CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL

Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar

Mark Latonero &
Aaina Agarwal

Carr Center
Discussion Paper

Image from United to End Genocide

Human Rights Impact Assessments for AI:
Learning from Facebook's Failure in Myanmar

Mark Latonero, PhD
2019-2020 Technology & Human Rights Fellow, Carr Center
Email: mark@latonero.com

Aaina Agarwal, JD
Human Rights Lawyer and Advisor, Algorithmic Justice League

Abstract

Human rights impact assessments (HRIAs) have recently emerged as a way for technology companies to identify, mitigate, and remedy the potential risks and harms of artificial intelligence (AI) and algorithmic systems. The purpose of this paper is to assess whether HRIAs are a tool fit for purpose for AI. Will HRIAs become an effective tool of AI governance that reduces risks and harms? Or, will they become a form of AI "ethics washing" that permits companies to hide behind a veneer of human rights due diligence and accountability? This paper finds that HRIAs of AI are only in their infancy. Simply conducting such assessments with the usual methods will miss the mark for AI and algorithmic systems, as demonstrated by the failures of the HRIA of Facebook in Myanmar. Facebook commissioned an HRIA after UN investigators found that genocide was committed in the country. However, the HRIA did not adequately assess the most salient human rights impacts of Facebook's presence and product in Myanmar. HRIAs should be updated if they are to be used on AI and algorithmic systems. HRIAs for AI should be seen as an analysis of a sociotechnical system wherein social and technical factors are inherently intertwined and interrelated. Interdisciplinary expertise is needed to determine the appropriate methods and criteria for specific contexts where AI systems are deployed. In addition, HRIAs should be conducted at appropriate times relative to critical stages in an AI development lifecycle and function on an ongoing basis as part of a comprehensive human rights due diligence process. Challenges remain, such as developing methods to identify algorithmic discrimination as one of the most salient human rights concerns when it comes to assessing AI harms. In addition, a mix of voluntary actions and mandatory measures may be needed to incentivize organizations to incorporate HRIAs for AI and algorithmic systems in a more effective, transparent, and accountable way. The paper concludes with considerations for the technology sector, government, and civil society.

Introduction

A human rights impact assessment (HRIA) is a tool to evaluate the potential or actual impact of an organization's strategy, practice, or product on people's human rights. While HRIAs can be used by any organization in any sector, these assessments are used primarily in business, and have recently emerged as a way for technology companies to identify, mitigate, and remedy the potential risks and harms of AI and algorithmic systems.

To date, HRIAs have been conducted by a number of leading artificial intelligence (AI) companies. For example, Microsoft conducted an HRIA with the goal of identifying human rights risks and developing mitigation strategies around its AI applications and impacts.[1] Facebook commissioned an HRIA for Myanmar after UN investigators found the genocide was committed in the country, which this paper will discuss in detail below. Facebook has also commissioned HRIAs for Indonesia, Cambodia, and Sri Lanka.[2] And, the CEO of Alphabet Inc. (Google) has conveyed that human rights assessments are a key component of AI regulation, stating, "...principles that remain on paper are meaningless. So we've also developed tools to put them into action, such as...conducting independent human-rights assessments of new products."[3]

The purpose of this paper is to assess whether HRIAs are a tool fit for purpose when applied to AI and algorithmic systems. Will HRIAs become an effective tool of AI governance that reduces risks and harms to human rights? Or, will they become a form of AI "ethics washing" that permits companies to hide behind a veneer of human rights due diligence and accountability?[4]

The first challenge when it comes to HRIAs for AI is epistemological: How can we *know* if an AI system—and the business decisions made to deploy such a system—has impacted the fundamental human rights of an individual or group? The existing knowledge of how to conduct HRIAs has been established over decades of practice in business sectors as varied as garment manufacturing and mining. AI and algorithmic systems are unique in ways that can confound existing HRIA methodology. Today, there is little standardized knowledge, methodology, or expertise for investigating the unique human rights impacts of AI and algorithmic systems, business models, and products.

In this paper, we find that HRIAs of AI are only in their infancy. Simply conducting these assessments with the usual methods will miss the mark for AI and algorithmic systems, as demonstrated by the failures of the HRIA of Facebook in Myanmar. In order for HRIAs to be established as a legitimate and effective tool for AI and algorithmic systems, they should:

- align with the UN Guiding Principles on Business and Human Rights and be regarded by companies as one tool in a comprehensive and organizationally embedded human rights due diligence process;

- be conducted on an on-going basis following an early baseline assessment of the human rights context relative to the lifecycle of AI system and products;

- focus on evaluating the technical performance, business models, and other defining features of AI, as a sociotechnical system, relative to the most salient human rights issues;

- integrate the necessary interdisciplinary expertise (including from civil society) and emerging approaches, such as algorithmic impact assessments and research on algorithmic discrimination; and

- be backed by global policy and legislation that supports ongoing HRIAs and human rights due diligence in businesses as well as organizations in government and other sectors that develop, procure, and use AI.

1. Microsoft Corporation, "Human Rights Annual Report: Fiscal Year 2018," Microsoft Corporation, 2018, 22–30, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE2FMZY.

2. See Alex Warofka, "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar," *Facebook Newsroom* (blog), November 5, 2018, https://about.fb.com/news/2018/11/myanmar-hria/; And Miranda Sissons and Alex Warofka, "An Update on Facebook's Human Rights Work in Asia and Around the World," *Facebook Newsroom* (blog), May 12, 2020, https://about.fb.com/news/2020/05/human-rights-work-in-asia.

3. Sundar Pichai, "Why Google Thinks We Need to Regulate AI," *Financial Times*, January 19, 2020, https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04.

4. See Karen Yeung, Andrew Howes, and Ganna Pogrebna, "AI Governance by Human Rights Centred-Design, Deliberation and Oversight: An End to Ethics Washing," in *The Oxford Handbook of Ethics of AI*, eds. Markus Dubber, Frank Pasquale, and Sunit Das (Oxford: Oxford University Press, 2020).

The paper begins with a brief discussion of HRIAs in the field of business and human rights in relation to the UN Guiding Principles on Business and Human Rights. Next is a discussion of Facebook in Myanmar and the HRIA commissioned by the company in the wake of genocide against the Rohingya people. We then provide examples of what is needed for HRIAs to be effective in the AI space, considering the timing and criteria for evaluating real-world product deployments and algorithmic operations. The paper discusses current regulatory efforts underway to mandate human rights due diligence. Finally, the paper provides considerations for stakeholders in business, government, and civil society.

ASSESSING IMPACT IN BUSINESS AND HUMAN RIGHTS

HRIAs measure impact against specific rights defined in the Universal Declaration of Human Rights; the International Covenant on Economic, Social and Cultural Rights; the International Covenant on Civil and Political Rights; and additional treaties signed and ratified by nation-states. An individual experiences a negative human rights impact as an infringement of their right to non-discrimination, privacy, life, political participation, assembly, education, freedom of opinion and expression, and additional rights. Also, members of certain groups, such as migrant workers, children, and persons with disabilities, have specific rights under international human rights law.

The use of HRIAs makes an implicit claim to legitimacy by appealing to human rights law, and the way it has evolved, to encompass business. When the modern human rights system began after World War II, it was focused on binding states to international human rights law through a network of treaty obligations. Generally, only nation-states can become parties to treaties. While human rights treaties operate as a form of hard law, the integration and enforcement of human rights law in the legal regimes of individual countries vary widely. Yet, the presumption is that a violation of human rights law carries greater international accountability.

Those working on AI might ask: if human rights produce legally binding obligations for states, how could it engender accountability for a tech company? Although states have historically been the primary duty holders of international human rights obligations, whether by treaty or customary international law, non-state actors—like companies and investors—also have obligations under various aspects of the human rights system. While the human rights system initially evolved to focus on state

obligations, over the past few decades there has been a significant global expansion of the private sector that has increased the business impact on human rights. As a result, human rights advocates began to consider how to articulate obligations for corporations.

A historical challenge for the field of business and human rights has been establishing a coherent, authoritative, and comprehensive system to govern stakeholder expectations and behavior.[5] In 2008, Special Representative John Ruggie, who was appointed by then Secretary-General Kofi Annan to investigate "the issue of human rights and transnational corporations and other business enterprises," recommended the "Protect, Respect and Remedy" framework.[6] In 2011, Ruggie operationalized this framework in the publication of the UN Guiding Principles on Business and Human Rights (UN Guiding Principles).[7] The three pillars of the framework refer to (1) the *state duty* to *protect* against human rights violations by businesses and other third parties, (2) the *corporate responsibility* to *respect* human rights by avoiding harming others rights and addressing adverse impacts, and (3) the need to enable victims to access effective *remedy*.[8]

A hallmark achievement of the UN Guiding Principles within the human rights system was to clarify the distinct but complementary *duties of states* and *responsibilities of businesses* related to business and human rights.[9] The UN Guiding Principles offer a baseline and shared reference point from which all the actors who play a role related to business and human rights—states, companies, and civil society organizations—can surface recommendations to develop further with one another.

The UN Guiding Principles call upon companies to meet their responsibility to respect human rights by adopting a "human rights due diligence process." The first requirement of this process is for companies to *assess* their actual and potential human rights impact, with the additional requirement that companies *integrate* and act upon these findings. Assessment and integration, then, are the key components of the corporate due diligence process that is required under the UN Guiding Principles.

Once an assessment identifies a human rights impact, the UN Guiding Principles offer a three-tiered framework to determine the company's relationship to the harm, distinguishing between the harms it may *cause*, those it has *contributed to*, and, those it may be *directly linked* to through a business relationship.[10] These descending distinctions are posed to identify the appropriate course of remedial action that a company is required to take following an assessment.[11]

5. John Ruggie, "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework," United Nations, March 21, 2011, 3 https://undocs.org/en/A/HRC/17/31.

6. John Gerard Ruggie, "Protect, Respect and Remedy: A United Nations Policy Framework for Business and Human Rights," *Proceedings of the Annual Meeting (American Society of International Law)* vol. 103 (2009): 282-87, http://www.jstor.org/stable/10.5305/procannmeetasil.103.1.0282.

7. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework," United Nations, June 16, 2011, https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.

8. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles."

9. John G. Ruggie, Letter to Ms. Saskia Wilkes and Mr. Johannes Blankenbach at the Business and Human Rights Resource Centre, September 19, 2019, https://media.business-humanrights.org/media/documents/files/documents/19092019_Letter_John_Ruggie.pdf.

10. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles," Principle 17.

11. "Integration" refers to the actions that a company is required to take to address the human rights impacts that have been identified by the assessment. See generally, United Nations Office of the High Commissioner for Human Rights, "Guiding Principles," Principle 19.

To meet these requirements, companies have turned to HRIAs as an assessment tool. Predating the publication of the UN Guiding Principles, HRIAs have been used to evaluate human rights impacts of business activities and products, for example, when a corporation's product is being released in an international market. In practice, HRIAs may have similarities to environmental and social impact assessments; however, HRIAs are based on benchmarks drawn from international human rights law as well as the UN Guiding Principles.[12] These benchmarks limit the extent to which tradeoffs are acceptable and reflect the core concepts and metrics that guide inquiry during the assessment process. It should be noted that the UN Guiding Principles do not lay out the precise methodology for conducting a human rights impact assessment, which would need to be tailored to the specific domain, industry, and context.[13]

AI GOVERNANCE AND HUMAN RIGHTS

What role does international human rights have in AI governance and algorithmic accountability? In recent years, a number of researchers and scholars have demonstrated the importance of human rights in addressing and mitigating AI risks and harms. There are numerous arguments for why human rights provide a strong foundation for AI governance out of the range of approaches and principles available today.

Existing research argues that human rights provide a global formulation of human values that can guide the design and deployment of AI.[14] Human rights provide a global vernacular that addresses the immediate challenge concerning the proliferation of AI standards and principles.[15] Human rights law maintains global legitimacy when it comes to holding both AI companies and governments using AI accountable.[16] Additionally, human rights can be seen as "AI governance with teeth" in its capacity to strengthen the enforcement of fairness, accountability, and transparency in AI,[17] while also offering an expansive network of existing multinational and multistakeholder fora and systems in which to deliberate the impact of AI.[18]

In addition, committing to effective human rights protection in the design of AI systems is part of ensuring the continuity of a democratic constitutional order wherein the adjudication of human rights conflicts can offer concrete guidance to those involved in the design and deployment of AI systems.[19] The human rights approach to algorithmic accountability can apply across the full algorithmic lifecycle—from conception to deployment—and complement technical solutions that are necessary albeit insufficient to address rights-based concerns.[20] For example, whereas a technical approach may surface a particular statistical bias, however defined, human rights can contribute to the solution by providing a method for understanding when this bias is prohibited or unlawful, and thereby constitutes harm.[21] Finally, a human rights approach requires AI designers to focus their attention to impacts on vulnerable and marginalized groups.[22]

The following section will consider the relevance of a human rights framework to address the impact of AI and algorithmic systems on the one hand, and the challenges of using HRIAs on the other, by examining the case of Facebook in Myanmar.

12. See generally, Nora Götzmann et al., "Human Rights Impact Assessment: Guidance and Toolbox," Danish Institute for Human Rights, 2016, https://respect.international/wp-content/uploads/2019/07/Human-Rights-Impact-Assessment-Guidance-and-Toolbox.pdf; Faris Natour and Jessica Davis Pluess, "Conducting an Effective Human Rights Impact Assessment: Guidelines, Steps, and Examples," Business for Social Responsibility, March 2013, http://www.bsr.org/reports/BSR_Human_Rights_Impact_Assessments.pdf.

13. For example, see John Ruggie, "Guiding Principles on Business and Human Rights," 18.

14. Mark Latonero, "Governing Artificial Intelligence: Upholding Human Rights and Dignity," Data & Society Research Institute, 2018, https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

15. Christiaan van Veen, "Artificial Intelligence: What's Human Rights Got to Do with it?" *Data & Society: Points*, May 14, 2018, https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5; Mark Latonero, "AI Principle Proliferation as a Crisis of Legitimacy," Carr Center for Human Rights Policy, August 2020, https://carrcenter.hks.harvard.edu/files/cchr/files/mark_latonero_ai_principles_6.pdf?m=1601910899.

16. Filippo Raso et al., "Artificial Intelligence & Human Rights: Opportunities & Risks," Berkman Klein Center for Internet & Society, 2018, http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439; Eileen Donahoe and Megan MacDuffee Metzger, "Artificial Intelligence and Human Rights," *Journal of Democracy* vol. 30, no. 2 (April 2019): 115–126, https://www.journalofdemocracy.org/articles/artificial-intelligence-and-human-rights/.

17. Vidushi Marda, "Governance with Teeth: How Human Rights Can Strengthen FAT and Ethics Initiatives on Artificial Intelligence," Article 19, April 2019, https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

18. Jason Pielemeier, "AI & Global Governance: The Advantages of Applying the International Human Rights Framework to Artificial Intelligence," Centre for Policy Research, February 26, 2019, https://cpr.unu.edu/ai-global-governance-the-advantages-of-applying-the-international-human-rights-framework-to-artificial-intelligence.html.

19. Yeung, Howes, and Pogrebna, "AI Governance by Human Rights-Centred Design, Deliberation and Oversight," 5-6.

20. Lorna McGregor, Daragh Murray, and Vivian Ng, "International Human Rights Law as Framework for Algorithmic Accountability," *International and Comparative Law Quarterly* vol. 68, no. 2, (2019): 325–26, https://doi.org/10.1017/S0020589319000046.

21. McGregor, Murray, and Ng, "International Human Rights Law as Framework for Algorithmic Accountability."

22. Corrine Cath et al., "Leap of FATE: Human Rights as a Complementary Framework for AI Policy and Practice," in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery)*, 702, doi:10.1145/3351095.3375665.

Facebook's HRIA and Genocide in Myanmar

In 2018, the Independent International Fact-Finding Mission on Myanmar, established by the United Nations Human Rights Council, issued a four-hundred-plus page report finding that the dominant ethnic Buddhist-led government of Myanmar had committed genocide against the Rohingya, an ethnic Muslim minority residing primarily in the country's Rakhine state.[23] The Mission established consistent patterns of serious human rights violations against the Rohingya, and recommended that the military be prosecuted for genocide, crimes against humanity, and other war crimes.[24]

The UN Report extensively cites a wide range of Facebook posts by officials and community leaders to demonstrate the significant role that Facebook's platform played in disseminating hate speech that was coordinated by government propaganda campaigns and which perpetuated a narrative of the Rohingya as an existential threat to the country. The report provides a detailed analysis of how authorities used the platform to "exacerbate a climate in which hate speech thrives and in which individuals and groups may be more receptive to calls of incitement to violence."[25] The campaign was led by some prominent Buddhist monks, and military and government officials, who used the platform to share fabricated stories that vilified the Rohingya to a wide audience. For example, the government was found to have posted photos on Facebook—that were later debunked—of dismembered children and dead babies, claiming they were attacked by the Rohingya, as well as images purporting to show Rohingya burning their own homes.[26]

Along with online content, offline activity continued to escalate in violence. Government-led "clearance operations" expelled the Rohingya from the Rakhine state in 2016 and 2017.[27] Over 650,000 Rohingya fled the country following a military-led attack that burned their villages and consisted of widespread execution and rape. One soldier who participated in these crimes testified that the forces were instructed to "[k]ill all you see, whether children or adults."[28]

The UN Report notes that these violations of international human rights law, undertaken in total disregard for human life and dignity, can only be understood given the historical context, including years of concerted hate campaigns against the Rohingya on Facebook, and decades of rights erosion under an institutionalized system of oppression.[29] The report describes the complex relationship between the historical context of Myanmar as an emerging democracy rife with ethnic tension; the country's relative unfamiliarity with the internet; the adoption of Facebook as the country's main mode of public and government communication, news dissemination, and general internet use; and the public's perception of Facebook as a reliable source of information.[30]

DISCRIMINATION AGAINST THE ROHINGYA

The Rohingya arrived in the Rakhine region of Myanmar as early as the fifteenth century, with more arrivals during the nineteenth and early twentieth centuries when the region was governed by colonial rule as part of British India.[31] Since independence in 1948, successive governments in Myanmar have refuted the Rohingya's historical claim to the land, and have denied the group citizenship or recognition as one of the country's official ethnic groups. These rights violations were compounded by restrictions affecting their movement, subsistence, and development, among other forms of discrimination.[32] The denial of legal status left the group effectively stateless and without the right to vote. In the years leading up to the genocide, the government implemented a process requiring the Rohingya to register as "foreigners," which further contributed to their administrative erasure and eliminated any ability to advocate for or protect themselves under the system of law.[33]

23. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," United Nations, September 17, 2018, https://undocs.org/en/A/HRC/39/CRP.2.

24. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 63.

25. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 333.

26. Timothy McLaughlin, "How Facebook's Rise Fueled Chaos and Confusion in Myanmar," *Wired*, July 6, 2018, https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/.

27. See Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 177.

28. Hannah Beech, Saw Nang, and Marlise Simons, "'Kill All You See': In a First, Myanmar Soldiers Tell of Rohingya Slaughter," *New York Times*, September 8, 2020, https://www.nytimes.com/2020/09/08/world/asia/myanmar-rohingya-genocide.html.

29. See Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 177.

30. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 340.

31. See, for example, Eleanor Albert and Lindsay Maizland, "The Rohingya Crisis," Council for Foreign Relations, January 23, 2020, https://www.cfr.org/backgrounder/rohingya-crisis.

32. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 110.

33. Fortify Rights, "Tools of Genocide: National Verification Cards and the Denial of Citizenship of Rohingya Muslims in Myanmar," Fortify Rights, September 2019, 9–10, fortifyrights.org/downloads/Tools%20of%20Genocide%20-%20Fortify%20Rights%20-%20September-03-2019-EN.pdf.

For decades the Rohingya have faced countless examples of brutal institutionalized discrimination, including restrictions on their rights to employment, education, health, expression, religion, and more.[34] As noted in the UN Report:

> The Rohingya are in a situation of severe, systemic and institutionalised oppression from birth to death. Their extreme vulnerability is a consequence of State policies and practices implemented over decades, steadily marginalising the Rohingya and eroding their enjoyment of human rights. The process of "othering" the Rohingya and their discriminatory treatment started long before the period covered by the [UN Fact-Finding] Mission.[35]

These instances of oppression have been documented in multiple forms of human rights reporting too numerous to review in detail here, and demonstrate a pattern of discrimination and gross violation of international human rights law.[36]

FACEBOOK ENTERS A FRAGILE DEMOCRACY FOLLOWING MILITARY RULE

Facebook's popularity in Myanmar rose in 2013 following the country's nascent transition from decades of military rule to democratic governance, and the corresponding deregulation of the telecommunications sector.[37] Prior to this shift in governance, only 1% of Myanmar's population used the internet, reflecting the country's lack of digital exposure and connectivity under an insular military rule.[38] Following deregulation, mobile access jumped to 56% in 2015, which supported the rapid adoption of Facebook.[39] As Facebook became the primary way for people to connect to the internet, it also became the newly connected country's primary source of news and information.

Despite overtures towards democracy including elections and increased foreign investment, the military retained a dominant role in politics and governance, and military officers continued to occupy positions of authority throughout all branches of government.[40] These facts are important to understand because even if Facebook appeared to introduce a democratic form of connectivity and information sharing, it was not in a context that supported media freedom nor freedom of expression and opinion; rather, the state retained significant control. This landscape included limited diversity in the ownership of online media outlets, and outlets that promoted state-sanctioned narratives while discrediting independent reporting.[41] These dynamics limited the growth of Myanmar's telecommunications market and attracted commercial interests as well. For example, a 2012 Deloitte report on the growth potential for Myanmar as one of the last undeveloped telecommunications markets in Asia noted:

> The potential in the telecommunications market is immense...there is a need to balance the desire for a speedy catch-up with the need to learn from other nations' experiences. The current environment is risky, with particular threats of civil unrest and corruption. Companies are keenly aware of Myanmar's regulatory ambiguity.[42]

In addition to the factors that limited free and informed media consumption, Myanmar continued to operate under a system of laws that was frequently used to arrest and silence journalists, human rights defenders, and others critical of the government, particularly those drawing attention to human rights violations against the Rohingya.[43] The use of such laws further limited the development of democratic space in Myanmar, with direct consequences for how information was derived and understood amongst communities, and in shared narratives across society. It demonstrates an atmosphere of repression combined with lack of strong watchdog institutions that could otherwise support independent journalism to counteract government control over media narratives.

As Facebook entered the market and quickly became "synonymous with the internet,"[44] suddenly a country without media pluralism or a free press was dominated by an algorithmically driven information ecosystem, characterized as it was by echo chambers, and coordinated misinformation and disinformation activity.

---

34. See generally, Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar."

35. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 110.

36. Fortify Rights, "Tools of Genocide," 9.

37. See, for example, Albert and Maizland, "The Rohingya Crisis."

38. Steve Stecklow, "Why Facebook is Losing the War on Hate Speech in Myanmar," *Reuters*, August 15, 2018, https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

39. See McLaughlin, "Facebook's Rise in Myanmar."

40. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 24.

41. Freedom House, "Freedom on the Net 2019: Myanmar," Freedom House, accessed February 9, 2021, https://freedomhouse.org/country/myanmar/freedom-net/2019.

42. Deloitte, "Myanmar: The Next Asian Telecommunications Greenfield?," Deloitte, 2013, https://www2.deloitte.com/content/dam/Deloitte/sg/Documents/technology-media-telecommunications/sg-tmt-Myanmar-next-Asian-telecommunications-greenfield.pdf.

43. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 18.

44. See, for example, Hannah Beech and Saw Nang, "In Myanmar, a Facebook Blackout Brings More Anger than a Genocide Charge," *New York Times*, August 31, 2018, https://www.nytimes.com/2018/08/31/world/asia/myanmar-genocide-facebook-military.html.

EARLY WARNING HOT SPOTS AND CIVIL SOCIETY RED FLAGS

When considered in a human rights context, a baseline assessment can indicate the range of risks and harms that may arise from introducing technological systems, particularly in fragile socio-political environments. For example, UN investigators found that the Myanmar government's propensity to control the information landscape and stifle human rights was supported by existing laws:

> These problematic laws constitute a veritable toolbox for State officials and representatives wishing to stifle dissent and evade legitimate scrutiny of their actions. It provides them with various options to curtail and punish the legitimate exercise of the rights to freedom of expression, association and peaceful assembly. Individuals who are critical of the government or security forces, and who expose or speak out against human rights violations, are particularly vulnerable. The use of such laws has curbed the democratic space in Myanmar.[45]

Had Facebook conducted a baseline assessment for potential human rights impacts prior to operating in Myanmar, such indicators would have served as an early warning. The potential human rights impacts derived from the legal, political, and information landscape, including the government's discrimination and marginalization of the Rohingya, would have been apparent from any human rights analysis of the market context prior to Facebook's entry.

While a baseline human rights assessment cannot definitively predict the social impact of a technological system on a population or geographic area, it can serve as an early warning of a human rights "hot spot." An assessment could determine if a transitional democracy like Myanmar, or an event like an election, could be designated as a hot spot that presents urgent human rights risks for any company contemplating entry into the market. Given the potential harms, Facebook conducting business as usual in such a hot spot amounts to reckless social experimentation on a national scale by a US-based technology company, which had neither expertise nor physical presence in Myanmar. A baseline assessment would have determined that Facebook's company mantra, "move fast and break things," could exacerbate human rights risks, and that both its AI products and business strategy should be reexamined for the Myanmar market. If a company determined nonetheless to move forward in a human rights hot spot, subsequent reports confirming the risks would be sufficient evidence to support pausing and redirecting the company's approach on human rights grounds.

To that point, for years following its entry into Myanmar, local civil society organizations raised red flags cautioning Facebook that its platform was being used to promote hatred against the Rohingya.[46] Civil society organizations on the ground can provide valuable insights and intelligence on technological impact at the granular community level, which company officials thousands of miles away could not possibly know.

In the Myanmar case, reports of civil society red flags were apparently heard and all but ignored by the company.[47] Civil society leaders communicated with company officials at length, even traveling to Facebook's California headquarters multiple times from 2015–2017. These leaders shared information about how Facebook was being used to foment violence, noting similarities to the way radio broadcasts were used to incite killings during the Rwandan genocide.[48] As the founder of a community tech hub in Yangon called Phandeeyar noted: "It couldn't have been presented to them more clearly, and they didn't take the necessary steps."[49] In an open letter written to Facebook CEO Mark Zuckerberg in April 2019, civil society leaders again reiterated specific instances that linked messages on the platform to real-world harms, similar to the kind they had been flagging for years:

> The messages...were clear examples of your tools being used to incite real harm. Far from being stopped, they spread in an unprecedented way, reaching country-wide and causing widespread fear and at least three violent incidents in the process.[50]

HRIA OF FACEBOOK'S PRESENCE IN MYANMAR

In the UN Report, investigators contextualized the relationship between Facebook's platform and the dynamics of Myanmar to demonstrate how the platform was positioned to spread hate speech and misinformation, and thus incite violence and hostility. The UN Report called for further independent investigation of the extent to which communication on Facebook led to violence in Myanmar, positioning it as a case study for the company's global operations to prevent similar potential harm from occurring again. The UN Report concluded by recommending that Facebook focus on the relationship between its platform and the human rights context of a new market prior to entering that market:

> Before entering any new market, particularly those with volatile ethnic, religious or other social tensions, Facebook and other social media platforms...should conduct in-depth human rights impact assessments for their products policies and operations, based on national context.[51]

45. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 318.

46. Stecklow, "Why Facebook is Losing the War on Hate Speech in Myanmar."

47. This paper is informed by interviews in Myanmar with civil society organizations conducted by Mark Latonero.

48. McLaughlin, "Facebook's Rise in Myanmar."

49. Stecklow, "Why Facebook is Losing the War on Hate Speech in Myanmar."

50. Letter from Civil Society Organizations in Myanmar to Mark Zuckerberg, April 5, 2018, https://assets.documentcloud.org/documents/4432469/Myanmar-Open-Letter-to-Mark-Zuckerberg.pdf; For an example of the types of messages spread through Facebook and used to incite violence, as highlighted by the letter, see https://www.documentcloud.org/documents/4432469-Myanmar-Open-Letter-to-Mark-Zuckerberg.html.

51. Independent International Fact-Finding Mission on Myanmar, "Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar," 429.

Facebook commissioned the non-profit organization Business for Social Responsibility (BSR) to undertake an HRIA of the company's presence in Myanmar, which was publicly and transparently released in its entirety in October 2018.[52] This HRIA appeared to follow a standard methodology, including direct stakeholder consultation in Myanmar and an analysis of human rights impacts for rights holders. The assessment attributed a level of responsibility to Facebook based on the "cause, contribute to and directly linked" framework set forth in the UN Guiding Principles.

However, this HRIA failed in terms of assessing the actual and potential human rights impact of Facebook's presence and product in Myanmar in two critical ways. First, the HRIA does not discuss the role of Facebook's News Feed algorithm and its human rights impacts. At a high level, the News Feed is an algorithmic system that orders posts and news stories shared by the people and pages within a user's Facebook community.[53] News Feed reportedly uses AI and machine learning algorithms to process the large amounts of behavioral data collected about users, and delivers a personalized content feed that is optimized for user engagement.[54] One method of optimizing for engagement is to prioritize posts that the algorithm determines are most likely to captivate user attention and promote engagement in the form of likes, comments, shares, and time spent with the content. This form of optimization supports the company's advertising business, but it can also amplify sensationalized misinformation and disinformation.[55]

Second, although the HRIA reviews certain parts of the history of Myanmar, it does not surface how key elements of this context—like the discrimination against, and oppression of the Rohingya, and the country's status as a fragile and emerging democracy—were related to decisions to deploy and operate Facebook's products like News Feed in Myanmar. In fact, the HRIA does not at all mention the decades-long history of systematic and comprehensive human rights violations committed against the Rohingya by the government of Myanmar. Although the history, treatment, and subsequent genocide of the Rohingya are discussed in detail throughout the UN Report, the Facebook-commissioned HRIA does not evaluate how this specific ethnic tension and history of oppression set the baseline conditions upon which its platform would be used. A breadth of human rights reporting, including the UN report on genocide, has revealed that Myanmar continued to impose escalating restrictions on the Rohingya—including forced labor, land confiscation, forced eviction, extortion and arbitrary taxation, and restrictions on movement, marriage, employment, health care, and education—that constitute violations of their human rights and fundamental freedoms.[56] This reflects systemic discrimination in violation of the Universal Declaration of Human Rights and primary human rights treaties.[57] The UN Guiding Principles propose that companies should prioritize addressing the most severe human rights impacts. The treatment of the Rohingya should be the first place that any HRIA in Myanmar begins.

The lack of analysis of the inherent social and technical factors of Facebook's presence in Myanmar led to a miscalculation in the HRIA's findings. Any such HRIA would fall short as an instrument of human rights accountability, without the interdisciplinary expertise needed to assess the impact of the AI and algorithmic system itself in relation to the chronic plight of the Rohingya—the most urgent human rights concern in Myanmar. For example, the HRIA findings overemphasize the social, cultural, and political environment as a potential justification for why Facebook was misused. This shifts the responsibility to others, such as the people of Myanmar for not following Facebook's community standards of behavior, rather than emphasizing Facebook's responsibility to adapt its AI products and business model to the human rights risks from the moment the company decided to operate in Myanmar.

Ultimately, the HRIA concluded that Facebook did not cause or contribute to any human rights harms in Myanmar. When it came to the right to nondiscrimination, the Rohingya were not mentioned. Instead, the assessment said that Facebook had leverage to affect discrimination on the platform through "implementation of the Community Standards, efforts to increase digital literacy, and raising awareness of the Community Standards." These findings reflect neither the most salient and serious impact, nor the impact of the algorithmic system and business model. As John Ruggie states:

> When can we say that a company like Facebook is 'contributing to' human rights harm?... Unwittingly getting even severely consequential cases wrong once or twice is one thing. But persistent refusal to substantially change what the company does to reduce its role in others' promotion of social strife and violence makes the attribution of 'contribution' inescapable.[58]

52. Business for Social Responsibility (BSR), "Human Rights Impact Assessment: Facebook in Myanmar," Business for Social Responsibility, 2018, https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf.

53. Facebook Careers, "The Impact of Machine Learning at Facebook: Community, Integrity, and Innovation," *Facebook Careers* (blog), March 27, 2019, https://www.facebook.com/careers/life/the-impact-of-machine-learning-at-facebook-community-integrity-and-innovation.

54. Will Ormeus, "Who Controls Your Facebook Feed," *Slate*, January 3, 2016, http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html.

55. See Amnesty International, "Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights," Amnesty International, November 2019, https://www.amnesty.org/download/Documents/POL3014042019ENGLISH.PDF; Joan Donovan, "Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech," Data & Society Research Institute, October 17, 2018, https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf; Julia Hass, "Freedom of the Media and Artificial Intelligence," Office of the OSCE Representative on Freedom of the Media, November 16, 2020, 3, https://www.international.gc.ca/campaign-campagne/assets/pdfs/media_freedom-liberte_presse-2020/policy_paper-documents_orientation-ai-ia-en.pdf.

56. Benjamin Zawacki, "Defining Myanmar's 'Rohingya Problem,'" *Human Rights Brief*, vol. 20, no. 3 (2013): 18–25, https://digitalcommons.wcl.american.edu/hrbrief/vol20/iss3/2.

57. See United Nations General Assembly, "Universal Declaration on Human Rights," United Nations, December 10, 1948, https://www.un.org/en/universal-declaration-human-rights; and United Nations General Assembly, "International Covenant on Civil and Political Rights," United Nations Office of the High Commissioner for Human Rights, March 23, 1976, https://www.ohchr.org/Documents/ProfessionalInterest/ccpr.pdf.

58. John G. Ruggie, "Facebook in the Rest of the World," John F. Kennedy School of Government, Harvard University, November 15, 2018, https://media.business-humanrights.org/media/documents/files/documents/John_Ruggie_Facebook_15_Nov_2018.pdf.

Developing Effective HRIAs for AI

The HRIA of Facebook in Myanmar illustrates several ways the existing methodology should be updated if it is to be used on AI systems. When an HRIA is promoted as a thorough assessment but fails to evaluate the relationship between the AI systems and the most salient human rights harms, it risks creating a shield for companies to claim that they have prioritized their responsibility to respect human rights. HRIAs that do not focus on this relationship risk providing recommendations for companies that amount to tinkering around the margins rather than addressing core problems.

For example, when Facebook published the HRIA of its presence in Myanmar in 2018, it issued an official statement acknowledging, "[w]e know we need to do more to ensure we are a force for good in Myanmar."[59] In its blog post summarizing the HRIA, Facebook notes several steps that it took in response to BSR's findings, which included establishing a separate policy for content moderation with respect to human rights, and increasing team capacity to focus on issues specific to Myanmar, including adding Burmese speaking content reviewers."[60] Yet, none of these changes reflected a recognition or reexamination of the role of Facebook's AI systems, products, or business model in the discrimination and genocide against the Rohingya, as well as other rights harms.

To effectively center the impact of AI and algorithmic systems in human rights contexts, HRIAs should be seen as an analysis of a *sociotechnical* system wherein social and technical factors are inherently intertwined and interrelated. Both computer and social scientists with expertise on the social impact of AI and algorithmic systems would buttress the work of human rights experts conducting HRIAs. Determining whether or not an AI system contributed to a human rights harm is not obvious to those without the appropriate expertise and methodologies.[61] Furthermore, without additional technical expertise, those conducting HRIAs would not be able to recommend potential changes to AI products and algorithmic processes themselves in order to mitigate existing and future harms.

The optimal team conducting an HRIA of an AI system would reflect the interdisciplinary expertise needed to determine the appropriate methods and criteria for assessing an actual or potential impact from AI and algorithmic systems in specific contexts. Creating a dialogue between experts to establish standardized knowledge and methodologies will be essential in order to conduct effective HRIAs for AI.

The precise steps such a team would take when embarking on an HRIA will depend on contextual factors such as the type of AI system, product, market, and domain (such as policing, hiring, or health care). These factors surface the nature and magnitude of salient human rights harms and, thereby, the thresholds for demonstrating algorithmic impact that should be considered by an HRIA. The following section will discuss recommendations like timing ongoing HRIAs and situating HRIAs within a due diligence process. Examples of new criteria for HRIAs for AI, such as assessing algorithmic discrimination, may be particularly useful for identifying human rights risks and harms to vulnerable and marginalized groups. These examples are not exhaustive. Research on best practices for evaluating algorithmic accountability [62] and AI system performance continues to evolve, and its relevance to human rights are only beginning to be considered by the field.

TIMING HRIAS FOR AI

In addition to expertise and coordination, for HRIAs to prove useful in the AI space they must be conducted at appropriate times relative to critical stages in an AI development lifecycle. In 2018, the research organization AI Now proposed an algorithmic impact assessment (AIA) framework focused on AI systems being used by public agencies.[63] They recommended AIAs that draw on "growing and important research that scientific and policy experts have been developing on the topic of algorithmic accountability" to help us better understand AI systems prior to deployment, as well as on an on-going basis where they have been previously sanctioned for deployment and are actively in use.[64]

59. Warofka, "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar." Facebook originally published the entire findings of this HRIA on its blog in November 2018, yet the link to the assessment was broken of December 2020 and can be found here: https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf. The original post was updated August 26, 2020 to note that Facebook was working with the UN Independent Investigative Mechanism for Myanmar (IIMM) to provide data regarding its investigation into "international crimes in Myanmar." As of this date, the government of Myanmar is facing charges of genocide in the International Court of Justice regarding the violence perpetuated against the Rohingya in 2017. See also Poppy McPherson, "Facebook Shares Data on Myanmar with United Nations Investigators," *Reuters*, August 25, 2020, https://www.reuters.com/article/us-myanmar-facebook/facebook-shares-data-on-myanmar-with-united-nations-investigators-idUSKBN25L2G4.

60. Warofka, "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar."

61. Recommendations for additional expertise are derived from the UN Guiding Principles, e.g. see United Nations Office of the High Commissioner for Human Rights, "Guiding Principles," Principles 18, 20.

62. We use "algorithmic accountability" as a general term that includes social, legal, computational, and philosophical research on AI, machine learning, and algorithmic systems in areas such as fairness, accountability, transparency, explainability, bias, and discrimination. See generally, "ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)," Association for Computing Machinery, accessed February 9, 2021, https://facctconference.org.

63. Dillon Reisman et al., "Algorithmic Impact Assessments: A Practical Accountability Framework for Public Agency Accountability," AI Now Institute, April 2018, https://ainowinstitute.org/aiareport2018.pdf; For an critical examination of AIAs see Emanuel Moss et al., "Governing with Algorithmic Impact Assessments: Six Observations," *SSRN*, April 24, 2020, https://ssrn.com/abstract=3584818.

64. Reisman et al., "Algorithmic Impact Assessments: A Practical Accountability Framework for Public Agency Accountability," 5–6.

In a report on the implications of AI for human rights in the information environment, the former UN Special Rapporteur on the Right to Freedom of Expression and Opinion, David Kaye, extended AI Now's recommendation to the private sector, stating that companies should also conduct assessments that leverage research on algorithmic accountability.[65] Kaye added that an HRIA of any given AI system should be performed by a company prior to the procurement or development of the AI, depending on where the company sits in the lifecycle of the system.[66] This prescription refers to an ex-ante or forward-looking evaluation of what is likely to occur. Conducting an ex-ante evaluation attempts to predict the future relationship between rights and an on-going or proposed business activity. On a standalone basis, this can be distinguished from an ex-post approach, which is when a company undertakes an HRIA to evaluate the relationship between a past or on-going business activity and its human rights impact. Facebook's 2018 HRIA of its presence in Myanmar is an example of a one-time ex-post approach.

LIMITS TO ONE-TIME EX-POST & EX-ANTE ASSESSMENTS

Rather than looking to one moment in time, effective HRIAs for AI should move beyond standalone assessments. The limitations of looking forward or back at a single moment in time are too significant for a standalone approach to be effective when it comes to addressing the human rights concerns posed by AI. These systems are running in real time, are frequently updated, use datasets that change over time, and are supported by machine learning algorithms and dynamic predictive models. As AI systems change course, so too does their potential relationship to human rights.

A major criticism of HRIAs is that they do not function to identify and address human rights impacts over time and become a compliance exercise.[67] If HRIAs are conducted as backward-looking standalone assessments, they will remain ineffective. Looking back at what happened can be important in stabilizing knowledge about the relationship between specific AI systems and human rights, as well as establishing a basis for the remedy of harms. However, ex-post HRIAs do not address the potential for AI systems to cause harm to vulnerable populations in the near term or in real time. In other words, conducting an HRIA after the genocide in Myanmar is far too little and too late for the Rohingya.

Conversely, although forecasting impact prior to deployment is a necessary component of the HRIA process, it is also insufficient as a standalone exercise. Once deployed, technological systems can become integrated into business activities, organizational practices, policies, and consumers' daily lives, which narrows the options for course corrections that reverse both engineering and business decisions. Using an assessment to ask whether a technology should be designed or deployed in the first place highlights the stakes of being able to identify and intervene in problematic systems before they become entrenched and politically constrained.

There is a limited frame of reference as to the foreseeable impacts of introducing AI systems. However robust the ex-ante methodology is at predicting potential human rights harms, evaluating an AI system at a single moment in time prior to deployment is unlikely to capture the evolving socio-technical system, which could lead to unforeseen consequences.

ESTABLISHING ONGOING HRIAS IN A DUE DILIGENCE PROCESS

Whether ex-ante or ex-post, time-bound HRIAs have notable limitations that are amplified in the context of AI. In order to identify and respond to changes in the relationship between AI and its operating context, companies should implement a dynamic HRIA process that functions on an ongoing basis as part of a comprehensive human rights due diligence process. A report by the Business Tech Project (B-Tech Project) at UN Human Rights emphasizes the importance of integrating HRIAs into a comprehensive human rights due diligence process effort that is "organizationally embedded" throughout the company.[68] The report notes that the continued failure to integrate HRIAs into this larger process can pose a number of risks that are counterproductive to supporting business respect for human rights, including "reducing organization-wide ownership of the human rights risks identified, leading to lack of meaningful action."[69]

Implementing findings and recommendations from an HRIA through complex organizational structures, like AI technology companies, would require multiple levels of internal teams coordinated around a respect for human rights. To demonstrate a meaningful commitment, due diligence needs to be supported by corporate decision-making from the top down to ensure that it is being taken seriously through all functions of the business. Ongoing HRIAs in the context of AI reflect an essential but single component of that larger process. The UN Human Rights B-Tech Project published recommendations for how technology companies should implement a comprehensive approach to human rights due diligence, including best practices to help ensure that HRIA findings are integrated throughout corporate practice.[70] Due diligence helps ensure that companies have the appropriate governance structures in place to identify and prioritize the most salient human rights issues in relation to the product and market. Without coordination from the top to promote alignment across teams that might otherwise operate independently—such as engineering, policy, and marketing teams—companies run the risk of failing to implement a comprehensive set of actions that are necessary to redirect AI development and deployment to mitigate harms.

65. Davis Kaye, "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression," United Nations, August 29, 2018, 18, https://www.undocs.org/A/73/348.

66. Kaye, "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression," 18, 20.

67. For a critique of HRIAs in tech, see United Nations Office of the High Commission for Human Rights, "Key Characteristics of Business Respect for Human Rights: A B-Tech Foundational Paper," United Nations Office of the High Commissioner for Human Rights, 2020, https://www.ohchr.org/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf.

68. United Nations Office of the High Commission for Human Rights, "Key Characteristics of Business Respect for Human Rights."

69. United Nations Office of the High Commission for Human Rights, "Key Characteristics of Business Respect for Human Rights."

70. United Nations Office of the High Commission for Human Rights, "Key Characteristics of Business Respect for Human Rights."

The ongoing approach is found in the UN Guiding Principles, per the recommendation that impact assessment should be undertaken at regular intervals due to the fact that human rights situations are dynamic.[71] The UN Guiding Principles suggest several moments that could be defined to mark the appropriate intervals for ongoing assessment, including anticipated changes in the operation or operating environment.[72] These intervals for conducting HRIAs should be framed in relation to the lifecycle of AI systems, taking into consideration the design, development, and deployment stages, as well as sale, use, integration, and maintenance.

A critical component for an ongoing assessment process is a baseline determination of the human rights status quo in a market prior to a company introducing its AI system.[73] Establishing the status quo is necessary to evaluate how an AI system may alter the human rights situation. As an assessment identifies potential impacts, both human rights and technical experts can aid in risk mitigation actions that involve a change to product design, strategy, or approach. Setting a baseline establishes a meaningful scope of responsibility for subsequent impacts. Baseline assessments help companies develop risk mitigation strategies and levels of responsibility relative to human rights situations as they are found.

Creating a baseline picture of human rights will require tapping into the expertise of civil society organizations and individuals who have on-the-ground knowledge of vulnerable groups and local sociopolitical tensions. Determining the picture of any given human rights context will require establishing *indicators* that can be tracked and measured over time. Establishing indicators is no small task, and requires interdisciplinary expertise to examine measurable sociotechnical relationships and categories that are context specific.

IDENTIFYING DISCRIMINATION AS AN ALGORITHMIC HARM

In addition to these recommendations on *timing*, HRIAs for AI should engage with emerging criteria for evaluating algorithmic decision-making, which should prioritize the most salient potential human rights concerns. The UN Guiding Principles require companies to prioritize the most severe human rights impacts that they face in the course of their operations.[74] For example, if the underlying human rights context reveals serious patterns of abuse or decades of systematic discrimination against an ethnic minority, algorithmic assessment should prioritize the understanding of disparate impacts and how the system might be weaponized in the service of further oppression and control.

Non-discrimination is a fundamental right that is indivisible and interrelated to all other human rights.[75] Human rights treaties emphasize that the rights enunciated must be respected without discrimination of any kind.[76] Non-discrimination is so important within the human rights system that it has formed the basis for treaties that focus specifically on protecting the rights of vulnerable or historically disadvantaged groups, for example, those based on gender, age, disability, or migrant worker status.[77] In the business context, the UN Guiding Principles reinforce that enterprises "should pay special attention to any particular human rights impacts on individuals from groups or populations that may be at heightened risk of vulnerability or marginalization."[78]

In the AI governance space, discrimination is a paramount concern and presents additional challenges. In general, discrimination can be challenging to detect, particularly in indirect situations in which a seemingly neutral provision or criteria disproportionately disadvantages a particular group. Indirect discrimination can be difficult to identify because it extends beyond the idea of formal equality to reveal how the universal application of rules can reinforce underlying social inequalities and thereby perpetuate discrimination.

AI systems introduce new risks of widespread yet subtle forms of discrimination as an "unintentional emergent property of the algorithm's use rather than a conscious choice by programmers."[79] These systems may discriminate against groups that are not defined by historically or legally protected characteristics (i.e., strictly along the lines of race, gender, age, and sexual orientation).[80] These new patterns of disparity may require human rights stakeholders and policymakers to redefine the scope of non-discrimination norms.

71. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights," Principle 18.

72. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights."

73. Raso at al., "Artificial Intelligence & Human Rights."

74. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights," Principle 18.

75. For example, Article 26 of the ICCPR provides that, "All persons are equal before the law and are entitled without any discrimination to the equal protection of the law…," while Article 24 of the ICCPR provides for rights of a child "without any discrimination as to race, colour, sex, language, religion, national or social origin, property or birth. See United Nations General Assembly, "International Covenant on Civil and Political Rights," Articles 24 and 26.

76. See, for example, United Nations General Assembly, "International Covenant on Economic, Social and Cultural Rights," United Nations Office of the High Commissioner for Human Rights, December 16, 1966, https://www.ohchr.org/en/professionalinterest/pages/cescr.aspx; United Nations General Assembly, "International Covenant on Civil and Political Rights."

77. See the core human rights instruments here: "The Core International Human Rights Instruments and their Monitoring Bodies," United Nations Office of the High Commissioner for Human Rights, accessed February 9, 2021, https://www.ohchr.org/EN/ProfessionalInterest/Pages/CoreInstruments.aspx.

78. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights," Principle 18.

79. Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* vol. 104 (2016): 671–732, http://dx.doi.org/10.2139/ssrn.2477899.

80. Title VII of the Civil Rights Act that prohibits employment discrimination based on protected characteristics including race, color, religion, sex, national origin, age, disability, or genetic information (which can be read as an expression of the U.S. constitutional guarantee of "equal protection" under the fourteenth amendment). Civil Rights Act of 1964 § 7, 42 U.S.C. § 2000e et seq (1964).

For example, 2018 research on commercial gender classification algorithms using facial recognition technology designed by IBM, Microsoft, and Face++ found substantial differences in error rates between different groups when accounting for both gender and skin tone.[81] While, overall, the systems tested performed better on lighter versus darker skin and men versus women, the error rates compounded significantly when considering the performance of the algorithm on the subgroup of darker skinned women.

These findings demonstrate how, in some contexts, it will be important for HRIAs to consider multiple characteristics that define subpopulations potentially related to identifying algorithmic discrimination. In other words, we might ask, what characteristics do we use to define a marginalized group for the purposes of evaluating discrimination? For the purposes of evaluating human rights impacts related to algorithms, groups may be identified based on a combination of several characteristics, such as age, gender, ethnicity, race, sexual orientation, disability, or refugee status. Evaluating the severity of impacts on different subgroups may reveal harms amounting to a human rights hot spot that would not have otherwise been apparent without analyzing the AI system itself. For example, while an algorithm might not generally demonstrate a statistical difference between members of two ethnic groups, this finding could change if you further define the groups by considering children. Given the particular considerations of children's rights, these performance differences could warrant additional precautionary measures if experts determine that rights are potentially implicated. An intersectional framework would help avoid narrow or binary definitions of groups when investigating harms arising from hidden patterns of algorithmic discrimination.

ASSESSING ALGORITHMIC DECISION-MAKING

HRIAs of AI presuppose, to some degree, that the relationship between an AI system and a human rights impact can be known and understood. Yet, AI and algorithmic systems are often seen as "black-box" technologies that defy our ability to know precisely how they work and how they make "decisions" that may impact human rights. HRIAs and human rights due diligence do not require absolute transparency and perfect knowledge. If causation cannot be determined—which would be exceedingly difficult in an AI context—an assessment can find that a business product or process *contributed to* a human rights harm or was *directly linked* to a harm through a business relationship.

Each of these assessments would attribute responsibility and initiate a course of corrective action. Nevertheless, black-box AI systems may challenge the ability to assess impact, which, in turn, can problematize assigning responsibility and ultimately accountability.

Looking at the decisions and processes behind AI systems can reveal the values that shape its design, which will often have direct implications for human rights-based concerns. Algorithms are not intuitive—meaning, they operate at a scale and level of complexity that is beyond human reasoning.[82] As a recent paper notes: "Intuition is poorly suited to detect automated discrimination that does not resemble human discrimination, and can be altogether more subtle, widespread and based on complex patterns and correlations perceived to exist between people."[83] An AI system will not readily reveal the reasoning behind its decision-making. Hence, effective HRIAs for AI will benefit from company policies implemented under a comprehensive approach to human rights due diligence that require AI designers and developers to "show their work" through documentation.[84]

Those conducting HRIAs will benefit from reviewing such documentation to understand developers' subjective decisions throughout the process of designing and deploying an AI system. For example, documentation of the types of data used to train an AI model might reveal that personal identifiable information on religion or ethnicity was included in the dataset, which may be contributing to unintended discriminatory effects. Establishing internal policies that require documentation and make it available for HRIAs would dovetail with the UN Guiding Principles directive that companies "know and show that they respect human rights" by putting such policies and practices in place.[85]

The need for documentation has also been explored by AI researchers for internal company audits of their systems to address or mitigate algorithmic risk and harms.[86] If human rights were considered the standard to guide and govern the development and deployment of AI systems, internal algorithmic audits could dovetail with HRIAs as tools that could help companies assess human rights alignment throughout the AI lifecycle.

Algorithms can also be audited externally. A number of academics have examined the issue of auditing algorithms that may be impacting social harms when a company is unwilling to provide necessary access or transparency.[87] "Auditability" is seen as a key

81. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* vol. 81, (2018): 1–15, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

82. Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI," *SSRN*, March 3, 2020, https://dx.doi.org/10.2139/ssrn.3547922.

83. Wachter, Brent Mittelstadt, and Chris Russell, "Why Fairness Cannot Be Automated," 48.

84. Andrew Selbst and Solon Barocas, "The Intuitive Appeal of Explainable Machines," *Fordham Law Review* vol. 87 (2018): 1085-1139, 1138, https://ssrn.com/abstract=3126971.

85. United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights," 16.

86. Inioluwa Deborah Raji et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," *FAT* *'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020, 33–44, https://doi.org/10.1145/3351095.3372873.

87. Christian Sandvig et al. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms" (presentation, *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*," preconference at the 64th Annual Meeting of the International Communication Association, Seattle, WA, May 22, 2014); Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (Cambridge, MA: Harvard University Press, 2015); Latanya Sweeney, "Discrimination in Online Ad Delivery," *Communications of the ACM* vol. 56, no. 5 (May 2013): 44–54, https://arxiv.org/abs/1301.6822.

principle for accountable algorithms,[88] although it is often the case that AI companies resist allowing external researchers to access or examine their systems and products, citing trade secret concerns or fearing adverse publicity.[89] "Auditing algorithms from the outside"[90] is seen as a potential way for civil society actors with the requisite expertise to investigate commercial AI products and publicize their findings to incentivize private sector action.

We do not know if HRIAs for AI could function like an external audit of algorithms. On one level, conducting effective ongoing HRIAs for AI as a tool of human rights due diligence assumes adequate resources and political will from company leadership in order to embed a respect for human rights throughout the organization. However, given the urgency of potential human rights harms, it is important to note that these recommendations can also be taken up by civil society actors to conduct HRIAs without needing a company's buy-in or support. Techniques, best practices, and funding for HRIAs for AI outside of corporate due diligence are an open area for research and development.

INCENTIVIZING ASSESSMENTS AND DILIGENCE THROUGH MANDATORY MEASURES

As more recommendations for effective HRIA methodology for AI emerge, issues with enforcement and accountability also arise. While companies may choose to voluntarily comply with human rights due diligence, governments can enforce human rights standards through regulation. With the UN Guiding Principles, checks on business are not intended to take on a mandatory character until they have been developed into norms crystallized by law. The UN Guiding Principles provide a cross-functional framework under which companies, governments, and other stakeholders could work together to develop norms that address the challenges of business and human rights. Consider the following summary from John Ruggie in a December 2019 keynote address:

> [The UN Guiding Principles] were intended to help generate a new regulatory dynamic, one in which public and private governance systems, corporate as well as civil, each come to add distinct value, compensate for one another's

weaknesses, and play mutually reinforcing roles—out of which a more comprehensive and effective global regime might evolve.[91]

While the UN Guiding Principles discuss self-regulatory action by companies, it also highlights that a "smart mix" of both voluntary and mandatory measures should be considered by states to foster businesses' responsibility to respect human rights.[92]

A number of states, and in particular the EU, have begun to implement mandatory human rights due diligence. These initiatives are focused on a range of business activities and assessment criteria, and have gained momentum, even from some investors and companies seeking a legal foundation that supports business respect for human rights.[93] In the same address referenced above, Ruggie makes two additional points regarding efforts underway to implement mandatory measures. First, such initiatives are an important step towards adding law and regulation into the "smart mix" of measures to protect against human rights harm and enable business respect for human rights. Second, more should be done to *specify* meaningful implementation "in order to avoid contributing to the proliferation of self-defined standards and storytelling by firms."[94]

Conclusion & Considerations for Stakeholders

The case of Facebook in the context of genocide in Myanmar demonstrates the value of using a human rights lens to address the most consequential risks and impacts of AI and algorithmic systems. Fragile human rights contexts represent high-risk and high-stakes domains where tech companies have a responsibility and duty of care to respect the rights of people who may be impacted by their systems and products. Human rights impact assessments represent a key tool for identifying potential risks, informing mitigation strategies, and providing a basis for remedy for those who have been harmed. Yet, applying the usual approaches and methodology to AI systems is a recipe for failure, as with the HRIA of Facebook in Myanmar.

88. Nicholas Diakopoulos et al, "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," Fairness, Accountability, and Transparency in Machine Learning, accessed February 9, 2021, https://www.fatml.org/resources/principles-for-accountable-algorithms.

89. For example, see Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System," *Stanford Law Review* vol. 70 (2018): 1343-1429, http://dx.doi.org/10.2139/ssrn.2920883.

90. See Mike Ananny et al., "Auditing Algorithms from the Outside: Methods and Implications," (workshop, International Conference on Web and Social Media (ICWSM), 2015), https://auditingalgorithms.wordpress.com/rationale/.

91. John G. Ruggie, "Conference on Business and Human Rights: Towards a Common Agenda for Action," (keynote address, Conference on Business and Human Rights: Towards a Common Agenda for Action, Brussels, Belgium, December 2, 2019), https://shiftproject.org/resource/john-ruggie-keynote-finland2019/.

92. See United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights," 5.

93. For example, a coalition of 70+ investors recently signed a memorandum urging states to mandate due diligence as "materially good for business, investors and the economy." See Investor Alliance for Human Rights, "The Investor Case of Mandatory Human Rights Due Diligence," Investor Alliance for Human Rights, April 21 2020, https://investorsforhumanrights.org/sites/default/files/attachments/2019-12/The%20Investor%20Case%20for%20mHRDD%20-%20FINAL%20for%2011.25%20launch.pdf; In 2017, France passed a corporate duty of vigilance law that mandates major companies conducting business in France to implement human rights due diligence that identifies and prevents adverse impacts resulting from their own activities or the activities of companies with which they have a business relationship. As of 2020, President of the EU Council, Finland, has also put forward a priority agenda on business and human rights that calls for EU-wide "regulation on mandatory human rights due diligence." See, for example, Phil Dawson, "Closing the Human Rights Gap in AI Governance," *Element AI*, November 2019, https://www.elementai.com/news/2019/supporting-rights-respecting-ai.

94. Ruggie, "Conference on Business and Human Rights."

In order to apply the learnings from Facebook in Myanmar, HRIAs for AI should start to include an emerging set of structural qualities. HRIAs of AI should undertake a sociotechnical approach that examines the inextricable relationships between AI systems and societal context defined by the most salient human rights concerns. Assessments should examine AI systems and their impact on human rights, as well the business models and decisions that guide how AI products are designed and used. HRIAs should be conducted on an ongoing basis at key times in the AI product lifecycle, and as part of a company's embedded human rights due diligence process. HRIAs should involve interdisciplinary experts to employ evaluative criteria and methodologies emerging from research, for example, on algorithmic accountability and algorithmic discrimination.

This is but the first step if HRIAs are to be used by companies or other organizations to legitimately assess the potential and actual human rights impact of AI systems. Only through continued practice and refinement can the field build the standardized knowledge and methodology necessary for HRIAs to uniquely identify and mitigate risk and harms arising from AI, particularly for vulnerable or marginalized people. Of course, HRIAs can also assess potential *positive* impacts of AI systems and products on the enjoyment of human rights, which can be an opportunity for companies and organizations to demonstrate how they could provide a public good.

In a speech after visiting Silicon Valley tech companies, the UN High Commissioner for Human Rights, Michelle Bachelet, stated: "We cannot expect Big Tech to self-regulate effectively, nor do I believe we would want them to. The onus is on both technology businesses and governments—and also civil society—to work together to identify effective and equitable policies."[95] A mix of voluntary actions and mandatory measures may be needed to incentivize organizations to incorporate HRIAs for AI and algorithmic systems in a more effective, transparent, and accountable way. In the meantime, there are a number of actions stakeholders should consider in the near term and in order to move the discussion forward.

TECH COMPANIES

Companies should expand their engagement with experts who possess the skills required to assess baseline contexts—including identifying high-risk human rights hotspots—and the impact and performance of AI as a sociotechnical system. Companies should explore models for continuous and *equitable* engagement with independent civil society experts on the ground, since they are better positioned to provide an early warning of AI impacts on vulnerable and marginalized populations.

AI company leaders should express publicly, and through corporate policy, their commitment to embed a respect for human rights in order to marshal resources and political will. While potential government regulation may require HRIAs, they should not be seen as a compliance exercise that induces complacency or one that can be used as a performative shield that does not result in necessary change. Whether AI risks and harms are identified by the company or civil society, company leaders must

be committed to integrating those findings through corporate decision-making structures. In order for the assessment process to be effective, companies need policies and mechanisms for both immediate course correction and remedy for those whose rights have already been harmed. As discussed throughout this paper, timing is of the essence when it comes to upholding human rights throughout the development of AI systems.

Mature companies with existing units focused on responsible or ethical AI should explore how HRIAs and human rights due diligence can be integrated in their development of responsible AI systems, products, and policies. For early-stage companies, founders and venture capital firms have a distinct opportunity to use HRIAs and due diligence to proactively address human rights issues prior to ideation, development, and deployment. While early-stage or smaller companies are often resource constrained and lack capacity to engage with human rights, they may have less barriers to implementing systemic changes that mitigate potential harms, particularly before their technologies have become widely integrated into regular usage.

Finally, companies should consider how to increase transparency of both successes and failures in conducting HRIAs for AI in order to build the needed knowledge base and methodologies to improve their own HRIA practices and those across the tech sector.

GOVERNMENT

Under the UN Guiding Principles, nation-states are called on to take action that protects against human rights harms by businesses or other organizations. State action is meant to complement and reinforce the voluntary measures recommended for private firms. It is important for legislative measures to be implemented with some level of specificity in order to provide a structure beyond self-defined standards.[96]

For example, governments developing regulation on AI and algorithmic systems can consider mandatory measures for human rights due diligence, which can include timing requirements for HRIAs. Regular HRIAs could help monitor human rights impacts as they develop in response to changes on the ground and should occur at regular intervals in the lifecycle of AI systems and throughout the design, development, deployment, and use stages. Additional requirements could be considered for companies that sell their AI technologies for clients, including government use. While legislators should consider the appropriate methods for effective HRIAs for AI, they should be wary of the unintended consequences. Any proposed regulations for HRIAs for AI should be made with the input from experts who are fluent with emerging methods, criteria, and best practices for evaluating AI system performance and accountability in the interest of human rights.

In addition to company regulation, policy and decision makers in government can apply lessons from the UN Guiding Principles, and the business and human rights field, internally. States can require ongoing HRIAs and due diligence processes on the part of government agencies and intergovernmental organizations developing and using AI technologies, as well as those procuring AI products from tech companies.

95. Michelle Bachelet, "'Smart Mix' of Measures Needed to Regulate New Technologies," United Nations Office of the High Commissioner for Human Rights, April 24, 2019, https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=24509.

96. See, generally, Ruggie, "Conference on Business and Human Rights."

CIVIL SOCIETY

Civil society should consider expanding its role in HRIAs for AI. While limited funding and capacity is a perennial problem, civil society organizations should advocate for more equitable models of tech company engagement as well as new philanthropic funding. Civil society organizations should also explore the strategic value (e.g. for advocacy) of conducting independent HRIAs of AI systems initiated externally when companies refuse or are unable to do so internally. Civil society organizations can request that companies be more transparent and publicly accountable for the HRIAs for AI that they conduct, particularly if those assessments were conducted by other civil society/non-profit organizations and involved consultations with local rights holders. Civil society organizations providing direct services to local communities can also help determine meaningful remedies for people who have been harmed by AI systems. Civil society actors, including international organizations, who are developing and using AI technologies for their programs can also implement HRIAs and due diligence processes internally.

Academic researchers would benefit from access to HRIA for AI policies, practices, and outputs in order to analyze the efficacy of HRIAs in relation to other AI and algorithmic impact assessment approaches and methodologies. Researchers can also analyze and improve the organizational dynamics of legal, human rights, social science, business, and computer science experts, who need to work together to assess the human rights impact of AI systems.

Carr Center Discussion Paper

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