HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

FALL 2021

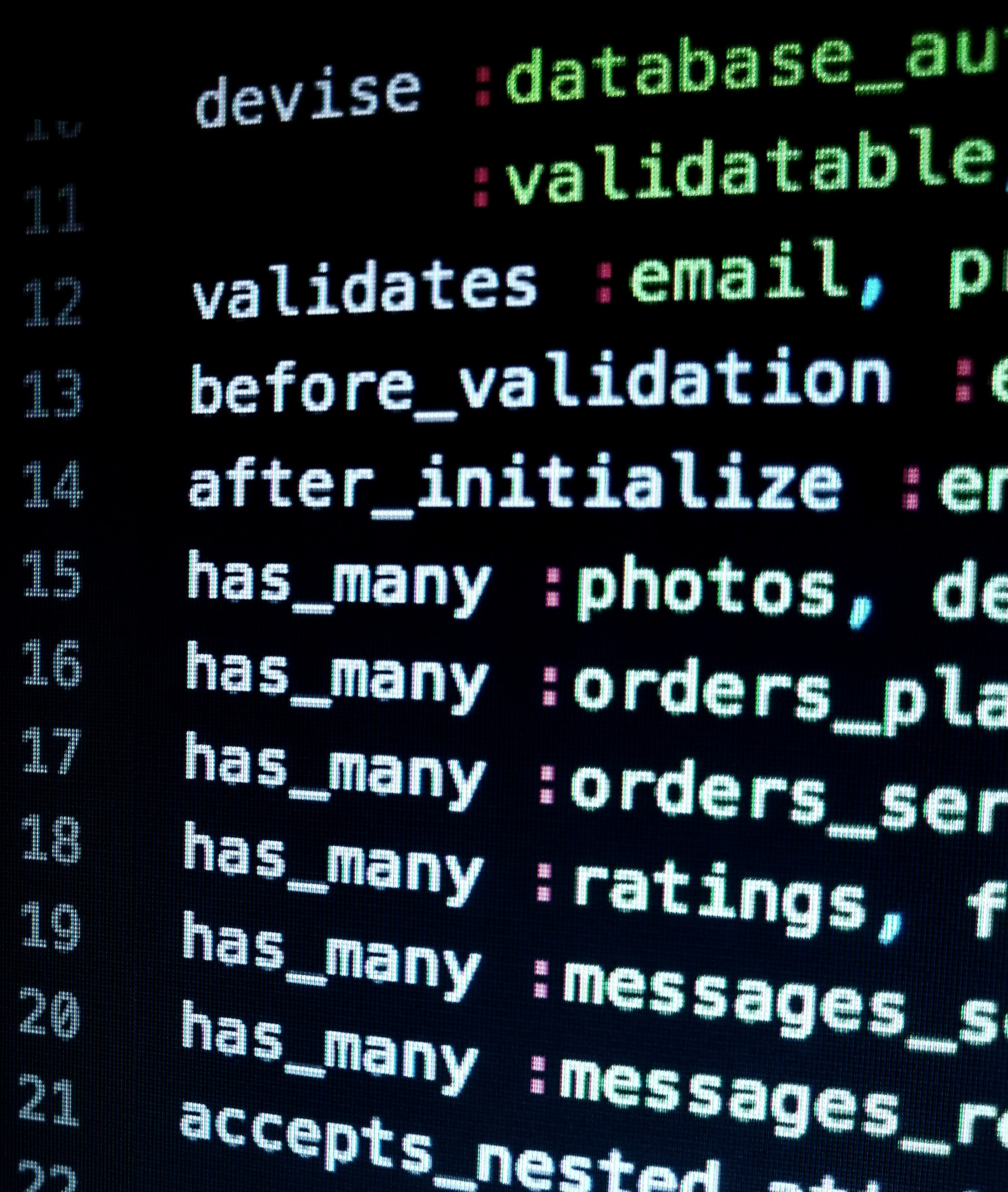
CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL

Human Rights Implications of Algorithmic Impact Assessments

Priority Considerations to Guide Effective Development and Use

Brandie Nonnecke
Philip Dawson

Carr Center
Discussion Paper



```
devise :database_au
       :validatable
validates :email, p
before_validation :e
after_initialize :er
has_many :photos, de
has_many :orders_pla
has_many :orders_ser
has_many :ratings, f
has_many :messages_s
has_many :messages_r
accepts_nested
```

Human Rights Implications of Algorithmic Impact Assessments: Priority Considerations to Guide Effective Development and Use

Carr Center for Human Rights Policy
Harvard Kennedy School, Harvard University
October 21, 2021

Brandie Nonnecke and Philip Dawson
Technology and Human Rights Fellows
Carr Center for Human Rights Policy

The views expressed in the Carr Center Discussion Paper Series are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

Table of Contents

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 AI. 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 the implementation of human rights impacts assessments (HRIAs) to assess AI. The benefits and drawbacks from recent implementations of AIAs and HRIAs to assess AI adopted by the public and private sectors are explored and considered in the context of an emerging trend toward the development of standards, certifications, and regulatory technologies for responsible AI governance practices. We conclude with priority considerations to better ensure that AIAs and their corresponding responsible AI governance strategies live up to their promise.

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 anticipate, 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, the purpose of an AIA is 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 takes a risk-based approach to AI governance, prohibiting certain harmful applications of AI and calling for the use of "conformity assessments" 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 in 2021).[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 risk assessments for AI to be implemented *ex ante* and for impact assessments to be conducted *ex post* to

[1] Artificial intelligence (AI) refers to a computer system capable of performing tasks that require human-level intelligence, such as decision-making, visual perception, and speech recognition. Methods for doing so are wide ranging and vary significantly in complexity, including algorithms, predictive models, computer vision, deep learning, machine learning, natural language processing, neural nets, and more.

[2] Bryan Casey, Ashkon Farhangi, and Roland Vogl, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise," *Berkeley Technology Law Journal* 34, no. 1 (2019): 143–188; Timnit Gebru et al., "Datasheets for Datasets," *arXiv* 1803.09010 (2018), https://arxiv.org/abs/1803.09010; Margaret Mitchell et al., "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019): 220–29, http://doi.org/10.1145/3287560.3287596; Emanuel Moss et al., "Governing with Algorithmic Impact Assessments: Six Observations," *AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (2020), https://dx.doi.org/10.2139/ssrn.3584818.

[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," *Boston University Law Review* 100 (2020): 1315–76, https://www.bu.edu/bulawreview/files/2020/09/SELBST.pdf.

[5] European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM/2021/206 (April 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

[6] US Congress, Senate, *Algorithmic Accountability Act of 2019*, S 1108, 116th Cong., introduced in Senate April 10, 2019, https://www.congress.gov/bill/116th-congress/senate-bill/1108; Grace Dill, "Sen. Wyden to Reintroduce AI Bias Bill in Coming Months," *MeriTalk*, February 19, 2021, https://www.meritalk.com/articles/sen-wyden-to-reintroduce-ai-bias-bill-in-coming-months/.

[7] US Congress, House, Committee on Appropriations, *Commerce, Justice, Science and Related Agencies Appropriations Bill, 2021 - Report Together With Minority Views*, 116th Cong., 2d sess., 2020, H. Rep. 116–455, 23, https://appropriations.house.gov/sites/democrats.appropriations.house.gov/files/July%209th%20report%20for%20circulation_0.pdf.

"increase public transparency 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 AIA tools 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 great promise to promote the development of regulatory, policy, and governance mechanisms by government and corporate actors to identify potential harms, civil society organizations have warned that using a risk-based AIA approach may be insufficient.[10] Most guidance for implementation of AIAs indicates that 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 evade proper oversight. This is especially problematic in the context of emerging legislation that places the onus of determining risk level on the entity developing the AI. 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 isn't to say that implementation of AIAs cannot provide benefits now, but 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 perpetuate the harms they seek to mitigate.

In this research, we explore the emerging integration of the human rights framework into AI governance strategies, the development and use of AIAs, and the potential benefits and risks they pose to human rights. 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 with a discussion of priority considerations to help guide the effective development and use of AIAs to better ensure that they live up to their promise.

AI Governance and Human Rights

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] 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.

Out of the over 170 existing sets of AI principles, there is seldom consensus in the interpretation of the principles in practice.[13] AI principles have been developed by diverse institutions (e.g., academia, civil society, governments) with varying multi-stakeholder 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

[8] US National Security Commission on Artificial Intelligence, *Final Report* (2021), 395, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[9] Government of Canada, *Directive on Automated Decision-Making* (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, February 17, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach/.

[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, 2020, http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420.

[13] Algorithm Watch, "AI Ethics Guidelines Global Inventory."

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 Organisation for Economic Co-operation and Development (OECD) AI Principles, which 42 countries have pledged to uphold, specifically call for the protection of human rights.[18] The European Commission's Artificial Intelligence Act (AI Act) seeks to ensure that AI systems respect human rights through development of a risk-based approach to AI governance and oversight.[19] The White House Office of Science and Technology Policy (OSTP) in its National AI Initiative called out the need to ensure that 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 AIA tool that seeks to measure and mitigate the human rights harms of ADS used in public services.[21] Data-intensive 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]

> **"The ill-conceived development and deployment of AIAs pose substantial risk themselves... If done inappropriately, their use may ultimately perpetuate the harms they seek to mitigate."**

Centering human rights within AI governance strategies can help operationalize AI principles across sectors, international contexts, and domain application areas.[24] Through their codification in charters, case law, regulation, and industry standards, human rights norms and values have gained

[14] Mark Latonero, *AI Principle Proliferation as a Crisis of Legitimacy*, Carr Center for Human Rights Policy, 2020, https://carrcenter.hks.harvard.edu/files/cchr/files/mark_latonero_ai_principles_6.pdf.

[15] Mark Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Data & Society, 2018, https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf; 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* 41 (July 2021), https://doi.org/10.1016/j.clsr.2021.105561; Eileen Donahoe and Megan MacDuffee Metzger, "Artificial Intelligence and Human Rights," *Journal of Democracy* 30, no. 2 (2019): 115–126.

[16] The Institute of Electrical and Electronics Engineers (IEEE) Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligence Systems*, 1st ed., IEEE, 2019, https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html.

[17] "Asilomar AI Principles," Future of Life Institute, 2017, https://futureoflife.org/ai-principles/.

[18] Organisation for Economic Co-operation and Development, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD%20-LEGAL-0449.

[19] European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*.

[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] Government of Canada, *Directive on Automated Decision-Making*.

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

[23] Microsoft, *Human Rights Annual Report*, 2018, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE2FMZY; "Intel Human Rights Impact Assessment," Article One Advisors, accessed August 22, 2021, https://www.articleoneadvisors.com/intel-hria.

[24] Mantelero and Esposito, "An evidence-based methodology"; Donahoe and MacDuffee Metzger, "Artificial Intelligence and Human Rights"; 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.

The human rights framework can provide the substantive foundation and governance architecture needed to produce greater specificity in defining and operationalizing AI principles.

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]

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 by 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 the transparency, explainability, and accountability, are not only desirable from a commercial or ethical standpoint, but also 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, pre-empting 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 HRIAs for AI are also on the rise.[29] As AIAs and HRIAs for AI continue to gain prominence, it is important to consider *how* these approaches 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 to support accountability.[30] 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).[31]

[25] Donahoe and MacDuffee Metzger, "Artificial Intelligence and Human Rights."

[26] UN General Assembly, 183rd Plenary Meeting, Resolution 217A, A Universal Declaration of Human Rights, A/RES/217 (December 10, 1948), https://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/217(III); "The Core International Human Rights Instruments and their monitoring bodies," United Nations Human Rights Office of the High Commissioner (OHCHR), 2021, https://www.ohchr.org/en/professionalinterest/pages/coreinstruments.aspx; OHCHR, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, June 16, 2011, https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf; Latonero, *Governing Artificial Intelligence*.

[27] Mantelero and Esposito, "An evidence-based methodology."

[28] Australian Human Rights Commission, *Human Rights and Technology Final Report (2021)*, 2021, https://tech.humanrights.gov.au/downloads.

[29] Australian Human Rights Commission, *Human Rights and Technology Final Report (2021)*; Council of Europe Ad Hoc Committee on Artificial Intelligence Policy Develop Group, *Human Rights, Democracy and Rule of Law Impact Assessment of AI Systems*, CAHAI-PDG(2021)05, May 21, 2021, https://rm.coe.int/cahai-pdg-2021-05-2768-0229-3507-v-1/1680a291a3.

[30] Jacob Metcalf et al., "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021): 735–46.

[31] 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/.

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 and enforce 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.[32]

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."[33] 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 reviews, risk management, and due diligence processes. 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 an AIA and how it is implemented in practice.[34]

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."[35] 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, before 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.

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 an AIA and how it is implemented in practice.

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.[36] Microsoft publishes a "Human Rights Annual Report" within which the human rights effects of its technologies are explored and risk mitigation strategies taken are discussed. Facebook commissioned an HRIA to evaluate its role in the genocide of the Rohingya in Myanmar.[37] The HRIA was found to be largely ineffective at uncovering the human rights harms of Facebook's AI-enabled tools and identifying appropriate mechanisms to mitigate those harms moving forward.

[32] Moss et al., *Assembling Accountability*, 28.

[33] Moss et al., *Assembling Accountability*, 26.

[34] Andrew D. Selbst, "An Institutional View of Algorithmic Impact Assessments," *Harvard Journal of Law and Technology* 35 (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867634.

[35] Mark Latonero and Aaina Agarwal, *Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar*, Carr Center for Human Rights Policy, 2021, https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf.

[36] "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/.

[37] Latonero and Agarwal, *Human Rights Impact Assessments for AI*.

In the remainder of this section, we explore proposed and existing AIAs 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 of an AIA to evaluate and mitigate harms of ADS in government public service delivery. We next consider the EU's risk-based approach to AI governance and its proposed implementation of "conformity assessments" to identify and mitigate AI risks emerging from the private sector. We then explore the implementation of HRIAs for AI and how these may differ from, complement, or should be integrated into AIAs to better ensure the protection of fundamental human rights.

In 2019, the Canadian government released its Directive on Automated Decision-Making (the Directive).[38] The Directive's principal objectives were to ensure that 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 that harmful effects of algorithms on administrative decisions are assessed and reduced.[39] To this end, the Directive includes an accompanying AIA 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."[40] 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.[41] 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.

In April 2021, the European Commission released its draft "Artificial Intelligence Act" (AI Act).[42] The AI Act takes a risk-based approach to AI regulation, establishing four levels of risk: minimal, limited, high, and unacceptable. The Act requires different levels of oversight for limited and high-risk AI applications. Applications that fall within the category of unacceptable risk are forbidden (e.g., use of AI that is capable of manipulating individuals through subliminal techniques).

The deploying entity can determine whether applications fall under the minimal-risk categorization. Applications posing limited risk would have transparency obligations. High-risk applications (such as use of AI in critical infrastructure, medical devices, and education) or applications that pose a risk to health, safety, and/or fundamental rights (such as remote biometric identification, credit scoring, or hiring decisions) would be subject to *ex ante* conformity assessments to be conducted by independent third parties. 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.

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 an AIA 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 technical 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

[38] Government of Canada, *Directive on Automated Decision-Making*.

[39] Government of Canada, *Directive on Automated Decision-Making*, 35.

[40] Government of Canada, *Directive on Automated Decision-Making*, 35.

[41] Tom Cardoso and Bill Curry, "National Defence 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 September 7, 2021, https://search.open.canada.ca/en/od/?search_text=AIA.

[42] European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*.

of AI-driven systems.[43] 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 an AIA 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. For one, 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 an HRIA tends to be broader and more forward-looking than that of an AIA 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 activity, the presence of adequate legal protections or governance structures, the track record of supply chain 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 stand 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 may be to hire external consultants to help adapt existing policies and procedures to the AI context, upskill employees, or conduct in-depth, standalone AIAs and/or HRIAs for applications believed to be higher risk.

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.

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."[44] 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.[45] In addition, the study concluded that HRIAs should be conducted at the earliest stages of the design or conception of AI systems.

[43] Metcalf et al., "Algorithmic Impact Assessments and Accountability."

[44] Metcalf et al., "Algorithmic Impact Assessments and Accountability," 41.

[45] Latonero and Agarwal, *Human Rights Impact Assessments for AI*.

The *ex ante* HRIA conducted on Alphabet-affiliate Sidewalk Labs's "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. 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 rapid acceleration and enhancement of existing human rights-based governance efforts related to the project in a relatively short period of time. 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.[46] 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 technologies that can help automate verification of compliance may be more appropriate as a complement to human oversight.[47] 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.[48]

Ultimately, design specifications and implementation tactics for AIAs and HRIAs will have to be tailored to the complexity, scale, 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. 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 risk, 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.

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, or 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." 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 Ethically Aligned Design 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

[46] Mantelero and Esposito, "An evidence-based methodology."

[47] Gillian Hadfield, *Rules for a Flat World* (Oxford: Oxford University Press, 2016); Jack Clark and Gillian K. Hadfield, "Regulatory Markets for AI Safety," *arXiv* 2001.00078 (2019), https://arxiv.org/abs/2001.00078.

[48] Daniel Zhang et al., *The AI Index 2021 Annual Report*, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, March 2021, https://aiindex.stanford.edu/report/.

international human rights framework and the relevance of the UNGPs. Additionally, the IEEE has established 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, which has the potential to raise consumer trust and market power.[49]

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).[50] 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.

In January 2021, Congress mandated that NIST identify and provide "standards, guidelines, best practices, methodologies, procedures, and processes for developing trustworthy AI systems."[51] 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.[52] 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.[53] 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."[54]

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.[55]

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."[56] In AI, certifications are emerging for the technology itself (e.g., training data and model attributes), the development

[49] "The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)," IEEE, accessed August 22, 2021, https://standards.ieee.org/industry-connections/ecpais.html.

[50] "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.

[51] 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. See US Congress, House, *William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021*, HR 6395, 116th Cong., introduced in House March 26, 2020, https://www.congress.gov/bill/116th-congress/house-bill/6395/text.

[52] US Congress, House, *National Defense Authorization Act*, 49.

[53] 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.

[54] 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.

[55] Stefano Nativi and Sarah De Nigris, *AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework*, EUR 30772 EN, Publications Office of the European Union, 2021, https://doi.org/10.2760/376602.

[56] Peter Cihon et al., "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries," *IEEE Transactions on Technology and Society* (2021), https://doi.org/10.1109/TTS.2021.3077595.

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 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.[57] 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.[58]

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.[59] 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, such as 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.[60] 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 of 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.

Conclusion

Risk-based approaches to AI governance, including tools such as AIAs, are likely to play a central role in AI governance strategies. 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, international human rights law 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 harmonized technical standards, conformity assessments, and certification schemes, as well as guidance for their implementation in a variety of contexts..

Acknowledgment

This report was supported in part and reviewed by Access Now, a global human rights organization. We also want to thank the following individuals for their thoughtful review: Kathy Baxter, Ashley Casovan, Danielle Cass, Peter Cihon, Camille Crittenden, Mark Latonero, Marc-Étienne Ouimette, Ed Santow, and Craig Shank.

[57] Cihon et al., "AI Certification," 54.

[58] Gillian K. Hadfield, "Regulatory technologies can solve the problem of AI," Schwartz Reisman Institute for Technology and Society, University of Toronto, April 19, 2021, https://srinstitute.utoronto.ca/news/hadfield-torstar-ai-debate. See also Hadfield, *Rules for a Flat World*; Clark and Hadfield, "Regulatory Markets for AI Safety."

[59] "RAI Certification Beta," Responsible Artificial Intelligence Institute, accessed September 1, 2021, https://www.responsible.ai/certification.

[60] Jess Whittlestone and Jack Clark, "Why and How Governments Should Monitor AI Development," *arXiv* 2108.12427 (2021), https://arxiv.org/abs/2108.12427.

Carr Center Discussion Paper Series

Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

Statements and views expressed in this report are solely those of the author and do not imply endorsement by Harvard University, the Harvard Kennedy School, or the Carr Center for Human Rights Policy.

Copyright 2021, President and Fellows of Harvard College
Printed in the United States of America

This publication was published by the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University.

Copyright 2021, President and Fellows of Harvard College
Printed in the United States of America

79 JFK Street
Cambridge, MA 02138

617.495.5819
https://carrcenter.hks.harvard.edu