involving engagement with multiple stakeholders and international collaboration. States can gather insights for relevancy and more effective and balanced governance solutions by engaging with multiple stakeholders, including tech enterprises, consumers, academia, experts, civil society, national human rights institutions, and national contact points, to name a few. International cooperation can lead to more consistent enforcement, especially in cases of harm committed across borders.

AI challenges and dynamics

The guidance on accountability and remedy models, while established and relevant, faces unprecedented challenges when applied to AI business. AI differs fundamentally from traditional business models due to several unique characteristics.

AI amplifies the "winners-take-all" phenomenon seen in digital and platform economies—where a few dominant players, like large tech companies, potentially control vast market shares due to network effects and data accumulation. In AI, this is exacerbated as these companies have superior access to data, enhancing their AI algorithms further and creating a loop where the winner takes all. Such dominance raises significant concerns for human rights, particularly regarding privacy, non-discrimination, and freedom of expression. How can we ensure that these companies safeguard individual rights when market dynamics incline so heavily towards monopolization?

The potential harms from AI are evolving and not fully understood. Unlike traditional technologies, the impacts of AI can be unexpected and new harms will need to be identified over time. This makes it challenging to fully address and mitigate these harms through existing regulatory frameworks or through self-regulation initiatives alone.

AI promises cognitive labor automation, just as past technological advancements liberated humans from much physical labor (Roehrl, 2022). This shift could lead to massive changes in the labor market and societal structure, similar to the industrial revolution, raising concerns for human as well as labor rights. Based on the IMF AI Preparedness Index[iii] (IMF, 2023), IMF's Managing Director Georgieva (2024) pointed out that in most scenarios, "AI will likely worsen overall inequality" between and within countries.

The EU has begun regulating AI through its AI Act (EU, 2024), which includes establishing a specialized regulatory body (art.64) to continuously monitor AI development and deployment. The Act uses risk-based assessment (art.9) and a classification system (unacceptable, high, limited, or minimal) (art.5-6) as a model to determine the level of risk of an AI technology on the health, safety, and fundamental rights of a person. The Act provides the specialized body with extraterritorial mandates (art.2(1)) and enforcement mechanisms of administrative fines (art.99). Together with the EU Data Governance Act (2022a), Digital Services Act (2022b), and the General Data Protection Regulation (GDPR) (2016)[iv], these legislative efforts complement one another and support the future of EU's digital governance.

Though the AI Act is viewed as a bold initiative, experts, international bodies, and NGOs have cautioned the 'blanket' exemptions (military and national security AI) (art.2(3)), the lack of access to remedy by rights holders, resource allocation and expertise for effective assessment and enforcement, among others. The most important concern, however, centers on whether the AI Act is robust and agile enough to adapt to the rapid pace of AI innovation and that such regulatory framework risks becoming outdated quickly as technologies rapidly evolve and become more widespread as is the case with general-purpose AI.

Addressing key concerns with an adapted process-oriented framework

These dynamics and concerns could be addressed by incorporating process-oriented rights, inspired by the Aarhus Convention (AC). The AC model links government accountability and environmental protection by granting three fundamental process-oriented rights: access to information (articles 4 and 5), access to public participation (articles 6, 7 and 8), and access to justice (article 9) prior to any implementation of environmental decisions, programs, and policies. Right holders (individuals as well as civil society, article 2) may bring complaints when their

rights to access information and/or participate in public processes are affected or denied. A similar process-oriented approach can be applied to AI governance for the following reasons:

1) The AC model respects democratic values and grants public rights by ensuring access to information and public participation in AI governance and it backs up these rights with access to justice provisions.
2) The AC model links environmental protection with public rights, including the rights to information and public participation in complex issues such as genetically modified organisms (GMOs). Similarly, by providing a legal basis for public access to information and participation, governments are held to account to tackle multidimensional, unknown impacts brought by AI, for example, an increase in technology-enabled gender-based violence, the amplification of discriminatory racial and ethnic stereotypes, the supercharging of online disinformation campaigns or the creation of child sexual abuse materials (UN, 2023b).
3) The AC model provides an important mechanism for multistakeholder engagement, including members of the public, civil society, the private sector, and government. Optional grounds for refusing disclosure are applied in a restrictive way, taking into account public interest in disclosure.
4) The AC model acknowledges that we owe an obligation to future generations and establishes that sustainable development can be achieved only through the involvement of all stakeholders.

While the Aarhus Convention has proven that it is fit for purpose to deal with rapidly evolving biotech issues, but its scope is limited to environmental issues. A Convention focused on AI and large-scale digital tech initiatives that guarantees the same process-oriented rights appears the main missing building block in AI governance that would be able to deal with rapid and unanticipated developments in this area.

Policy recommendations / conclusions

To address a broad spectrum of concerns, I propose an initiative grounded in established international standards of business conduct and adapted to the unique characteristics of AI, which should comprise of:

1) Supporting the UN B-Tech Project and the work of the OECD Working Party on Responsible Business Conduct and the OECD Working Party on AI Governance to apply and adapt due diligence frameworks and remedy mechanisms based on the UNGPs and OECD Guidelines for the rapidly evolving AI context.
2) Applying the procedural rights and remedy mechanisms model of the Aarhus Convention to enhance transparency, public participation, and accountability in AI development and deployment (ideally leading to an Aarhus-style Convention on AI and large-scale digital tech initiatives).
3) Incorporating rights aspects to massive and continuous training and education at all levels, in order to manage the AI impacts on the job market and to foster a socially just and sustainable AI-driven world.

Acknowledgments

I am grateful to Prof. Christine Kaufmann of the University of Zurich for her encouragement, to my friends at Human Rights Now, and to the UN Secretariat for this opportunity to propose these ideas to the STI Forum.

References

Beduschi, Ana and Ebert, Isabel (2021), Working Paper on *The relevance of the Smart Mix of Measures for Artificial Intelligence - Assessing the Role of Regulation and the Need for Stronger Policy Coherence*, Geneva Academy, September 2021.

Coalition of NGOs (2024), "EU's AI Act fails to set gold standard for human rights", 3 April, 2024, https://www.amnesty.eu/wp-content/uploads/2024/04/EUs-AI-Act-fails-to-set-gold-standard-for-human-rights.pdf.

EU (2016), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation (GDPR)), OJ L 119, 4.5.2016.

EU (2022a), Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), OJ L 152, 3.6.2022.

EU (2022b), Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277, 27.10.2022

EU (2024), Artificial Intelligence Act, European Parliament 'Corrigendum', 16th April 2024.

Feingold, Spencer (2023), "The European Union's Artificial Intelligence Act, explained", World Economic Forum Blog, 30 June 2023.

Georgieva, Kristalina (2024), "AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity," IMF Blog, 14 January 2024.

Hoffmann, Mia (2023), "The EU AI Act: A Primer", CSET (Center for Security and Emerging Technology Blog, 26 September 2023.

IMF (2023), *Artificial Intelligence: What AI means for economies*, IMF F&D publication, December 2023.

OECD (2023a), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en.

OECD (2023b), Recommendation of the Council on Artificial Intelligence, adopted on 21 May 2019, amended on 7 November 2023.

OECD (2023c), "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI", *OECD Digital Economy Papers*, No. 349, OECD Publishing, Paris, https://doi.org/10.1787/2448f04b-en. OECD proposed a typical AI value chain to include, for example, 1) suppliers of AI knowledge and resources (content creators; data providers and data annotators; digital infrastructure providers; hardware manufacturers. 2) Actors in the AI lifecycle - companies, States, research institutions involved in planning & design of the system; collecting & processing of data; building & using the model; verifying & validating the model; deploying the system, regardless of the distribution channel (including the distribution of open-source software); and operating & monitoring the system. 3) Users/operators of the AI system - businesses, including financial institutions, investors, and businesses in the 'real' economy (e.g., manufacturing, purchases, and flows of goods and services); individuals or other actors using AI for personal use, commercial, or research activity; and States.

Roehrl, R. (2022), Conceptualizing future scenarios of artificial intelligence: from energy servants to AI servants, Science-Policy brief for the STI Forum 2022, https://sdgs.un.org/sites/default/files/2022-05/1.2.5-46-Roehrl-Concept%20AI%20scenarios.pdf

Schuett, Jonas (2023), "Risk Management in the Artificial Intelligence Act", *European Journal of Risk Regulation* (2023), DOI: https://doi.org/10.1017/err.2023.1.

UNECE (1998), Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters (Aarhus Convention), Treaty Series, vol. 2161, 1998, p. 447 and The Aarhus Convention, An Implementation Guide, second edition, 2014.

UN (2011), Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework, 16th June 2011.

UN (2012), The Corporate Responsibility to Respect Human Rights, An Interpretive Guide, 2012.

UN (2020-2021), B-Tech Project Foundational Papers:

- *Addressing Business Model Related Human Rights Risks,* July 2020;
- *An Introduction to the UN Guiding Principles in the Age of Technology,* September 2020;
- *Key Characteristics of Business Respect for Human Rights,* September 2020;
- *Identifying Human Rights Risks Related to End-Use*, September 2020;
- *Taking Action to Address Human Rights Risks Related to End-Use*, September 2020;
- *Access to remedy and the technology sector: basic concepts and principles*, January 2021;
- *Access to remedy and the technology sector: a "remedy ecosystem" approach*, January 2021;
- *Designing and implementing effective company-based grievance mechanisms*, January 2021;
- *Access to remedy and the technology sector: understanding the perspectives and needs of affected people and groups*, January 2021; and
- *Bridging Governance Gaps in the Age of Technology – Key Characteristics of the State Duty to Protect*, May 2021.

UN (2023a), Third Committee resolution 78/213 entitled "*Promotion and protection of human rights in the context of digital technologies,*" 22 December 2023, https://www.undocs.org/A/RES/78/213.

UN (2023b), B-Tech Project Foundational Paper: *Advancing Responsible Development and Deployment of Generative AI, The value proposition of the UN Guiding Principles on Business and Human Rights*, November 2023.

UN (2023c), B-Tech Project Supplemental Paper: *Taxonomy of Human Rights Risks Connected to Generative AI, Supplement to B-Tech's Foundational Paper on Advancing Responsible Development and Deployment of Generative AI*, November 2023.

UN (2023d), Open Letter from the United Nations High Commissioner for Human Rights to European Union institutions on the European Union Artificial Intelligence Act ("AI Act"), 8 November 2023.

UN (2024), Second Committee resolution A/78/L.49, entitled "*Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*," 11 March 2024, https://www.undocs.org/A/78/L.49.

[i] *The Business and Human Rights in Technology Project (B-Tech Project)* of the UN OHCHR applies the UN Guiding Principles on Business and Human Rights to digital technologies. The project's latest phase started in 2023 with a focus on generative AI. It collaborates with the OECD.AI Network of Experts under the purview of the OECD Working Party on Responsible Business Conduct and the OECD Working Party on AI Governance. The OECD.AI network develops practical recommendations for AI actors under an overarching due diligence framework by incorporating existing AI risk management frameworks, such as the OECD Due Diligence Guidance for Responsible Business Conduct, the NIST AI Risk Management Framework, the G7 Code of Conduct

for the Development of Advanced AI Systems, IEEE 7000 series, ISO 31000, and ISO/IEC 23894.

[ii] Elements of AI, a series of free online courses (and teaching kit) created by MinnaLearn and the University of Helsinki in Finland, https://www.elementsofai.com/

[iii] IMF has developed an AI Preparedness Index that measures readiness in areas such as digital infrastructure, human-capital and labor-market policies, innovation and economic integration, and regulation and ethics.

[iv] The DA Act sets up a new European Data Innovation Board; the GDPR sets up the European data protection supervisor (EDPS). The oversight bodies are responsible for monitoring the application of rules, facilitating the exchange of best practices, and investigating complaints.