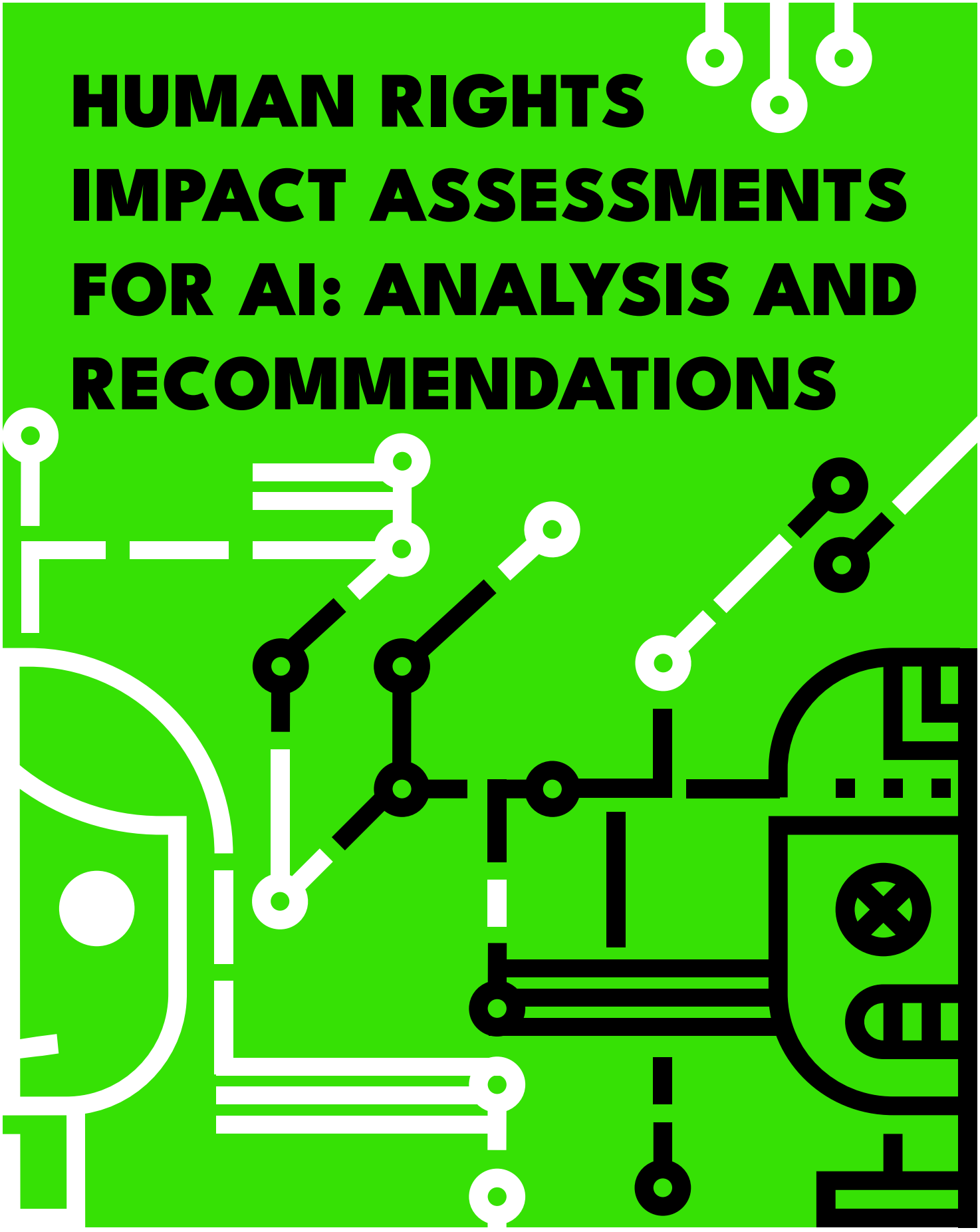
HUMAN RIGHTS IMPACT ASSESSMENTS FOR AI: ANALYSIS AND RECOMMENDATIONS



accessnow

Access Now (accessnow.org) defends and extends the digital rights of people and communities at risk. As a grassroots-to-global organization, we partner with local actors to bring a human rights agenda to the use, development, and governance of digital technologies, and to intervene where technologies adversely impact our human rights. By combining direct technical support, strategic advocacy, grassroots grantmaking, and convenings such as RightsCon, we fight for human rights in the digital age.

Human rights impact assessments for AI: analysis and recommendations

Brandie Nonnecke, Director, CITRIS Policy Lab, UC Berkeley
Philip Dawson, 2020-2021 Technology & Human Rights Fellow, Harvard Kennedy School Carr Center for Human Rights Policy

ABSTRACT

The public and private sectors are increasingly turning to the use of algorithmic or artificial intelligence impact assessments (AIAs) as a means to identify and mitigate harms from Artificial Intelligence (AI) Systems. While promising, lack of clarity on the proper scope, methodology, and best practices for AIAs could inadvertently perpetuate the harms they seek to mitigate, especially to human rights. We explore the emerging integration of the human rights legal framework into AI governance strategies, including within Canada's Directive on Automated Decision-Making and the European Commission's proposed Digital Services Act and Artificial Intelligence Act, as well as the implementation of human rights impact assessments (HRIAs) for AI Systems. The benefits and drawbacks from recent implementations of AIAs and HRIAs to assess AI systems adopted by the public and private sectors are explored and considered in the context of an emerging trend toward the development of standards and certifications for responsible AI governance practices. We conclude with priority recommendations for how the human rights framework can help better ensure AIAs and their corresponding responsible AI governance strategies live up to their promise.

October 2022

This report is a publication of Access Now, which commissioned this research as part of our work on the intersection of human rights law and Artificial Intelligence (AI) Systems. It is written by Brandie Nonnecke, Director, CITRIS Policy Lab, UC Berkeley and Philip Dawson, 2020-2021 Technology & Human Rights Fellow, Harvard Kennedy School Carr Center for Human Rights Policy. A previous version of the report was published in October 2021 as part of the Harvard Kennedy School Carr Center for Human Rights Policy Discussion Paper Series. It has been updated and expanded in this version.

Access Now would like to thank Brandie and Phil for their excellent and insightful work.

Access Now (https://www.accessnow.org) defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots grantmaking, legal interventions, and convenings such as RightsCon, we fight for human rights in the digital age.

INTRODUCTION

In response to growing recognition of the societal risks of artificial intelligence (AI) broadly and automated decision-making systems (ADS) in particular, algorithmic or AI impact assessments (AIAs) are increasingly being considered by the public and private sectors to identify, prevent, and mitigate harms, or as a means to improve the quality of AI products and services.[1] The term "algorithmic impact assessment" currently lacks definitional clarity. In general, AIAs aim to identify potential risks and impacts—including to health, safety, ethics and, in some implementations, to human rights —arising from the development and deployment of an algorithmic system as well as appropriate risk mitigation strategies, such as use of "algorithmic audits", "datasheets for datasets", and "model cards."[2]

Implementations of AIAs are gaining momentum as a viable AI governance strategy, finding their way into binding regulation and legislation.[3] Corporate policies are also requiring implementation of AIAs as a mechanism to reduce legal risks stemming from liability and negligence.[4] The European Commission's Artificial Intelligence Act suggests a risk-based approach to AI governance, prohibiting certain harmful applications of AI and calling for developers to go through a form of impact assessment (called a "conformity assessment") for high-risk applications to identify necessary oversight mechanisms.[5] The Algorithmic Accountability Act proposed in the United States Congress in 2019 would have required companies with large user bases to conduct impact assessments of highly sensitive ADS (the Act is expected to be reintroduced).[6] In 2021, the National Institute of Standards and Technology (NIST) was tasked by Congress to develop an "AI risk management framework" to guide the "reliability, robustness, and trustworthiness of AI systems" used in the federal government.[7] In 2021, the National Security Commission on Artificial Intelligence issued a report recommending that government agencies deploying AI systems conduct *ex ante* risk assessments and *ex post* impact assessments to "increase public transparency

[1] The term artificial intelligence (AI) is typically used to refer to a computer system capable of performing tasks that would ordinarily require some form of intelligence to accomplish, such as decision-making, visual perception, speech recognition, and more. Methods for doing so are wide ranging and vary significantly in complexity, such as algorithms, predictive models, computer vision, deep learning, machine learning, natural language processing, neural nets, and more. For more on the problematic nature of the term itself, see Daniel Leufer, "The term "AI" has a clear meaning," https://www.aimyths.org/the-term-ai-has-a-clear-meaning.

[2] Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking Explainable Machines: The GDPR's' Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Tech. LJ*, *34*, 143; Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for datasets." *arXiv preprint arXiv:1803.09010* (2018); Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting." In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229. 2019; Emanuel Moss et al. "Governing with algorithmic impact assessments: six observations." *Available at SSRN* (2020).

[3] Kent Walker and Jeff Dean, "An Update on Our work on AI and Responsible Innovation," *Google,* July 9, 2021, https://blog.google/technology/ai/update-work-ai-responsible-innovation.

[4] Andrew D. Selbst, Negligence and AI's Human Users, 100 B.U.L.REV. 1315 (2020);

[5] "Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)." European Commission, last modified April 21, 2021, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[6] Algorithmic Accountability Act, 116th Cong. (2019). https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf; Grace Dill, "Sen. Wyden to Reintroduce AI Bias Bill in Coming Months," *MeriTalk,* Feb. 19, 2021, https://www.meritalk.com/articles/sen-wyden-to-reintroduce-ai-bias-bill-in-coming-months/.

[7] "Commerce, Justice, Science And Related Agencies Appropriations Bill, 2021 - Report Together With Minority Views," House Committee on Appropriations, July 2020, https://appropriations.house.gov/sites/democrats.appropriations.house.gov/files/July%209th%20report%20for%20circulation_0.pdf, 23.

about AI use through improved reporting."[8] In this instance, risk assessments and impact assessments are differentiated, with risks being identified at the outset and impacts being evaluated after deployment to quantify and mitigate the identified risks. Canada's "Directive on Automated Decision-Making", which came into effect in 2020, led to the development of one of the first AIAs to identify and mitigate a range of risks—to individual rights, economic interests, health and well-being, and sustainability—arising from ADS developed and deployed in the public sector.[9]

While AIAs hold promise to promote the development of regulatory, policy, and governance mechanisms by government and corporate actors to identify potential harms, human rights organizations have warned that using risk-based AIAs may be inadequate.[10] Most guidance for implementation of AIAs indicates their use should be reserved for "high-risk" AI applications (e.g., use of AI in biometric identification or judicial sentencing). However, applications wrongly categorized as "low-risk" can thereby evade proper oversight. This would be especially problematic in a case where the onus of determining risk level is placed on the entity developing the AI. This could lead to a troubling scenario where developers artificially reduce the perception of risk in order to evade oversight. Further, a lack of common or internationally standardized approaches to the development of AIAs could lead to confusion and complicate their effectiveness.

As a risk-based approach increasingly dominates AI governance strategies, important questions emerge regarding the proper scope, methodology, and best practices that might protect AIAs from inadvertently becoming smokescreens for human rights and other abuses. In short, the ill-conceived development and deployment of AIAs pose substantial risk themselves. This is not to say that the implementation of AIAs cannot provide benefits now, but rather that significant work remains to determine how to appropriately develop and apply AIAs to ensure long-term effectiveness. If done inappropriately, their use may ultimately enable and perpetuate the harms they seek to mitigate.

We explore the emerging integration of the human rights framework into AI governance strategies, including the human rights implications of AIAs. We rely on the international human rights law framework, including the UN Declaration of Human Rights (UDHR) as well as the UN Guiding Principles on Business and Human Rights (UNGPs), to provide an analysis of emerging proposals for the use of AIAs, including in recommendations made by international and intergovernmental organizations, regulatory and legislative proposals from government bodies, and usage to date in the private sector. We conclude by analyzing the integration of human rights into emerging legislative proposals, such as in Canada's Directive on Automated Decision-Making and the European Union's Artificial Intelligence Act and Digital Services Act, and offer recommendations to help guide the effective development and use of human rights-based AIAs more broadly.

[8] Eric Schmidt et al. "National Security Commission on Artificial Intelligence (AI) Final Report". *National Security Commission on Artificial Intelligence*, (2021), 395, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.
[9] "Directive on Automated Decision-Making" Canadian Government, last modified April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592&section=html.
[10] Fanny Hidvegi, Daniel Leufer, and Estelle Massé, "The EU Should Regulate AI on the Basis of Rights, Not Risks," *Access Now,* Feb. 17, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach/.

AI GOVERNANCE AND HUMAN RIGHTS

Over the last few years, at least 170 sets of ethical or human-rights based AI principles, frameworks, and guidelines have been developed to support responsible AI development and deployment within the public and private sectors.[11] Research has shown that a growing consensus is forming around core principles, including the need for accountability, privacy and security, transparency and explainability, fairness and non-discrimination, professional responsibility, human control, and the promotion of human values.[12] As these AI principles gain acceptance within the public and private sectors, the focus is now shifting to the development of appropriate strategies to operationalize the principles into responsible practices. Yet this process is not straightforward.

Despite the relative convergence of AI principles proposals there is seldom consensus in the interpretation of the principles in practice, especially when it comes to the details.[13] AI principles have been developed by diverse institutions (e.g., academia, civil society, governments) with varying multistakeholder representation. Because these institutions have differing priorities and needs and have often applied different ethical frameworks (e.g., deontological, consequentialist, utilitarian approaches) to evaluate the benefits and risks of AI, there is great heterogeneity in how AI principles are defined and in recommendations for their appropriate operationalization. Certain scholars have argued that "AI principle proliferation" has perpetrated a crisis of legitimacy, complicating the already complex task of identifying and mitigating risks and harms of AI-enabled technologies.[14] In response, the international human rights framework and its normative and legal guidance has been proposed as a mechanism to support more consistent framing and operationalization of AI principles, and many prominent professional associations, consortia, intergovernmental organizations, governments, and companies seem to agree.[15]

The Institute of Electrical and Electronics Engineers (IEEE), the world's largest technical professional organization, issued a report in 2017 stating as its first principle that AI should be "created and operated to respect, promote, and protect internationally recognized human rights" and emphasized that human rights should be part of AI risk assessments.[16] The Asilomar Principles, with over 5,000 signatories from the public and private sectors, include the need to protect human rights in the design and deployment of AI systems.[17] The OECD AI Principles, which

[11] "AI Ethics Guidelines Global Inventory," Algorithm Watch, accessed June 20, 2021, https://inventory.algorithmwatch.org/.
[12] Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI," Berkman Klein Center for Internet & Society, Harvard University, 2020, https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf.
[13] "AI Ethics Guidelines Global Inventory," Algorithm Watch, accessed June 20, 2021, https://inventory.algorithmwatch.org/.
[14] Mark Latonero, "AI Principle Proliferation as a Crisis of Legitimacy." *Carr Center Discussion Paper Series,* (Sept. 30, 2020).
[15] Mark Latonero, "Governing Artificial Intelligence: Upholding Human Rights & Dignity." *Data & Society* (2018): 1-37; Alessandro Mantelero and Samantha Esposito. "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems." *Computer Law & Security Review* (2021); Eileen Donahoe and Megan MacDuffee Metzger. "Artificial Intelligence and Human Rights." *Journal of Democracy* 30, no. 2 (2019): 115-126.
[16] "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligence Systems," *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,* https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined
[17] "Asilomar AI Principles," *Future of Life Institute* (2017), https://futureoflife.org/ai-principles/

42 countries have pledged to uphold, specifically call for the protection of human rights.[18] The European Commission's High-Level Expert Group on AI has also established a set of principles, one of which calls for ensuring AI respects fundamental rights.[19] The White House Office of Science and Technology Policy (OSTP) in its National AI Initiative identified the need to ensure AI systems do not infringe upon human rights, especially rights to privacy, civil rights, and civil liberties.[20] Canada, through its Directive on Automated Decision-Making, is one of the first countries to develop an Algorithmic Impact Assessment tool that seeks to measure and mitigate the human rights harms of ADS used in public services.[21] In the private sector, technology companies like Salesforce have explicitly identified protecting human rights in their AI ethics strategy.[22] And Microsoft and Intel are among the first global tech companies to conduct HRIAs on their development and use of AI.[23]

Centering human rights within AI governance strategies can help operationalize AI principles across sectors, international contexts, and domain application areas.[24] Through codification in charters, case law, regulation and industry standards, human rights norms and values have gained broad global consensus.[25] The UDHR and corresponding international human rights instruments and guiding principles, UN treaties and commentaries, national laws, and related policies and guidelines have helped to clarify core definitions and interpretations of human rights over decades.[26] As such, international human rights norms and values may be "clearer, better defined, and [more] stable" than AI principles alone. Applying a human rights framework "facilitates better harmonization and reduces the risk of uncertainty" in defining and applying AI principles in practice.[27]

---

[18] "OECD/LEGAL/0449: Recommendation of the Council on Artificial Intelligence," OECD Legal Instruments, OECD, last modified May 21, 2019, https://legalinstruments.oecd.org/en/instruments/OECD%20-LEGAL-0449.

[19] "Ethics Guidelines for Trustworthy Artificial Intelligence," European Commission High-Level Expert Group on AI, Last modified April 8, 2019, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[20] "Advancing Trustworthy AI," *National AI Initiative Office,* (2021), https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/#Metrics-Assessment-Tools-and-Technical-Standards-for-AI

[21] "Directive on Automated Decision-Making" Canadian Government, last modified April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592&section=html

[22] "AI Ethics," *Salesforce,* accessed Aug. 22, 2021, https://einstein.ai/ethics.

[23] "Human Rights Annual Report," *Microsoft,* (2018), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE2FMZY; "Intel Human Rights Impact Assessment," *Article One Advisors,* (2018), https://www.articleoneadvisors.com/intel-hria;

[24] Alessandro Mantelero and Samantha Esposito. "An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems." *Computer Law & Security Review* (2021); Eileen Donahoe and Megan MacDuffee Metzger. "Artificial intelligence and human rights." *Journal of Democracy* 30, no. 2 (2019): 115-126; Charles Bradley, Richard Wingfield, and Megan Metzger. "National Artificial Intelligence Strategies and Human Rights: A Review. Second Edition." *Global Partners Digital and Stanford Global Digital Policy Incubator* (April 2021): 1-70.

[25] Eileen Donahoe and Megan MacDuffee Metzger. "Artificial intelligence and human rights." *Journal of Democracy* 30, no. 2 (2019): 115-126.

[26] "Universal Declaration of Human Rights," United Nations, (1948), https://www.un.org/en/about-us/universal-declaration-of-human-rights; "The Core International Human Rights Instruments and Their Monitoring Bodies," *United Nations Human Rights Office of the High Commissioner,* (2021), https://www.ohchr.org/en/professionalinterest/pages/coreinstruments.aspx; "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework," *United Nations Human Rights Office of the High Commissioner*, (2011), https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf; Mark Latonero. "Governing Artificial Intelligence: Upholding Human Rights & Dignity." *Data & Society* (2018): 1-37.

[27] Alessandro Mantelero and Samantha Esposito. "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems." *Computer Law & Security Review* (2021)

Take, for example, the principle of "non-discrimination," which exists in Article 2 of the UDHR and has also been widely adopted as an AI principle in the public and private sectors. The operationalization of "non-discrimination" is complicated due to the absence of a shared understanding of what it means in the development and deployment of AI systems. By applying a human rights framework and relevant charters, case law, and regulation to identify *how* "non-discrimination" has been interpreted in a particular domain, appropriate strategies to move the concept of "non-discrimination" from the abstract to the concrete can become clearer.

Human rights principles also highlight that the responsible design of AI systems, including transparency, explainability, and accountability, are not only desirable from a commercial or ethical standpoint, but prerequisites to upholding existing legal obligations. For instance, a lack of transparency regarding the use of AI systems can make it difficult to determine whether a violation of human rights or any other legal obligation has occurred, undermining the ability to seek redress. Similarly, and especially in the public sector, the reliance on a recommendation, decision, or insight provided by an AI system that is not explainable or accountable is at odds with human rights principles incorporated into national administrative law, which generally requires that an individual be provided with reasons for a decision made against them, as well as an opportunity to contest that decision and receive remedy(ies).[28]

The human rights framework can provide the substantive foundation and governance architecture needed to produce greater specificity in defining and operationalizing AI principles. As the public and private sectors increase their efforts to implement AIAs, calls to require Human Rights Impact Assessments (HRIAs) for AI are also on the rise.[29] Legislative approaches, such as the EU AI Act, are beginning to codify human rights-based principles in governance and oversight practices.[30] While promising, it is important to consider *how* AIAs and HRIAS should be implemented to better identify and mitigate risks. We next evaluate the design and scope of AIAs and HRIAs for AI and then turn to a discussion of the challenges associated with their implementation.

AlGORITHMIC IMPACT ASSESSMENTS AND HUMAN RIGHTS IMPACT ASSESSMENTS

There is a long history of using impact assessments in a variety of domains, including to assess and mitigate harms to the environment, data security, privacy, and human rights. For each, the appropriate scoping and implementation methods must be carefully negotiated and constructed

[28] "Human Rights and Technology Final Report (2021)," *Australia Human RIghts Commission,* 2021, https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-final-report-2021
[29] "Human Rights and Technology Final Report 2021," Australian Human Rights Commission, last modified May 27, 2021, https://tech.humanrights.gov.au/downloads?_ga=2.118186762.280577679.1628185379-267672886.1624901321; "Human Rights, Democracy and Rule of Law Impact Assessment of AI systems", Council of Europe Ad Hoc Committee on Artificial Intelligence Policy Develop Group (CAHAI-PDG), last modified May 21, 2021, https://rm.coe.int/cahai-pdg-2021-05-2768-0229-3507-v-1/1680a291a3
[30] "Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)." European Commission, last modified April 21, 2021, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

to support accountability.[31] In a recent study of impact assessments in different sectors, researchers noted that the methodology is largely driven by 10 constitutive components, including criteria such as source(s) of legitimacy (e.g., legislative or regulatory mandates that define who must implement an impact assessment and when), identifying potential "impacts" to be assessed and mitigated (e.g., risks to non-discrimination), and the appropriate methods for doing so (e.g., consultation with diverse subject matter experts and those directly affected).[32]

The design and implementation of impact assessments in the field of AI is nascent. As such, there is a lack of consensus or common standards regarding the appropriate configuration or application of such constitutive components, including which entities should administer, enforce and oversee AIAs or HRIAs to support legitimacy; how to adopt meaningful governance and engagement processes to support accountability; and what are the appropriate methods for implementation, including how to effectively define, identify, and mitigate risks.[33]

Metcalf et al. (2021) define AIAs as "emerging governance practices for delineating accountability, rendering visible the harms caused by algorithmic systems, and ensuring practical steps are taken to ameliorate those harms."[34] Typical sources of risk to be identified include the presence of bias in datasets used to train an AI system, as well as the fairness and explainability of the model; identification of potential impacts can include contextual considerations related to equity and justice, as well as the economic interests, health, and well-being of users or populations potentially affected by the proposed system. Companies may integrate AIAs in whole or in part into traditional product design, reviews, risk management, and due diligence processes. Further, implementation of AIAs early, perhaps during the initial design process, is likely more effective at identifying and mitigating risks before widespread investment and deployment. AIAs should also consider leading to the termination of an AI application if appropriate safeguards cannot be put in place. Like "privacy by design" concepts, AIAs could support "AI responsibility by design."

The goal of an AIA, as with other impact assessments, is ultimately to identify technical adjustments that can be made to the AI system in order to eliminate the risks identified or to reduce them to an acceptable level. Because of their deep expertise and knowledge of the AI system being assessed, technology firms will likely be the primary administrators of AIAs, creating a potential situation where these firms have an outsized effect on what is included in AIAs and how they are implemented in practice.[35] Thus, transparency in how AIAs are developed and implemented by technology firms is critical.

---

[31] Jacob Metcalf et al.. "Algorithmic impact assessments and accountability: The co-construction of impacts." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 735-746. 2021.
[32] Emanuel Moss et al., "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest," *Data & Society,* June 29, 2021, https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/
[33] *Ibid*., 28.
[34] *Ibid*., 26.
[35] Andrew Selbst. "An Institutional View of Algorithmic Impact Assessments." *Harvard Journal of Law and Technology, 35*(10), 78, (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867634.

An HRIA is "a tool to evaluate the potential or actual impact of an organization's strategy, practice, or product on people's human rights."[36] Endorsed by the UN Human Rights Council in 2011, the UNGPs underpin much of the criteria and guidance applicable to best practices of HRIAs. The UNGPs recommend that assessments of human rights impacts should be undertaken regularly and at appropriate stages of a business's operations as part of its human rights due diligence processes, for instance, prior to a new activity or relationship; major decisions or changes in its operations (e.g., market entry, product launch, policy change, or wider changes to the business); and periodically throughout the life of an activity or relationship. In general, the assessment should include identifying who may be affected; cataloging the relevant human rights standards and issues; projecting how the proposed activity and associated business relationships could have adverse human rights impacts on those identified; and identifying mitigations that might eliminate or reduce the level of risk to an acceptable level.

Large technology companies like Microsoft and Facebook have begun conducting HRIAs to identify and address technology-related human rights risks, including those emanating from AI.[37] Microsoft publishes a "Human Rights Annual Report" within which the human rights effects of its technologies are explored and certain risk mitigation strategies are discussed. However, the company has no obligation to publish reports outlining the details of any mitigation actions it undertakes, or how feedback from civil society organizations has been addressed. Facebook commissioned an HRIA to evaluate its role in the genocide of the Rohingya in Myanmar. Yet scholars have criticized the HRIA for failing to uncover the most salient human rights harms of Facebook's AI-enabled tools and appropriate mechanisms to mitigate those harms moving forward.[38]

In the remainder of this section, we explore proposed and existing AIAs and related impact assessment strategies, such as conformity assessments and data protection impact assessments, in the public and private sectors to better understand emerging trends in their scope and structure and the corresponding benefits and risks associated with their implementation, especially to human rights. We first review Canada's "Directive on Automated Decision-Making" and its development and use of AIAs to evaluate and mitigate harms of ADS in government public service delivery. We next consider the EU's AI governance strategy through an evaluation of the EU AI Act and its risk-based approach to AI governance, including proposed implementation of "conformity assessments" to identify and mitigate AI risks emerging from the private sector. We then explore the relationship between the EU AI Act and the EU's General Data Protection Regulation (GDPR) and the feasibility of GDPR-mandated data protection impact assessments (DPIAs) to evaluate and mitigate human rights-based risks of AI. We then evaluate the EU's Digital Services Act (DSA) and its oversight mechanisms to mitigate risks of very large, AI-driven online platforms for human rights.

---

[36] Mark Latonero and Aaina Agarwal. "Human rights impact assessments for AI: Learning from Facebook's failure in Myanmar." *Carr Center Discussion Paper Series, March 19 (2021)*

[37] "Microsoft Global Human Rights Statement." Microsoft Corporation, last modified December 11, 2020, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4JIiU; "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar." Facebook, last modified November 5, 2018, https://about.fb.com/news/2018/11/myanmar-hria/

[38] Mark Latonero and Aaina Agarwal. "Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar." *Carr Center Discussion Paper Series,* March 19, 2021, https://carrcenter.hks.harvard.edu/publications/human-rights-impact-assessments-ai-learning-facebook%E2%80%99s-failure-myanmar.

We conclude with a discussion of the implementation of HRIAs for AI and how these may differ from, complement, or should be integrated into AIAs and other impact assessment strategies to better ensure the protection of fundamental human rights.

Canada's Directive on Automated Decision-Making

In 2019, the Canadian government released its Directive on Automated Decision-Making (the Directive).[39] The Directive's principal objectives were to ensure the incorporation of ADS into external public service delivery respects "core administrative law principles such as transparency, accountability, legality, and procedural fairness" and to ensure harmful effects of algorithms on administrative decisions are assessed and reduced.[40] To this end, the Directive includes an accompanying impact assessment tool in the form of a questionnaire that must be completed prior to the development of any ADS. Completion of the questionnaire helps internal teams compute a raw impact score that measures the risk of the automation, for instance, to the rights of individuals or communities, their health, well-being or economic interests, as well as effects on the overall "sustainability of the ecosystem."[41] Depending on the level of impact, the Directive provides for increasingly rigorous mitigation requirements, such as extensive peer review, notice, human intervention in the decision-making process, the provision of a "meaningful explanation", or personnel training.

While the Directive received attention both within Canada and globally, the government has been criticized for failing to enforce its requirements. Since the Directive came into force in May 2020, few AIAs have been completed and published per its requirements.[42] In a sense, Canada's experience with the Directive highlights a challenge that is well-known to global technology companies—obtaining institutional support and deploying the resources and expertise necessary to support the implementation of organization-wide compliance tools is not a straightforward process, particularly for emerging and poorly understood technologies such as ADS and AI.

EU AI Governance Strategy

In April 2021, the European Commission released its draft "Artificial Intelligence Act" (AI Act).[43] While the AI Act is the most comprehensive approach to AI governance proposed in the EU, it is part of a larger EU AI governance strategy, including the GDPR's mandated evaluation and oversight of automated decision systems and the DSA's proposed oversight of algorithmic-driven very large online platforms. We first provide a high-level overview of the AI Act, then discuss its

[39]"Directive on Automated Decision-Making" Canadian Government, April 1, 2021, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592&section=html
[40] *Ibid*., 35
[41] *Ibid*., 35
[42]Tom Cardoso and Bill Curry, "National defense skirted federal rules in using artificial intelligence, privacy commissioner says", The Globe and Mail, last modified February 8, 2021, https://www.theglobeandmail.com/canada/article-national-defence-skirted-federal-rules-in-using-artificial/; "Open Government Portal," Government of Canada, accessed Sept. 7, 2021, https://search.open.canada.ca/en/od/?search_text=AIA
[43]"Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)." European Commission, last modified April 21, 2021, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

connection with oversight requirements of automated decision systems instituted through the GDPR and proposed in the DSA.

The draft AI Act takes a risk-based approach to AI regulation, establishing three primary levels of risk: low/minimal, high, and unacceptable.. The Act proposes different levels of oversight for low/minimal and high-risk AI applications. Additionally, applications posing a risk of manipulation, such as AI applications that manipulate images, audio, video content (Title IV), would have transparency obligations. Applications that fall within the category of unacceptable risk are prohibited (e.g., uses of AI that are capable of manipulating individuals through subliminal techniques). High-risk applications, such as use of AI in critical infrastructure, medical devices, and education or which pose a risk to health, safety, and/or fundamental rights, such as credit scoring or hiring decisions, must undergo pre-market conformity assessments attesting to their compliance with the Act. Providers of high-risk AI systems must have a post-market monitoring system in place, in which they actively collect, document, and analyze relevant data throughout the AI system's lifetime. The development and use of harmonized technical standards, such as those in relation to bias mitigation, risk or quality management, is encouraged to facilitate the implementation of conformity assessments.

Other than certain remote biometric identification systems, however, which must be assessed by independent third parties ("notified bodies"), providers of most high-risk AI systems can attest to their conformity with requirements of the Act through a self-assessment. Moreover, Article 43 of the Act indicates that high-risk systems that are in conformity with "harmonized standards" shall be presumed to be in compliance with the Act, potentially enabling a self-assessment regime for all high-risk systems. Commentators have expressed concern that barriers to participation by human rights experts and civil society organizations in the development of technical standards could lead to the recognition of "harmonized standards" that do not address the Act's human rights objectives. [44]

To strengthen accountability for the protection of human rights, Article 9 of the EU AI Act could be revised to make the assessment of AI systems' human rights risks an explicit feature of high-risk providers' risk management systems. This would help incentivize the harmonized development of human rights-based approaches to risk management standards and conformity assessments both within and outside the EU. Furthermore, lawmakers may consider requiring providers of high-risk systems to submit to a conformity assessment conducted by an independent third party in certain cases, for instance, in the event of non-compliance with post-market monitoring requirements, or where serious incidents are reported (in addition to a potential requirement to remove the system from the market until such a conformity assessment or re-evaluation has been completed). Finally, while the Commission has chosen to designate a defined list of AI systems as high-risk *a priori*, it should consider developing and publishing a clear, objective methodology for assigning a level of risk to new AI systems, or for re-evaluating the initial risk designation for existing systems in light of new evidence. As per recommendations by civil society organizations,[45] policy makers should also

[44] Michael Veale and Frederik Zuiderveen Borgesius. "Demystifying the Draft EU Artificial Intelligence Act." SocArXiv, 6 July 2021
[45] Access Now, European Digital Rights (EDRi), Panoptykon Foundation, epicenter.works, AlgorithmWatch, European Disability Forum (EDF), Bits of Freedom, Fair Trials, ANEC (European consumer voice in standardisation) Platform for Undocumented Migrants, "An EU

consider allowing for the list of prohibited practices (Title II) and the list of practices in Title IV to be updated, and provide clear criteria to guide such decisions about risk designation. Such criteria could be based, for instance, on the OECD's ongoing work to develop a framework for classification and risk identification.[46] The methodology should take into account the criteria listed by Mantelero and Esposito for HRIAs for AI,[47] and include guidance to clarify situations in which the deployment of an AI system poses unacceptable risks to society and should not be allowed to proceed.

In light of the need for harmonized techniques, proposals to use the GDPR's Data Protection Impact Assessments (DPIAs) have been offered as an effective mechanism to fulfill the risk and impact assessment obligations put forth in the EU AI Act, especially for high-risk applications involving personal data.[48] Article 35 of the GDPR states that DPIAs are required where a type of data processing "is likely to result in a high risk to the rights and freedoms of natural persons."[49] DPIAs are especially required if an entity is implementing a new technology or if the data processing is used to make automated decisions.[50] This supports Article 22 of the GDPR, which affords data subjects the right to *not* be subject to a decision based solely on automated processing.[51] Providing data subjects with insights from a DPIA can help to inform their decision for whether they approve of the implementation of automated decision-making.

While DPIAs may be effective at helping to identify and mitigate risks of AI, we believe there are three primary limitations that must be considered. First, DPIAs are primarily flagged for applications that use personal data. While these applications are often high-risk, this scoping may inadvertently overlook applications that pose significant human rights risks without the use of personal data. Additional high-risk applications, such as use of AI in control systems for critical infrastructure, may not be flagged for more strenuous oversight and evaluation even though their failure poses catastrophic human rights implications. Second, DPIAs are framed to the individual rather than groups, including broader societal or environmental risks. DPIAs are primarily focused on evaluating individual rights (e.g., to privacy, human agency) and as such may not adequately cover the identification and mitigation of risks of AI for collective rights (e.g., cultural identity). Third, while there are overarching guidelines for what should be included and assessed in DPIAs,

---

Artificial Intelligence Act for Fundamental Rights - A Civil Society Statement," https://www.accessnow.org/cms/assets/uploads/2021/11/joint-statement-EU-AIA.pdf. For specifc details on the proposals to update the risk categories, see this issue paper on future-proofing the risk-based approach" https://accessnow.org/AIAct-risk-approach

[46] OECD, Network of AI Experts, Classifying and Risk, (2021) https://oecd.ai/en/network-of-experts/working-group/1137

[47] Alessandro Mantelero and Samantha Esposito. "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems." *Computer Law & Security Review* (2021)

[48] Access Now. "Here's how to fix the EU's Artificial Intelligence Act." Sept. 7 , 2021, https://www.accessnow.org/how-to-fix-eu-artificial-intelligence-act/; European Data Protection Board (EDPB) and European Data Protection Supervisor (EDPS). "Joint Opinion on the proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), May 2021, https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf. United Kingdom Information Commissioner's Office (ICO) "The Information Commissioner's response to the European Commission's white paper on artificial intelligence - a European approach to excellence and trust. June 10, 2020, https://ico.org.uk/media/about-the-ico/consultation-responses/2617826/ico-response-to-eu-commission-white-paper-on-ai.pdf.

[49] European Union General Data Protection Regulation. "Article 35. Data protection impact assessment." https://gdpr.eu/article-35-impact-assessment/.

[50] European Data Protection Board (EDPB) and European Data Protection Supervisor (EDPS). "Joint Opinion on the proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), May 2021, https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf.

[51] European Union General Data Protection Regulation. "Article 22. Automed individual decision-making, including profiling." https://gdpr.eu/article-22-automated-individual-decision-making/

there is great variability in how these guidelines are implemented in practice across the EU.[52] Further, DPIAs are not strictly mandated to be made publicly available. As such, this poses significant risk to the realization of transparency and accountability in AI governance. The EU AI Act could complement the GDPR by providing measures addressing these points.

Another effort underway to provide greater transparency and oversight over AI-driven systems is the EU Digital Services Act (DSA). The DSA represents a significant development in ongoing efforts to identify and address the human rights impacts of very large online platforms (VLOPs), as well as the algorithmic systems they employ. In particular, the DSA could impose an obligation on VLOPs (currently defined as having 45 million active monthly users in the EU, but still under negotiation) to conduct annual risk assessments specific to their services, which appears to combine aspects of AIAs and HRIAs. Specifically, Article 26 of the DSA provides that VLOPs must analyze and assess "systemic risks" to fundamental rights enshrined in the EU Charter, taking into account the impact of content moderation and recommender systems on such rights.[53] In this way, platforms are required to evaluate the impacts of these algorithmic systems on the rights enumerated. VLOPs must also submit to independent audits assessing and reporting on their compliance with obligations under the DSA, including the requirement to conduct annual risk assessments. The platforms must report on the outcome of these risk assessments, mitigations, as well as the implementation of mitigations and audit recommendations every six months. As compared to DPIAs, the public reporting requirements of risk assessments proposed in the DSA are a promising step forward.

Several improvements to certain provisions of the DSA could help close potential gaps in oversight and mitigation of human rights harms. First, Article 26 of the DSA could be amended to clarify that the requirement to conduct annual risk assessments flows from platforms' more general obligation to institute organization-wide human rights due diligence processes, including conducting risk assessments regularly and at critical stages of the AI lifecycle, (as informed by the UNGPs and Latonero and Agarwal's emerging scholarship on HRIAs for AI).[54] This could help avoid some of the pitfalls and misconceptions associated with HRIAs, as noted above, including their potential to be manipulated through decisions with respect to timing and scope, and promote the implementation of ongoing human rights-based risk management processes by platforms.

Second, given their complexity, scale, and scope of impact, risks assessments performed by VLOPs should follow Mantelero and Esposito's recommendation for HRIAs conducted in complex multi-factor scenarios (e.g., as in the case of Sidewalk Labs) and include consultations with independent human rights experts, civil society, and stakeholder groups.[55] This would add a higher level of accountability and legitimacy to the risk assessments as well as to the mitigations

[52] United Kingdom Information Commissioner's Office. "Data protection impact assessments." Last accessed Oct. 21, 2021, 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/

[53] European Commission. "EU Charter of Fundamental Rights" (2012). https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en

[54] Mark Latonero and Aaina Agarwal. "Human rights impact assessments for AI: Learning from Facebook's failure in Myanmar." *Carr Center Discussion Paper Series, March 19 (2021)*

[55] Alessandro Mantelero and Samantha Esposito. "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems." Computer Law & Security Review (2021)

proposed. Similarly, Article 28 could be amended to include a requirement for organizations performing the independent audits of platforms' obligations to have specific expertise in human rights, given its identification as a "systemic risk" (currently, this provision requires "proven expertise in the area of risk management, technical competence and capabilities" as well as "professional ethics").

Finally, given the global scale of platforms and the likely impact of the DSA outside of the EU, the supervisory authorities that will be tasked with the enforcement of the DSA should consider using the UDHR and the UNGPs as a basis for developing guidelines for content moderation-specific HRIAs and human rights due diligence. This would help safeguard against the problem of scoping risk assessments too narrowly, and help promote greater consideration of collective and societal issues, such as public health / mental health, the environment, and electoral integrity.

Towards a Systemic Approach

If AIAs are to be relied upon to protect society from potential AI harms, inclusion of risks to fundamental human rights will be critical to their success. Generally, the object of AIAs consists of the algorithmic or AI system(s), including the datasets used to train these systems. One of the current trends associated with AIAs is to focus on assessing the sociotechnical aspects (e.g., the potential for bias, fairness, or explainability of the system) and their immediately foreseeable and measurable risks or consequences. In doing so, Metcalf et al. (2021) caution that AIAs may lead to an "ontological flattening" of the risks of AI-driven systems.[56] Approaching AIAs in this manner may inadvertently lead to overlooking human rights risks altogether, or a failure to identify the connection between technical weaknesses and downstream, context-dependent impacts, including to human rights—especially those that occur secondarily (e.g., the chilling effect of misidentification by facial recognition systems on an individual's freedom of assembly and expression or the tendency of misinformation to amplify online misogyny and radicalization). In this sense, the range of issues to consider in the context of AIAs can be far more extensive than for traditional product reviews. As such, scoping AIAs too narrowly can lead to a false sense of due diligence in risk identification and mitigation, allowing tools with non-trivial risks to human rights to operate freely.

Defining the scope of an HRIA also presents specific challenges. Because the focus of the exercise shifts from an assessment of the quantifiable technical risks of an AI system to the potential for real-life impacts on the rights and freedoms of individuals and communities, the scope of HRIAs tend to be broader and more forward-looking than that of AIAs by default. Accordingly, while the subject of an HRIA could be the AI system itself, the assessment is more likely to require consideration of risk and impacts at a higher level, for example, resulting from the deployment of the product in different contexts; the nature of the overarching business or public activity; the presence of adequate legal protections or governance structures, including whether there is a history of human rights abuses where the AI is to be deployed; the track record of supply chain

[56] Jacob Metcalf et al. "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 735-746. 2021.

partners; or all of the above. Furthermore, HRIA guidance cautions against preemptively narrowing the scope of human rights and freedoms to be investigated at the outset of an assessment, for instance, to consider only risks or impacts related to the right to privacy or equality and non-discrimination.

In addition to the need to design appropriate methodologies for conducting AIAs and HRIAs for AI in different contexts, their operationalization also raises important considerations, for instance, in light of the administrative burden and costs involved. One approach taken by companies is to set up a central unit that develops internal policies and procedures for AI governance, which may incorporate components of AIAs and/or HRIAs. This requires hiring additional personnel with appropriate socio-technical expertise, consequently increasing operating costs. Even with a central "responsible AI" unit in place, additional hurdles arise with respect to training different teams to identify and mitigate potential AI risks, in particular on account of the distinct skill sets, roles, and responsibilities of personnel at various stages of the AI lifecycle (e.g., design, development, or deployment). Companies may opt to conduct training one multidisciplinary workshop at a time and struggle to administer AI governance at the enterprise level. Scalability challenges may be further compounded by the potential for AI systems to exhibit different risks depending on the context of deployment, and the global scale at which systems may operate. Alternatively, another approach taken by companies, especially small- to medium-size enterprises that may not have the financial backing to develop a standalone "responsible AI" unit, may be to hire external consultants to help adapt existing policies and procedures to the AI context; upskill employees; or to prioritize conducting in-depth, standalone AIAs and/or HRIAs for applications believed to be higher risk.

In the absence of proper guidance, the timing of impact assessments can also have significant effects on their outcomes and credibility. For example, a recent study of the HRIA commissioned by Facebook regarding its potential implication in the genocide in Myanmar cautioned against the use of HRIAs as one-time, *ex post* exercises, which could become a form of AI "ethics washing."[57] Rather, and as instructed by the UNGPs, HRIAs should be conducted at appropriate intervals, aligned with critical stages of the AI lifecycle and as part of ongoing risk management processes such as human rights due diligence.[58] In addition, the study concluded that HRIAs should be conducted at the earliest stages of the design or conception of AI systems.

The *ex ante* HRIA conducted on Alphabet-affiliate Sidewalk Labs' "smart-city" project in the City of Toronto represents one potential example of this approach. More than 50 proposed digital solutions, including some anticipated to leverage the use of AI, were assessed prior to the confirmation of the project. The project was ultimately abandoned and some experts involved in the consultation pointed to human rights flaws of the proposed plans. [59]While the final report of this HRIA was never publicly released, the exercise, which included extensive consultation with subject matter experts and local stakeholders, contributed to the acceleration and enhancement of

[57] *Ibid*., 41.
[58] Mark Latonero and Aaina Agarwal. "Human rights impact assessments for AI: Learning from Facebook's failure in Myanmar." *Carr Center Discussion Paper Series, March 19 (2021)*
[59] Gabrielle Canon. "'City of surveillance': privacy expert quits Toronto's smart-city project." *The Guardian, October 2018.* https://www.theguardian.com/world/2018/oct/23/toronto-smart-city-surveillance-ann-cavoukian-resigns-privacy

existing human rights-based governance efforts related to the project.[60] However, as Mantelero and Esposito (2021) point out, while labor-intensive HRIAs that involve extensive research and field work, including consultations with local stakeholders and subject matter experts, may be desirable in complex multi-factor scenarios (e.g., large smart-city projects), they are likely too burdensome and costly to serve as appropriate models for projects of a smaller scale.[61] Consideration should be given to developing light touch HRIAs, with methodologies calibrated to the nature of the context, risk profile, and/or stage of the AI lifecycle.

In light of the dynamic nature of AI systems, which can evolve, drift, or adapt in unpredictable ways, reliance on static governance tools, such as AIAs, may capture only a snapshot of an AI system's operations upfront and be ineffective at identifying potential downstream risks and necessary mitigations. Rather, continuous monitoring and auditing of deployed systems by regulatory authorities may require the development of technologies that can help automate verification of compliance and complement human oversight.[62] Given that AI's technical capabilities are progressing at a pace that greatly outstrips the ability to govern their harms through primarily manual risk management processes, the adaptation of policy frameworks and increased investment by both the public and private sectors could help incentivize the development of technologies that can help implement AI governance at scale more effectively.[63]

Ultimately, design specifications and implementation tactics for AIAs and HRIAs will have to be tailored to the complexity, scale, context, and scope of the projects they are intended to assess, including their phase of development. Without sector-specific guidance, standards, or training of qualified personnel, the operationalization of AIAs and HRIAs is likely to face significant hurdles and be inadequate to address specific impacts on rights. In this context, poor outcomes associated with conducting AIAs or HRIAs for AI, whether due to their administrative burden or failure to identify and mitigate risks, should be expected to have negative feedback effects on their legitimacy. At least part of the solution to this problem could reside with standards bodies, such as the IEEE, International Organization for Standardization (ISO), NIST and national counterparts, which are beginning to develop standards and conformity assessments to guide the responsible development and deployment of AIAs and related risk management processes. These soft law tools may have significant effects on human rights due diligence in the context of AI, providing enterprise-level guidance regarding best practices and clarifying expectations for accountability.

---

[60] *Corporation of the Canadian Civil Liberties Association and Lester Brow*n v. *Toronto Waterfront Revitalization Corporation, et al.*, (Ontario Superior Court of Justice File No. 211/19), Affidavit of Kristina Verner, January 17, 2020, https://ccla.org/wp-content/uploads/2021/06/Affidavit-of-Kristina-Verner_TSC.pdf;

[61] Alessandro Mantelero and Samantha Esposito. "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems." *Computer Law & Security Review* (2021).

[62] Gillian Hadfield, Rules for a Flat World, Oxford University Press, May 14 (2020); Jack Clark & Gillian K. Hadfield, 2019. "Regulatory Markets for AI Safety," Papers 2001.00078, arXiv.org

[63] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault, "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

THE ROLE OF STANDARDS AND CERTIFICATIONS

In parallel with the development of AI principles and the exploration of regulations, standard development organizations (SDOs), at both the national and international levels, have been actively working on developing AI standards and conformity assessments. The standards may provide helpful guidance on creating and implementing effective AIAs by offering definitional clarity on how to operationalize responsible AI principles in practice. Conformity assessments will be used to verify that a company's product, service, management/governance process meets the normative and/or technical requirements contained in those standards. As an additional step, certification schemes are being developed to enable accredited third-party assessors to certify conformity with AI standards by issuing a certification "mark" or "label." However, caution must be taken to ensure certifications are not confusing or deceptive, leading to a sense of "false trust" in AI products and services, as has been witnessed in other industries.[64]

As these processes mature, it is likely that certain AI-related industry standards and conformity assessments will be incorporated into legislation or regulation as a condition of compliance. With diverging approaches to AI regulation being proposed, in Europe and elsewhere, the international harmonization and mutual recognition of AI standards and conformity assessments will emerge as significant geopolitical issues, which is critical to the protection against AI harms but also to the international trade of AI goods and services.

In recognition of the global importance of AI standards, the IEEE has demonstrated a commitment to the development of a human rights-driven approach. Its report outlines a conceptual framework for addressing universal human values, data agency, and technical dependability through a set of principles to guide developers and users engaged in the design, development, and deployment of AI systems. Human rights are identified as the first General Principle, with explicit reference to the international human rights framework and the relevance of the UNGPs. Additionally, the IEEE is developing an Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). The ECPAIS is currently developing a set of standards focused on bias, transparency, and accountability. If a developer implements the ECPAIS standards, it can add a quality assurance mark to its products and services with the intent to raise consumer trust and market power.[65]

The ISO and the International Electrotechnical Commission (IEC) are advancing a conformity assessment standard for AI risk management through the work of a joint committee on artificial intelligence (ISO/IEC JTC1/SC 42).[66] The proposed ISO/EC 42001 - Artificial Intelligence Management System (AIMS) standard will enable organizations to show they have implemented and continually work on improving processes to address bias, fairness, inclusiveness, safety, security, privacy, accountability, applicability, and transparency in AI.

---

[64] Lesley Fair, "Deceptive 'Certified Organic' claims leave consumers verklempt," *Federal Trade Commission,* https://www.ftc.gov/news-events/blogs/business-blog/2019/09/deceptive-certified-organic-claims-leave-consumers-verklempt

[65] "The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)," *IEEE,* 2021, https://standards.ieee.org/industry-connections/ecpais.html

[66] Standards by ISO/IEC JTC 1/SC 42 Artificial intelligence, International Organizations for Standardization, accessed August 2021, https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

In January 2021, the US Congress mandated that NIST identify and provide "standards, guidelines, best practices, methodologies, procedures, and processes for developing trustworthy AI systems."[67] Within two years, NIST is required to develop an AI risk management framework that enables the assessment of "trustworthy" AI and identification of appropriate risk mitigation strategies on a voluntary basis in the public and private sectors.[68] NIST is to establish common definitions and characterizations for AI principles, such as explainability, transparency, and fairness. In June 2021, NIST issued a draft report defining different types of bias and mitigation strategies—an important first step in establishing standards for appropriate oversight and risk mitigation.[69] Given the important role that standards and conformity assessments are expected to play in supporting compliance with the proposed EU AI Act, more explicit linkages should be made between the technical assessments of AI systems and their potential downstream human rights impacts as these efforts evolve.

In March 2021, the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) established Joint Technical Committee 21 on Artificial Intelligence (CEN-CLC/JTC 21) to proceed with the development and adoption of standards for AI and related data, including international standards already available or under development from organizations like ISO/IEC JTC 1 and its subcommittees, such as SC 42 Artificial Intelligence. CEN-CLC/JTC 21 will focus on producing standardization deliverables that address "European market and societal needs, as well as underpinning EU legislation, policies, principles, and values."[70]

The European Commission issued a report in 2021 outlining relevant standards that support compliance with its AI Act, including standards from the IEEE and ISO to guide appropriate data governance; risk management; technical data and record keeping; transparency and accountability; human oversight; accuracy, robustness, and cybersecurity; and implementation of a quality management system to ensure compliance with regulation.[71]

As AI standards and conformity assessments mature, implementation of certification schemes designed to operationalize them are gaining prominence. Certification can be defined as the "attestation that a product, process, person, or organization meets specified criteria."[72] In AI, certifications are emerging for both the technology itself (e.g., training data and model attributes) and the development process (e.g., organizational ethics and risk management processes), or a combination of both. Certifications can be voluntary or mandatory, self-assessed or third-party assessed. At this stage, self-certifications are the most common with third-party certifications

[67] NIST was assigned the task of creating an AI risk management framework in the National Artificial Intelligence Initiative Act of 2020 (the AI Act), which was included in the 2021 National Defense Authorization Act; "H.R.6395 - National Defense Authorization Act for Fiscal Year 2021: Actions," Congress.gov, Library of Congress, accessed April 2, 2021, https://www.congress.gov/bill/116th-congress/house-bill/6395/text.
[68] *Ibid*., 49.
[69] Reva Schwartz et al. "A Proposal for Identifying and Managing Bias in Artificial Intelligence," *National Institute of Standards and Technology (NIST),* June 2021, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf
[70] Other national standards organizations are undertaking similar efforts. In Canada, the national counterpart to NIST and CEN-CENELEC recently received additional funding from the Canadian government to advance the development and adoption of AI standards, including risk management standards and conformity assessment schemes for AI.
[71] S. Nativi, and S. De Nigris. "AI Watch: AI Standardization Landscape." (2021).
[72] Peter Cihon et al., "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries." *IEEE Transactions on Technology and Society* (2021).

being proposed for high-risk applications of AI. In the EU's AI Act, for example, developers of "low-risk" applications can perform voluntary self-assessments and certain "high-risk" applications are required to complete mandatory third-party "conformity assessments." Self-assessments or self-certifications are widely used in many industries but may lack legitimacy due to the inherent potential for conflicts of interest and low accountability for lack of enforcement. Third-party assessments are more rigorous, but can be extraordinarily costly and require qualified assessors, which can be difficult to find for complex AI systems.[73] The development of software-based assessment and certification methods that automate and streamline regulatory compliance is one way that researchers and industry are investigating new ways of implementing AI governance at scale.[74]

While AI certification processes are still at an early stage, initiatives like the Responsible Artificial Intelligence (RAI) Certification developed by the Responsible AI Institute in collaboration with the World Economic Forum hold promise.[75] One of the first independent, accredited certification programs to emerge, the RAI Certification seeks to support the implementation of responsibly built AI systems through an objective third-party review process. Certification can incentivize implementation of appropriate risk identification and mitigation strategies; however, there are significant challenges to successful implementation. For example, false positives where certification is provided even though certain criteria have not been met or false negatives where certification is not provided even though all criteria have been satisfied.

Development of appropriate standards and certifications will depend on access to high quality data about AI operations in specific contexts.[76] Data collection from AI monitoring and measurement, therefore, will be critical to the effectiveness of standards and certifications and essential to protecting human rights in high-risk contexts and applications. For human rights, appropriately defining evaluation criteria, assessment, and verification processes contained in standards and certifications will be critical. In a field where concepts of "fair," "accountable," and "trustworthy" AI are still under development, defining and enforcing appropriate procedures to uphold human rights in AI is equally muddled. While promising to uncover human rights risks of AI and whether strategies are in place to mitigate these risks, use of standards and certifications to indicate human rights due diligence should be cautiously implemented.

CONCLUSIONS AND RECOMMENDATIONS

Given the important human rights considerations raised by the use of AI systems, closer linkages should be made between the study and practice of AIAs and lessons learned from the implementation of HRIAs. In particular, AIAs can play an important role in identifying the technical foundations needed to promote respect for human rights. In turn, international human rights law

[73] *Ibid*., 54.
[74] Gillian Hadfield: Regulatory technologies can solve the problem of AI, University of Toronto Schwartz Reisman Institute for Technology and Society, last modified April 19, 2021; see also Gillian Hadfield, Rules for a Flat World, Oxford University Press, May 14 (2020); Jack Clark & Gillian K. Hadfield, 2019. "Regulatory Markets for AI Safety," Papers 2001.00078, arXiv.org
[75] "RAI Certification Beta," Responsible Artificial Intelligence Institute, accessed Sept. 1, 2021, https://www.responsible.ai/certification.
[76] Jess Whittlestone and Jack Clark. "Why and How Governments Should Monitor AI Development." *arXiv preprint arXiv:2108.12427* (2021).

can serve as a helpful guide for identifying connections between AI systems' technical features and human rights implications, especially for vulnerable individuals and communities.

Significant work remains to develop best practices for successful implementation of AIAs and HRIAs for AI, including considerations related to how AIAs should integrate features of HRIAs and their appropriate scope, structure, scalability, timing, and administrative burden. In this respect, however, the emergence of common approaches and methodologies for AIAs and HRIAs for AI will be aided by the development of human-rights based technical standards, conformity assessments, and certification schemes, as well as customized guidance for their implementation in a variety of contexts.

As policymakers advance discussions on draft legislation covering algorithmic and AI systems, the following recommendations could help ensure respect for human rights—within the EU and beyond.

1. Legislation requiring large platforms to conduct risk assessments of their operations should require considerations of risks to international human rights and freedoms and disclosure of actions taken to mitigate these risks.
2. AI quality management and risk management standards developed should explicitly address risks to international human rights and freedoms.
3. Organizations and consortia empowered by legislation to perform independent risk assessments, conformity assessments, and audits should include personnel and civil society organizations with proven human rights expertise.
4. Self-assessment regimes for AI systems should be complemented by post-market monitoring that triggers independent conformity assessments or audits in certain cases; for instance, in situations where a company's violation of its obligations raises human rights concerns. Clear avenues should also be established for people affected by AI systems, or groups representing them, to flag harms and thereby trigger investigations by enforcement bodies.
5. A model risk assessment methodology that explicitly addresses human rights concerns triggered by AI systems should be developed with the involvement of all relevant stakeholders.
    a. This should begin by defining approaches to human rights risk assessments of specific applications of algorithmic and AI systems, notably in the context of large digital platforms and high-risk AI systems by building off of best practices identified in the UNGPs, the technical assessment defined by the EU High-level Expert Group on AI, and emerging scholarship.
    b. The methodology should include procedures for linking the technical performance of AI systems with potential downstream human rights impacts on individuals and communities, in particular, those at a higher risk of vulnerability.
6. Given the critical role of human rights considerations in the development of AI standards, conformity assessments, and certification schemes, governments should establish meaningful opportunities and support mechanisms (e.g., subsidies for travel) for human

rights experts and civil society organizations to participate in these processes at the national and international levels.

7. To facilitate rights-respecting AI governance at scale, policymakers should ensure appropriate coordination of research & development spending with AI standardization pilot programs, such as regulatory sandboxes, to accelerate:
    a. the adoption of privacy-enhancing technologies, technical tools that enable bias and fairness detection and mitigation, or the continuous monitoring of AI system performance; and,
    b. The customization of certification schemes to ensure their robust implementation in a variety of contexts, beginning with high-risk applications of AI systems.

ACKNOWLEDGMENT

An earlier version of this paper was published as part of the Harvard Kennedy School Carr Center for Human Rights Policy Discussion Paper Series.