The Use of Artificial Intelligence in International Human Rights Law

Anne Dulka*

26 STAN. TECH. L. REV. 316 (2023)

ABSTRACT

Over the last decade, AI has become an increasingly important tool in the enforcement of international human rights law. This Note provides a comprehensive overview of the implementation and application of AI technologies in the international human rights law space with a particular focus on how AI is being used to track and report on the enjoyment or violation of human rights. From the use of thermal imaging to monitor ethnic violence in Myanmar to the use of AI satellite imaging to quantify village destruction in Darfur, AI is already being used in impactful ways in this space. With new technologies emerging each day, the use of AI will continue to expand into the human rights space. The use of AI for the "good" of human rights may be transformative. At the same time, the same technologies may be used to perpetuate harms, so it is equally important to understand where AI must be mitigated by thoughtful legal and regulatory intervention. Based on the information gleaned from the case studies and exploration of limits and harms, this article proposes a framework for assessing impact of AI in the international human rights monitoring and reporting context. Using a cost-benefit analysis, practitioners can use this framework to determine where in the human rights space AI should or should not be used.

* J.D., UCLA School of Law, 2023. My deepest thanks to Professor Michael Karanicolas for his continued guidance, support, and feedback. Without him, this piece would not be possible. I am grateful to my many professors at UCLA Law who have pushed me to think more critically and creatively about the law. A special thank you to my parents, my sister, my grandma, and my closest friends for their continued support during the research and writing of this Note. For over a year, they have listened to me talk about AI every chance I get. I am grateful to the *Stanford Technology Law Review* editorial team for their thoughtful work editing this piece. The opinions expressed in this Note are mine alone. They do not purport to reflect the opinions or views of anyone else, including past or present employers.

TABLE OF CONTENTS

INTRODUCTION

AI is everywhere. Each time a new AI technology is introduced, legal and ethical questions arise. Law and policy struggle to keep pace as we confront unforeseen possibilities and heightened risks.[1] This is particularly true as it relates to human rights.

For example, in November 2022, OpenAI launched ChatGPT,[2] a longform, question answering AI chatbot that uses Reinforcement Learning from Human Feedback to answer questions conversationally.[3] With each new user, OpenAI is better able to deliver on its promise of gathering and using data to further train and fine-tune its program.[4] Already the technology has been called "highly capable"[5] with Microsoft committing to its potential by announcing a $10 billion investment into the technology.[6] On February 6, 2023, Google announced Bard AI in response to (and in direct competition with) ChatGPT.[7] The use of ChatGPT raises interesting questions across industries. For example, in academic settings, there is already widespread debate over whether the use of AI will boost cheating and disrupt education, or whether it can be used as a tool

---

[1] Government officials began flagging the dangers of AI and raising the need for laws to regulate its use even before ChatGPT launched. Though an exact approach has not been solidified, some progress has been made towards developing strategies to mitigate its potential harms. Lucy Papachristou & Jillian Deutsch, *ChatGPT Advances Are Moving So Fast Regulators Can't Keep Up*, BLOOMBERG (Mar. 17, 2023, 1:00 AM), https://perma.cc/H2VT-BEA6; *see also* Olivia Solon, *The Tech Behind Those Amazing, Flawed New Chatbots*, BLOOMBERG (Mar. 22, 2023, 4:11 AM), https://perma.cc/Z8JZ-K8Z7.

[2] *Introducing ChatGPT,* OPENAI (Nov. 30, 2022), https://perma.cc/5D6X-HG3F. On November 30, 2022, OpenAI introduced ChatGPT as a "trained . . . model . . . which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests." *Id.*

[3] *Id.*

[4] *Id.* OpenAI made ChatGPT free during the research phase with a promise that the technology will continue to learn and perfect in addition to getting users' feedback about its strength and witnesses. *Id.*

[5] *See, e.g.*, Alex Hughes, *ChatGPT: Everything You Need to Know About OpenAI's GPT-3 Tool*, BBC SCI. FOCUS (May 5, 2023, 4:53 PM), https://perma.cc/JQD6-VBWX.

[6] Cade Metz & Karen Weise, *Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I.*, N.Y. TIMES (Jan. 12, 2023), https://perma.cc/V5YU-XB55; *see also* Jagmeet Singh & Ingrid Lunden, *OpenAI Closes $300M Share Sale at $27B-29B Valuation*, TECHCRUNCH (Apr. 28, 2023, 4:10 PM), https://perma.cc/4ZQ3-BA2R ("The size of Microsoft's investment is believed to be around $10 billion, a figure we confirmed with our source.").

[7] Zoe Kleinman, *Bard: Google Launches ChatGPT Rival*, BBC (Feb. 6, 2023), https://perma.cc/7P4T-UPNW.

for good.[8] Already, institutions are being forced to enact new policies and procedures to respond to the critical changes ChatGPT presents.[9]

But these types of generative AI technologies pose additional and unique legal and ethical questions in the realm of human rights law. For example, what if governments could use technologies like ChatGPT to generate human rights reports? Conceivably, it can be used to solve resource challenges by reducing time spent drafting or reviewing reports. At the same time, it forces one to think about the consequences; what is lost when you take the human element out of report generation? What is gained from forcing governments—real human beings who hold positions of power—to engage with human rights data? How does AI change this dynamic?

This Note aims to advance emerging conversations on AI and international human rights law by providing a comprehensive mapping of the implementation and application of AI technologies in human rights monitoring and reporting. The Note uses a series of case studies as a way of (1) demonstrating the transformative potential of AI; (2) exploring the limitations of these technologies and the risks and potential harms associated with their use; and (3) deciphering what this tells us about our legal structures and institutions responsible for the setting, monitoring, and enforcement of human rights. The prominence and pervasiveness of AI will continue to expand into the human rights space, and the potential of its use for the "good" of human rights may be transformative. Where human rights are at risk, it is critical to understand how and when AI can—and even *should*—be used as a

---

[8] *See, e.g.*, Kevin Roose, *Don't Ban ChatGPT in Schools. Teach with It.*, N.Y. TIMES (Jan. 12, 2023), https://perma.cc/37HB-S8AR. Here, Roose, a technology columnist, summarizes the discussion around the use of AI technologies—particularly ChatGPT—in classrooms. Roose raises concerns—like the technology producing wrong or misleading answers or the propensity for cheating and misuse—and existential questions about the role of teachers. However, ultimately, Roose advocates for schools embracing technologies like ChatGPT as a teaching aid to unlock creativity, offer tutoring services, and better prepare students for the future.

[9] For example, Stanford University adopted guidance in February 2023, and in April 2023, the UCLA Dean of Students sent an email to students setting forth "Expectations Regarding ChatGPT and Other AI Tools in Academic Work." *Generative AI Policy Guidance*, STANFORD UNIVERSITY, https://perma.cc/MVE5-J77S; Email from UCLA Dean of Students, Graduate and Undergraduate Divisions and Councils, to UCLA Students (Apr. 4, 2023, 01:00 PM PDT) (on file with author) ("Unless an instructor indicates otherwise, the use of ChatGPT or other AI tools for course assignments is equivalent to receiving assistance from another person. Individual instructors have the authority to establish course policies for the use of ChatGPT and other AI tools. Acceptable use may vary from one course to another, and indeed from one assignment to another.").

way to advance rights. And it is equally important to understand the areas in which AI must be mitigated by thoughtful legal and regulatory intervention.

This Note proceeds in four parts. Part I provides a brief overview of existing literature on AI and human rights and articulates how this Note contributes to this scholarship. Many scholars have taken up the interesting legal questions raised by AI, human rights, and the connection between the two. However, they have primarily approached these questions from two angles: (1) exploring the ways in which AI technologies may pose risks to human rights and how they can be better designed to ensure rights; and (2) proposing how to use human rights law to evaluate and address the complex impacts of AI on society. This Note takes a different approach, coming at the question from the opposite direction. Rather than looking at AI's impact on human rights, I am looking specifically at how (and to what extent) AI can be used as a tool in the practice and application of international human rights law. Here, much of the focus is on AI's use in monitoring and reporting on rights violations.

Part II, through the use of case studies, looks at how AI is currently being used in international human rights law. Many civil society organizations have already begun piloting AI technologies to monitor human rights violations.[10] Organizations use AI to capture violations (immediately and over time), submit shadow reports of this data to international enforcement bodies, and ultimately aid in holding states accountable for major violations.[11] This Note uses case studies to explore the benefits of AI in this space such as by increasing efficiency, expanding capacity, providing better tracking of data over time, shielding humans from sensitive or traumatic information, authenticating data, and boosting accountability through use as an adversarial tool.

Part III also uses case studies to explore the limits of AI in the human rights space and identify harms associated with the use of AI to further perpetuate human rights abuses. Focusing on the specific concerns AI raises in regard to human rights, this also includes a deep dive into the ways in which AI-related harms might challenge existing institutions and structures that currently set, monitor, and enforce international human rights law.

Part IV of this Note proposes a framework for assessing AI's impact, exploring factors such as who is using the AI, how the AI technically works, how the AI is being used and in what context, to what extent individual rights are at risk, and the degree to which victims might be harmed. Using a cost-benefit

---

[10] *See, e.g.*, Cornebise et al., *infra* note 80.
[11] *E.g.*, *id.*

analysis, practitioners can use this framework to determine: (1) low risk areas where in the human rights space AI can be used with little concern; (2) areas where AI can be used with appropriate measures of caution and constraint; (3) high risk areas where the major costs (and limited constraints) might still be outweighed by the potential benefits; and (4) areas where the use of AI is never appropriate.

In closing, this Note outlines a handful of ideas for moving forward in developing policy and research in this space.

I. BACKGROUND AND LITERATURE REVIEW

A. An Introduction to International Law and the Accountability Challenge

International law "consists of [the] rules and principles governing the relations and dealings of nations with each other, . . . [nations] and individuals, and . . . international organizations."[12] Article 38 of the Statute of the International Court of Justice, which forms part of the U.N. Charter, identifies the four primary sources of international law widely recognized as the foundation of international law:[13] (1) international conventions (or treaties); (2) customary international law; (3) general principles of law recognized by states; and (4) judicial decisions and teachings (otherwise known as soft law).[14]

Most of the default norms for treaty-making are set forth in the Vienna Convention on the Law of Treaties.[15] The basic premise is that states choose to enter into agreements, and as a result undertake treaty obligations knowing that such commitments will have the force of law.[16] By contrast, customary international law is not based on direct state consent.[17] Rather, customary international law is established when states engage in a general and consistent practice out of a sense of legal obligation.[18]

---

[12] Legal Info. Inst., *International Law*, CORNELL L. SCH., https://perma.cc/3MBC-HB6B.

[13] JEFFREY L. DUNOFF ET AL., INTERNATIONAL LAW: NORMS, ACTORS, PROCESS: A PROBLEM-ORIENTED APPROACH 31-32 (5th ed. 2020).

[14] Statute of the International Court of Justice art. 38, ¶ 1, June 26, 1945, 59 Stat. 1055 (outlining sources of international law).

[15] Vienna Convention on the Law of Treaties, *opened for signature* May 23, 1969, 1155 U.N.T.S. 331.

[16] CHIMÈNE KEITNER, INTERNATIONAL LAW FRAMEWORKS 13 (5th ed. 2021).

[17] *Id.* at 32.

[18] *Id.*

States increasingly use non-traditional forms of lawmaking to supplement treaties and custom in regulation of international activities.[19] This soft law stems from standard setting activities of international organizations, regional bodies, multinational enterprises or multinational corporations, and non-governmental organizations (NGOs).[20] While not binding, soft law plays a critical role in international law, functioning as a gap-filler that gives guidance to states, civil society, and other stakeholders in the absence of binding legal norms.[21] Soft law has the potential to transform into norms that become widely accepted and may develop into customary international law over time.[22]

Much of the scrutiny and skepticism that international law faces derives from the fact that international law is a consent based system with major accountability gaps and limited enforcement mechanisms: "there is no centralized legislature to enact the law, centralized executive to apply or enforce it, or centralized judiciary with general and compulsory jurisdiction to interpret and adjudicate associated disputes under it."[23] This is especially true in the realm of human rights law where it is often the same states who are responsible for ensuring rights who are engaging in oppressive and exploitative behavior.[24]

B. What is International Human Rights Law?

The protection of human rights is a crucial objective of international law and the international legal system. Following World War I, the protection of human rights became a central issue of concern to the international community,[25] and the idea of establishing a clearly articulated set of international human rights laws emerged.[26] International human rights law is both the articulation of obligations which states are bound to respect in the promotion of human welfare[27] and the establishment of an international

---

[19] *See* DUNOFF ET AL., *supra* note 13, at 63.
[20] *See Id.* at 63-64.
[21] *See* KEITNER, *supra* note 16, at 37.
[22] *See generally* Mauro Barelli, *The Role of Soft Law in the International Legal System: The Case of the United Nations Declaration on the Rights of Indigenous Peoples*, 58 INT'L & COMP. L.Q. 957 (2009).
[23] DUNOFF ET AL., *supra* note 13, at 31.
[24] *See*, KEITNER, *supra* note 16, at 149.
[25] *Id.* at 6-7, 149.
[26] *Id.* at 153.
[27] *Id.* at 149.

human rights system to protect and ensure the enjoyment of such rights.[28] The international human rights system has a broad but critical objective: to provide both the normative framework and support for institutional systems to ensure that states promote and protect the human rights and fundamental freedoms of individuals and groups within their jurisdiction.[29] Like the rest of international law, international human rights law faces challenges in accountability and enforcement.[30]

A series of international human rights treaties and other instruments have been adopted since 1945 to create a body of internationally recognized human rights.[31] There are a wide range of issues that are implicated in the international human rights framework, encompassing both civil and political rights as well as economic, social and cultural rights. The core international human rights instruments that confer these legal obligations on states are the U.N. Charter,[32] the Universal Declaration of Human Rights (UDHR)[33] and associated U.N. system,[34] nine subsequent multilateral human rights treaties,[35] and regional

---

[28] The International Human Rights system consists of both international instruments and institutions. Instruments cover "all the different documents that embody human rights standards: legally binding treaties, covenants and conventions (hard law), as well as commitments expressed in declarations, resolutions, guiding principles, codes of conduct etc. (soft law)." WILLIAM G. O'NEILL & ANNETTE LYTH, *The International Human Rights System*, *in* MANUAL ON HUMAN RIGHTS MONITORING: AN INTRODUCTION FOR HUMAN RIGHTS FIELD OFFICERS (3d ed. 2008). Institutions include U.N. charter and treaty-based organs (like the General Assembly, Security Council, Human Rights Council, OHCHR, etc.); investigatory, thematic, or special mechanisms (such as country and thematic special rapporteurs and thematic working groups); and treaty-based organs (like the nine human rights treaty bodies' accompanying committees, made up of independent experts). *Id.*

[29] KEITNER, *supra* note 16, at 149.

[30] *Id.* at 162-63.

[31] *Id.* at 152-53 (discussing timeline of adopted treaties).

[32] *See generally* U.N. Charter.

[33] G.A. Res. 217A (III), at 152-53 (Dec. 10, 1948).

[34] In addition to the United Nation's principal organs, the United Nations has also established human rights-specific legal instruments to ensure the rights in the UDHR. These instruments include the Office of the High Commissioner for Human Rights (OHCHR) with offices in six regions, the Human Rights Council which serves as the key independent U.N. intergovernmental body responsible for human rights, thematic and country-specific Special Procedures and Independent Experts, and the U.N. Development Group's Human Rights Working Group. *See How Does the UN Promote and Protect Human Rights?*, UNITED NATIONS, https://perma.cc/VBW3-WGNF.

[35] These nine treaties, together with the UDHR, form the International Bill of Rights. OFF. OF THE U.N. HIGH COMM'R FOR HUM. RTS., THE UNITED NATIONS HUMAN RIGHTS TREATY SYSTEM: FACT SHEET NO. 30/REV. 1, 14 (2012), https://perma.cc/S8E4-NCK5 (outlining nine human rights treaties: (1) the International Convention on the Elimination of All Forms of Racial Discrimination; (2) the International Covenant on Economic, Social and Cultural Rights; (3) the International

human rights treaties,[36] all of which require domestic incorporation of the legal rights and obligations enshrined in them. Each system has a complex structure and requires the involvement of states at various levels. Common to all of them is a requirement that states monitor and report on the enshrined human rights,[37] though the reporting requirements may vary among instruments or bodies.

As a result, international human rights law faces major accountability and enforcement problems.[38] In a system that relies predominantly on treaty and custom as the binding sources of law, international law is inherently a system based on a foundation of consent; the dominant understanding of legal positivism is that states are sovereign entities, only bound by the legal norms and obligations to which they agree.[39] Even when states consent to be bound by international law, there is still widespread, routine non-compliance; states fail to ensure that human rights are protected and themselves violate the rights of individuals and groups.[40] To facilitate accountability, one of the major mechanisms built into major human rights treaties is the creation of an independent treaty body to monitor state compliance, the crux of which is a mandatory requirement that states regularly monitor and report on the implementation of rights within their state.[41] Individuals are able to bring claims of specific human rights abuses to international treaty bodies through the use of individual complaints and communications—these can serve as a way to hold national governments accountable when domestic proceedings fail to redress

---

Covenant on Civil and Political Rights; (4) the Convention on the Elimination of All Forms of Discrimination against Women; (5) the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment; (6) the Convention on the Rights of the Child; (7) the International Convention on the Protection of the Rights of All Migrant Workers and Members of their Families; (8) the Convention on the Rights of Persons with Disabilities; and (9) the International Convention for the Protection of All Persons from Enforced Disappearance).

[36] African Charter on Human and Peoples' Rights, *opened for signature* June 1, 1981, 1520 U.N.T.S. 26363; Charter of the Organization of American States, *opened for signature* Apr. 30, 1948, 119 U.N.T.S. 1609; American Declaration of the Rights and Duties of Man, *opened for signature* Oct. 5, 1948, 122 U.N.T.S. 4; Protocol of Buenos Aires, Feb. 27, 1967, 21 U.S.T. 607, 721 U.N.T.S. 324; American Convention on Human Rights, *opened for signature* Nov. 22, 1969, 1144 U.N.T.S. 17955; European Convention for the Protection of Human Rights and Fundamental Freedoms, *opened for signature* Nov. 4, 1950, 213 U.N.T.S. 2889.

[37] KEITNER, *supra* note 16, at 164.

[38] ELENA KATSELLI PROUKAKI, THE PROBLEM OF ENFORCEMENT IN INTERNATIONAL LAW 1-2 (2010).

[39] ANDREW CLAPHAM, BRIERLY'S LAW OF NATIONS 49-51 (7th ed. 2013).

[40] *See* PHILIP ALSTON & RYAN GOODMAN, INTERNATIONAL HUMAN RIGHTS 768-71 (2013).

[41] *Id.* at 768-69.

harms.[42] International human rights law hinges on regular monitoring, tracking, and reporting on rights.

Research on state reporting has shown that the more states participate in the reporting process, the greater the improvement in the enjoyment of rights for individuals.[43] Importantly, state reports have benefited from an increase in state capacity to collect, systematize, and analyze data.[44] As a result, reports are more thorough, candid, and ultimately more responsive to treaty obligations.[45] This suggests that if states had better tools for collecting data and compiling it into reports, they would include that information into their reports. A comprehensive system of regular monitoring and reporting seeps into domestic politics, providing the crucial connecting point between international obligations and domestic incorporation. Ideally, the reports do more than force states to collect data, spurring state self-reflection and resulting in substantive changes. For treaty obligations to materialize into the realization of domestic rights, it is critical that states are equipped with tools and technologies to make monitoring and reporting easier, more efficient, and more streamlined, while still maintaining the ability to capture specific information related to a broad range of rights.

C. What Is AI?

Defining AI is not a simple feat. While Encyclopedia Britannica defines AI as "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings,"[46] AI is more of a term of art than a specific definable "thing."

For our purposes, it may be helpful to break down and understand AI under the following frequently questioned categories: (1) the types of artificial

---

[42] *Instruments & Mechanisms: International Human Rights Law*, OFF. OF THE U.N. HIGH COMM'R FOR HUM. RTS., https://perma.cc/2HAZ-UKGH.

[43] Cosette D. Creamer & Beth A. Simmons, *The Proof Is in the Process: Self-Reporting Under International Human Rights Treaties*, 114 AM. J. INT'L. L. 1, 1 (2020).

[44] *Id.* at 2.

[45] *Id.* at 3.

[46] B.J. Copeland, *Artificial Intelligence*, ENCYC. BRITANNICA, https://perma.cc/35ND-WMAJ.

intelligence;[47] (2) the major categories of AI technologies;[48] and (3) the types of problems AI can solve.[49]

Broadly, artificial intelligence is typically separated into four types: (1) reactive machines; (2) limited memory; (3) theory of mind; and (4) self-awareness.[50] Most existing AI technologies fit into the first two types of reactive machines (where AI makes basic inferences based on data inputs) and limited memory (where AI can take inputs and use predictive modeling to "learn").[51] Limited memory is characterized by the technology's ability to absorb large amounts of training data and improve over time.[52]

AI technologies can also be separated into different, often overlapping, categories. These categories include technologies like automation, machine learning, natural language processing (NLP), computer vision, deep learning, and robotics.[53] Oxford Language defines machine learning as "the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data."[54] The AI "learns" insofar as the algorithms "improve their performance by examining more data and detecting additional patterns in that data that assist in making better automated decisions."[55] Deep learning is a subset of machine learning that

---

[47] Arend Hintze, *Understanding the Four Types of Artificial Intelligence*, GOV'T TECH. (Nov. 14, 2016), https://perma.cc/NMQ2-BPXU.

[48] Ed Burns et al., *What is Artificial Intelligence?*, TECHTARGET, https://perma.cc/JY9F-F95M (Mar. 2023).

[49] Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1321-24 (2019).

[50] Hintze, *supra* note 47.

[51] *See id.*

[52] *Id.*

[53] Burns et al., *supra* note 48. Automation is the use of repetitive, basic rules-based data processing to complete tasks traditionally done by humans; AI is paired with automated technologies to expand the types and number of tasks performed. *Id.* NLP is the human language processing by a computer program based on machine learning and might include tasks like spam detection, text translation, or speech recognition. *Id.* Machine vision uses analogue-to-digital conversion in order to capture and analyze visual information—one of its major uses is medical image analysis. *Id.* Robotics is the use of robots to perform human tasks, such as on assembly lines. *Id.*

[54] Richard Gate, *Machine Learning vs Artificial Intelligence*, OBJECTSPECTRUM (OCT. 1, 2022), https://perma.cc/9A5E-C9L8.

[55] Surden, *supra* note 49, at 1312. There are four types of machine learning algorithms: (1) supervised learning, where the algorithm is trained by human experts—data sets are labeled, patterns are then detected, and then those patterns are used to label new data sets; (2) unsupervised learning, where data sets are not labeled but are sorted according to

"uses artificial neural networks to recognize patterns and relationships in data" by dividing networks into different layers—it is often used in the context of image and speech recognition.[56]

Others have suggested that it is easier to define AI in terms of the problems it is trying to address, describing AI as a technology focused on automating specific tasks that normally require or involve human intelligence when being performed.[57] This can be a helpful way to think about AI as one gets into the weeds of how the technology itself works. Applied to the international human rights monitoring and reporting system, AI could be used for the organization of data, computational capabilities, repurposing of existing data sets to model and forecast, and even learned decision-making.

It is helpful from the onset to articulate the difference between AI and data. These two terms are easy to conflate or confuse, but it is important to understand the distinction between the two and how they work together. AI technologies require large data sets in order to be developed and deployed effectively;[58] the AI must learn from somewhere. Every algorithm has an input and output, but "with machine learning, computers write their own programs, so we don't have to."[59] For machine learning,[60] the data comes first: "the development of a machine learning algorithm depends on large volumes of data, from which the learning process draws many entities, relationships, and clusters."[61] Data sources discussed in this Note include satellite images, social media posts, and local news reports. At the same time, AI technologies may be

---

similarities or differences—often these are used in pattern detection or descriptive modeling; (3) semi-supervised learning, which falls in between these two; and (4) reinforcement learning, where the algorithm uses observations gathered through interactions with the environment to take actions that would maximize its performance, increasing reward or minimizing risk. *See* Jose Fumo, *Types of Machine Learning Algorithms You Should Know*, TOWARDS DATA SCI. (June 15, 2017), https://perma.cc/3AWK-4SNZ.

[56] Bastian Maiworm, *Deep Learning vs. Machine Learning—Understanding the Differences*, MORE THAN DIGIT. (Jan. 14, 2023), https://perma.cc/QK2V-AZ7E.

[57] *See* Surden, *supra* note 49, at 1307 ("A few examples will help illustrate this depiction of AI. . . . including playing chess, translating languages, and driving vehicles").

[58] Surden, *supra* note 49, at 1316.

[59] PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD 13-14 (2015) ("Machine learning is the scientific method on steroids. It follows the same process of generating, testing, and discarding or refining hypothesis . . . in a fraction of a second. Machine learning automates discovery.").

[60] *See id.* at 6, 8. ("Machine learning takes many different forms and goes by many different names: pattern recognition, statistical modeling, data mining, knowledge discovery, predictive analytics, data science, adaptive systems, self-organizing systems, and more."). Much like Domingos, this Note uses machine learning to refer broadly to all of these forms.

[61] Joe McKendrick, *The Data Paradox: Artificial Intelligence Needs Data; Data Needs AI*, FORBES (June 27, 2021, 11:59 AM EDT), https://perma.cc/MLY5-CSW5.

able to make inferences and come to conclusions about the data such as identifying gaps in data, collating or making sense of the data, or making algorithmic decisions based on the data, ultimately producing new data and new algorithms.[62] This paradox will be discussed more in Part III.

D. The Connection Between AI and Human Rights: A Brief Literature Review

Much legal scholarship has explored the impacts of AI on human rights. Many have looked at the human rights implications of using AI technologies, including ethical considerations,[63] such as how AI should be designed "in a rights-respecting manner."[64]

One of the most prominent areas of focus has been on understanding how AI impacts human rights broadly, identifying the challenges, vulnerabilities, and harms.[65] Data privacy[66] and bias[67] have been two focal points in legal discussion. One such study looked at the following list of harms: "lack of algorithmic transparency; cybersecurity vulnerabilities; unfairness, bias, and discrimination; lack of contestability; legal personhood; intellectual property; adverse effects on workers; privacy and data protection issues; liability for damage; and lack of accountability for harms."[68] The study provides a helpful overview of the types of harms that exist broadly within AI and human rights.

Others have looked at AI and human rights impacts in the context of varying industries. For example, a team of ethics and governance scholars at Harvard's Berkman Klein Center for Internet and Society mapped the human rights impacts of AI systems in the following six fields of endeavor: Criminal Justice, Access to the Financial System, Healthcare, Online Content Moderation, Human Resources, and Education.[69] The team looked at various effects of AI use on human rights (both positive and negative) to assess how their disparate impacts can

---

[62] DOMINGOS, *supra* note 59, at 6.

[63] Giovanni Sartor, *Artificial Intelligence and Human Rights: Between Law and Ethics*, 27 MAASTRICHT J. EUR. & COMP. L. 1 (Dec. 17, 2020), https://perma.cc/KA42-LHS6 (examining the way in which AI is addressed by ethical and legal rules, principles, and arguments).

[64] FILIPPO RASO ET AL., ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS: OPPORTUNITIES AND RISKS 4, 57 (2018).

[65] *See, e.g.*, *id.*

[66] *See, e.g.*, Sylvia Lu, Note, *Data Privacy, Human Rights, and Algorithmic Opacity*, 10 CALIF. L. REV. 2087 (2022).

[67] *See, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019).

[68] Rowena Rodrigues, *Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities*, 4 J. RESPONSIBLE TECH. 1, 1 (2020).

[69] RASO ET AL., *supra* note 64, at 2, 17-51.

be addressed using a human rights framework which relies on and provides shared language and infrastructure.[70]

Legal scholarship has also focused on providing frameworks by which to assess human rights impacts in AI design. Scholars have looked at the ways that traditional rules of law and ethics (both together and distinct from one another) can be used to evaluate and approach AI and its impacts.[71] Some have taken a more comprehensive look at the relationship between AI and human rights; ultimately articulating a new discipline where the two intersect.[72] Others have taken an evidence-based approach undertaking and advocating for a human rights impact assessment in the development of data-intensive AI systems.[73]

As noted, this Note takes a different approach than existing literature. Rather than looking at AI's impact on human rights (though this is relevant to my discussion broadly), this Note looks specifically at how and to what extent AI can be used as a tool in the practice and application of international human rights law. Much of the focus is on AI's use in monitoring and reporting on rights violations, including positive applications, potential limitations, and an exploration of harms specific to AI's use in this space.

II. POSITIVE APPLICATIONS OF AI IN INTERNATIONAL HUMAN RIGHTS LAW

Within the international human rights law context, AI has been used predominantly by civil society organizations as a tool to monitor, track and report on both the enjoyment of rights and more prominently the violation of rights.[74] Civil society organizations then use that information to try and hold governments accountable through a number of different advocacy strategies.[75] For example, an organization may submit these reports to international human rights treaty bodies as a supplement to state self-reporting. Outside of the formal reporting system, organizations can also publish reports and release information to the public with the goal of putting pressure on states to adjust

---

[70] *Id.* at 8.

[71] Sartor, *supra* note 63, at 1.

[72] Emmanuel Kabengele Mpinga et al., *Artificial Intelligence and Human Rights: Are There Signs of an Emerging Discipline? A Systematic Review*, 15 J. MULTIDISCIPLINARY HEALTHCARE 235, 235-36 (2022), https://perma.cc/94VV-F9DE.

[73] Alessandro Mantelero & Maria Samantha Esposito, *An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems*, 41 COMPUT. L. & SEC. REV. 1, at 9-10 (2021), https://perma.cc/9Q9W-ND9L.

[74] *See, e.g.*, Cornebise et al., *infra* note 80.

[75] *E.g.*, *Id.*

their behavior or bring to the attention of the international community for external pressure.

As indicated, there are major benefits of using these types of AI technologies in this way. Broadly, AI technologies can be used as a tool to more efficiently, effectively, and comprehensively monitor and report on rights—both for civil society organizations and for state actors. For example, technology has the capability of using open-source data on the internet to populate databases that keep track of rights violations. It can weed through massive amounts of data to identify major trends by key issue areas or demographic markers over time.[76] It can even identify gaps in data to indicate where information might be missing.[77] Humans will be able to make sense of greater amounts of information more quickly as AI technologies become more and more capable of synthesizing information and drafting summaries.[78] AI can boost accuracy of reporting by checking reports or data from different sources against one another and comparing data against outside data sources to validate and authenticate results.[79] At the same time, ethical questions are raised about what is lost when AI takes on the burden of reporting. Is there a value to having human rights workers and state actors conduct the reporting? What is lost if they do not engage in this process?

In the sections that follow, I have identified a series of case studies of how AI is already being used within the international human rights space. These examples showcase the different ways civil society organizations have developed and deployed technologies that monitor and track violations (or progress on the enjoyment) of specific rights.

A. Case Study: Quantifying Village Destruction in Darfur

In 2018, Amnesty International partnered with professors from University College London and the University of Amsterdam to quantify village destruction

---

[76] MICHAEL L. LITTMAN ET AL., GATHERING STRENGTH, GATHERING STORMS: THE ONE HUNDRED YEAR STUDY ON ARTIFICIAL INTELLIGENCE (AI100) 2021 STUDY PANEL REPORT 9 (Sept. 2021), https://perma.cc/VU9S-AXZB.

[77] *Id.*

[78] HOROWITZ ET AL., ARTIFICAL INTELLIGENCE AND INTERNATIONAL SECURITY (July 10, 2018), https://perma.cc/AVP2-9XSF.

[79] *See* PRICEWATERHOUSECOOPERS CONSULTING, ARTIFICIAL INTELLIGENCE FOR REPORTING 6 (2020), https://perma.cc/99NP-P8VS.

in Darfur.[80] They tout the project as the first ever use of machine learning for human rights led by an NGO partnership between civil society and technical experts.[81] The organization, with help from a team of AI experts, created an AI algorithm trained to scan satellite images for the rest of the country to identify additional human habitats and detect destruction using a multi-task binary classification.[82] The goals were to better understand the conflict at a larger scale, garner public outrage, and prompt public engagement in the form of an online petition.

The algorithm "learned" from open-source data that had been collected by Amnesty during their Eyes on Darfur campaign.[83] The data collected (and then used to teach the AI) was crowdsourced images of ongoing destruction of civilian villages.[84]

The success of the algorithm relied extensively on crowdsourced labeling efforts (by humans) that took place in 2016 when Amnesty launched a campaign asking the public to help label the satellite images.[85] In just three weeks, 28,600 volunteers from 147 countries took part in analyzing 2.6 million satellite images covering 326,000 square kilometers to identify Darfur's remote villages and key buildings and structures.[86]

Amnesty then turned to AI to scale up research and cover the whole of Darfur. Based on information collected and indexed previously, Amnesty International and the AI experts created a machine learning model to scan satellite images for the rest of the country to identify additional human habitats and detect destruction. Initially, the AI used a multi-task binary classification—looking at the satellite images and was able to quantify and determine both habitation and destruction.[87] According to the report, "the model was able to identify 'human presence,' as well as differentiate 'destroyed' and 'mixed' (i.e., partially destroyed) villages, at a country-wide scale, covering 500,000 square kilometers."[88] Following the initial study, the team went back and used a multi-

---

[80] Julien Cornebise et al., *Witnessing Atrocities: Quantifying Villages Destruction in Darfur with Crowdsourcing and Transfer Learning*, Proc. AI for Soc. Good NeurIPS2018 Workshop (2018).
[81] *Id.*
[82] *Id.*
[83] *Id.*
[84] *Id.*
[85] Project Complete, *Decode Darfur*, AMNESTY INT'L (Oct. 13, 2016), https://perma.cc/T8AD-3TRR.
[86] *Id.*
[87] Cornebise et al., *supra* note 80, at 3.
[88] *Id.* at 2-3.

label classification system which the report identifies as the "model produc[ing] three probabilities: the probability that nothing is in the tile, the probability that intact buildings are in a tile, and the probability that destroyed buildings are in the tile."[89]

The model was extremely successful, both in terms of its ability to reduce the time and resources it would have taken humans to conduct the research and to improve accuracy when assessed against human expert performance. Compared to the tiles with destroyed buildings that human experts identified, the AI was able to successfully identify 81% of the same tiles.[90] By contrast, of those the AI flagged, 85% were also flagged by human experts.[91]

There is great potential for future projects based on machine learning and satellite imaging. Amnesty published a whitepaper in 2019 detailing the potential of using this type of satellite imagery ("computer vision and earth observation data") and AI machine learning for human rights monitoring in other localities and at a larger scale.[92] For example, Amnesty tested their models (without retraining them) along the border between South Sudan and Uganda, and the models worked even though the terrain is distinct.[93] Though it is clear the models would need to be further trained to improve accuracy in the region, these types of tools could be easily modified for use in other places.[94] In the report, Amnesty notes the potential the AI creates for future projects: "[We] learned how to curate datasets from different satellite imagery platforms, and also what an efficient data pipeline looks like. A lot of handy in-house tooling [were] built for this project, which can be reused in similar satellite imagery projects."[95]

The report identifies some of the major challenges associated with the project, including sensitivity of the images and graphics, the need for data validation, transparency of data and risks associated with widespread dissemination, and risks of malicious attacks on vulnerable communities.[96] The AI is designed to provide specific information about specific villages in Darfur,

---

[89] Milena Marin et al., *Using Artificial Intelligence to Scale Up Human Rights Research: A Case Study on Darfur*, CITIZEN EVIDENCE LAB (July 6, 2020), https://perma.cc/5YMV-MFJP.
[90] *Id.*
[91] *Id.*
[92] *Id.*
[93] *Id.*
[94] *Id.*
[95] *Id.*
[96] *Id.*

and as a result, divulges the precise location of vulnerable villages.[97] There is a risk that insurgent groups might use the information to target these villages for future harms. The report acknowledges that it "touches upon a larger societal discussion of scientists' responsibility in the use of their tools" and proposes NGOs might work together to help navigate this issue and mitigate harms.[98] These types of harm will be discussed in greater detail in Part III.

B. Case Study: Using AI to Forecast International Displacement

In 2021, the Danish Refugee Council (DRC) used AI and machine learning technology to predict displacement trends for 2021 and 2022.[99] The tool was initially developed and deployed in 2021 and was again deployed in 2022.[100]

The tool has collected, compiled, and analyzed data on 148 indicators related to conflicts, governance, economy, climate, human rights, and societal trends.[101] Based on this information, combined with displacement trends from 2020 and 2021 as well as data going as far back as 1995,[102] the model is able to correctly forecast how many people would be displaced annually over the following three years.[103] The model uses open access data from sources like the World Bank and NGOs to predict forced displacement in a given country.[104]

The forecast predicted that 3.7 million more people would be displaced in 2021 and 7.2 million would be displaced by the end of 2022.[105] The 2021 model

---

[97] Cornebise et al., *supra* note 80, at 1-2, 4.

[98] *Id.*

[99] DANISH REFUGEE COUNCIL, GLOBAL DISPLACEMENT FORECAST 2021, at 3-4 (July 2021), https://perma.cc/YR6W-3N3G [hereinafter 2021 FORECAST].

[100] *See id.*; DANISH REFUGEE COUNCIL, GLOBAL DISPLACEMENT FORECAST 2022 (Feb. 2022), https://perma.cc/Q288-EN9F [hereinafter FEBRUARY 2022 FORECAST]; DANISH REFUGEE COUNCIL, GLOBAL DISPLACEMENT FORECAST 2022 JULY UPDATE (July 2022), https://perma.cc/8ZRS-KCGR [hereinafter JULY 2022 FORECAST].

[101] 2021 FORECAST, *supra* note 99, at 32; FEBRUARY 2022 FORECAST, *supra* note 100, at 50.

[102] 2021 FORECAST, *supra* note 99, at 33; FEBRUARY 2022 FORECAST, *supra* note 100, at 14, 51.

[103] 2021 FORECAST, *supra* note 99, at 5-6; FEBRUARY 2022 FORECAST, *supra* note 100, at 3, 7.

[104] 2021 FORECAST, *supra* note 99, at 32 ("The data is all derived from open source. The main data sources are the World Bank development indicators, the Armed Conflict Location & Event Data Project (ACLED), the Uppsala Conflict Data Program (UCDP), EM-DAT, U.N. agencies (UNHCR, the World Food Programme, The Food and Agriculture Organization), Internal Displacement Monitoring Center (IDMC), etc. In total, the system aggregates data from 18 sources, and contains 148 indicators."); FEBRUARY 2022 FORECAST, *supra* note 100, at 50 (same).

[105] 2021 FORECAST, *supra* note 99, at 3-4, 10.

uses data from twenty-four countries covering 84% of the world's displaced population in 2020 (roughly 69 million people).[106]

The DRC was able to compare these predictions with the actual numbers from 2021 once they were released to see how the model performed. They found that the average margin of error was 14% for the 2021 displacement forecasts. The highest margin of error was 45% for Libya, while the lowest margin of error was 1% for Libya. Overall, the margin of error across all 188 forecasts was 19%.[107]

Again in 2022, DRC used the tool to make predictions about 2022 and 2023.[108] The forecast covers twenty-six countries and predicts that the cumulative number of people displaced would increase by 4.6 million people in 2022 and by another 4.1 million in 2023, with a total increase of 8.7 million people displaced between 2021 and 2023.[109]

This use case provides a salient example of how AI technologies can be used to model and predict future trends for better strategic planning and response. Not only does the technology work more efficiently, but it provides organizations with information that enables them to allocate (often limited) resources more effectively. At the same time, it is a good example of an AI technology that uses open-source data from major international actors like the IMF and World Bank and then identifies and makes sense of patterns and trends within and across these data sources. It is easy to imagine this same type of forecasting tool being replicated and used in other human rights contexts beyond just displacement.

One of the major benefits of this type of AI forecasting technology is its ability to make sense of and find trends across a variety of data sources. AI can increase an organization's ability to make sense of information by analyzing large datasets and finding patterns likely unseen by the human eye (given the massive amount of data being analyzed). It can also identify patterns and trends over time. Here, the technology provided detailed data modeling for each state and then aggregated that information to find trends and forecast across states. The model was able to identify conflict as one of the major drivers of

---

[106] *Id.* at 3-4, 6.

[107] JULY 2022 FORECAST, *supra* note 100, at 3; *see also* FEBRUARY 2022 FORECAST, *supra* note 100, at 51 ("50% of the forecasts have a margin of error below 10% and almost 2/3 of the forecasts are less than 15% off the actual displacement.").

[108] JULY 2022 FORECAST, *supra* note 100, at 7, 60; FEBRUARY 2022 FORECAST, *supra* note 100, at 3.

[109] JULY 2022 FORECAST, *supra* note 100, at 6.

displacement across states.[110] While this might seem obvious, the AI is able to validate an assumption through the use of detailed data analysis, using data to back its claims and identify how conflict paired with other indicators result in different trends and forecasts. AI has the capacity to expand our vectors of understanding, making novel interpretations and identifying correlations not automatically noticeable to the human eye within data sets.

Ideally, predictive modeling can be helpful in advocating for increased humanitarian aid in places where the models predict high levels of displacement and low levels of humanitarian aid. The DRC has been able to take the tool's findings, compile them into a report for the European Union (who financed the project), and provide specific recommendations for how displacement should be addressed, including where resources should be allocated.[111] Specifically, the report aims to provide guidance on future humanitarian response allocation and draws connections between the forecasting and specific commitments of the Refugee Convention and the Global Compact on Refugees (i.e., where states must do their part to meet treaty obligations).[112]

One of the limitations of the AI model is that the most current data it uses is still from the previous year and the model is therefore largely unable to take into account unexpected developments or major changes in the geopolitical realities of a country in real time.[113] A larger discussion of the data gaps and challenges in this case study can be found in Part III.

C. Case Study: Media Monitoring and the Tracking of Death Penalty Cases

In 2018, Amnesty International and Element AI developed a tool to help track information and news related to death penalty cases.[114] The technology was designed to automate Amnesty's existing process for monitoring media

---

[110] 2021 FORECAST, *supra* note 99, at 11.

[111] *See* JULY 2022 FORECAST, *supra* note 100, at 9.

[112] *Id.*

[113] The 2021 report uses the example of Chad to demonstrate this point. While the model initially forecasted limited increases in displacement, by mid-2021, 65,000 people from Chad had already been displaced. This was largely a result of the president's unexpected death and tension within the country and region as a result. 2021 FORECAST, *supra* note 99, at 10-11. Another example is Ukraine, where over 12 million were forced to displace as a result of the war. JULY 2022 FORECAST, *supra* note 100, at 4.

[114] AMNESTY INT'L & ELEMENT AI, AI-ENABLED HUMAN RIGHTS MONITORING 10 (2019), https://perma.cc/GYU8-NDG7.

reports of executions and death penalty cases.[115] Prior to the use of the AI technology, volunteers would scan the internet, collect news articles, and then manually input the relevant information, such as name and country, into a database.[116] Amnesty was able to automate at least part of process, using the technology to "identify hundreds of potential stories per day from across the globe and automatically cluster them and use entity extraction techniques to identify country and victim."[117] The tool was able to successfully identify news articles mentioning executions with 79% accuracy.[118]

One of the major benefits of this tool is that it automates a typically laborious process, reducing the time and energy undertaken by staff and volunteers. For example, a volunteer might spend unnecessary time finding the same information over and over online with different media reporting the same case. For saliency, that volunteer would have to look through each to ensure the reporting matches up and is truly a duplicate, whereas an AI tool might be able to do the same thing instantaneously.[119]

Amnesty predicted that with additional development, the AI tool could reduce what would have been the work of four volunteers to that of one volunteer to validate the AI's "work" (i.e., checking to see if the technology missed anything, eliminating duplicates, and verifying results).[120] As indicated, there is still the need for human intervention; the AI is not perfect. But rather than having volunteers do all the work, the AI tool could do an initial pass followed by human intervention to authenticate it and correct any errors.[121]

D. Case Study: Using AI to Track Deforestation

In 2019, researchers from ETH Zurich launched a project using AI and drones to track deforestation in the Valdivian rainforest and are now looking at how the same technology can be used to track deforestation in the Amazon.[122] Deforestation is detrimental to the environment and causes multiple and

---

[115] *Id.*
[116] *See id.*
[117] *Id.*
[118] *Id.*
[119] PRICEWATERHOUSECOOPERS CONSULTING, *supra* note 79, at 3.
[120] AMNESTY INT'L & ELEMENT AI, *supra* note 114, at 10.
[121] *Id.*
[122] Lydia Skrabania, *Artificial Intelligence and Drones Join the Fight to Save the Rainforest*, RESET (Mar. 3, 2020), https://perma.cc/JT8G-MLRU.

severe human rights violations.[123] Community resistance to land grabs and forest clearing frequently results in violence against populations[124] with disparate impacts on indigenous communities.[125] Beyond direct violence, deforestation—and climate change more broadly—may destroy biodiversity and deplete water sources, which ultimately harm human health.[126]

The AI was taught to evaluate a series of aerial images from satellites and drones to see forest changes. This data can then be used to create models which predict the areas that will be most vulnerable to deforestation moving forward. The study also incorporated drone images[127] in order to identify types of tree species being destroyed, giving more detail regarding the effects on carbon dioxide levels.[128] Like the predictive modeling in the displacement forecasting example, being able to predict the areas where deforestation will have the most severe impacts can allow state and nonstate actors better and more targeted strategic planning and response.

E. Case Study: Use of Thermal Data to Monitor Ethnic Violence in Myanmar

In 2017, Human Rights Watch partnered with Element AI to create a machine learning tool that could use satellite imagery and remote sensing

---

[123] *See, e.g.*, *Climate Change in the Amazon*, WORLD WILDLIFE FOUND., https://perma.cc/4KKC-P5BX ("Coupled with land-use changes, we can expect the degradation of freshwater systems, loss of ecologically and agriculturally valuable soils, increased erosion, decreased agricultural yields, increased insect infestation, and spread of infectious diseases.").

[124] Violence may include: "forced evictions, police harassment, intimidation, death threats and violent attacks, arbitrary arrest, and retaliatory litigation and criminalization of community leaders, human rights defenders and activists. Community leaders also suffer intimidation and public smear campaigns in the media, while lawyers, local and international non-governmental organizations (NGOs) and journalists who seek to denounce violations and crimes against land defenders are subject to legal persecution and lawsuits by companies (often for libel or slander)." *Human Rights Impacts of Deforestation*, CLOSING THE GAP, https://perma.cc/4VPM-UFEN.

[125] The Yanomami peoples provide a salient example of the disparate impacts experienced by indigenous peoples where thousands of gold miners threaten their land, culture, and identity as they "lust for gold and other valuable minerals that lay beneath their ancestral territory [which] has in recent years attracted a wave of illegal prospectors who have cut down forests, poisoned rivers and brought deadly diseases to the tribe." *In the Amazon Rainforest, An Indigenous Tribe Fights for Survival*, OFF. OF THE U.N. HIGH COMM'R FOR HUM. RTS. (Aug. 9, 2022), https://perma.cc/S3DQ-W8T6.

[126] *Climate Change in the Amazon*, *supra* note 123.

[127] *Id.*

[128] *Id.*

thermal data to spot, track, and catalog human rights violations against Rohingya populations in Myanmar.[129]

Once news of the conflict broke, the team at Human Rights Watch immediately began using thermal data to monitor ethnic violence by tracking smoke plums and other indicators of destruction or burning captured on environmental satellites.[130] They were able to then use AI to combine the thermal data with aerial images to identify where destruction had taken place, finding the burning targeted Rohingya villages.[131] They were able to use AI to combine this data with information they found on public domains (like social media photos and videos) to pinpoint when burnings took place, allowing them to corroborate the testimony of individuals whose human rights were violated and identify specific perpetrators.[132] Compare this tactic to traditional human rights investigations where a researcher typically needs to travel to the region to conduct interviews, collect court and other records, and visit crime scenes to collect evidence.

Remote sensing and AI could help solve a major challenge in human rights monitoring by giving civil society organizations and external actors access to territories that on the ground researchers cannot access, such as a conflict zone or closed country. The inaccessibility or high risks associated with accessing these situations to track violations is a major hurdle in human rights monitoring and reporting.[133] Technology that allows organizations who might not have regional access greater access into these jurisdictions to collect information could have monumental impacts. Human Rights Watch emphasized the importance of AI tools that can analyze sensitive datasets where data sets might contain sensitive or classified information, such as forensic photographs or personal information.[134]

This case study provides a direct example of an adversarial use of AI technology in the international human rights law context. Human Rights Watch, a civil society organization, is monitoring human rights abuses and then using that information to hold state actors accountable. AI can be used as a tool to

---

[129] Isha Salian, *AI in the Sky Aids Feet on the Ground Spotting Human Rights*, NVIDIA (Apr. 4, 2019), https://perma.cc/J8TE-BH27.

[130] *Id.*

[131] *Id.*

[132] *Id.*

[133] *Id.* ("'We can't document it if we can't get there,' said Josh Lyons, director of geospatial analysis at Human Rights Watch. 'If the people are in hiding or they're dead, there's no way to document that case.'").

[134] *Id.*

boost transparency and accountability by shifting the locus of control away from the state entirely.

F. Case Study: Using Machine Learning to Track Abuse Against Women on Twitter

In 2018, Amnesty International and Element AI used machine learning to understand online abuse against women in the United States and United Kingdom.[135] They surveyed 778 journalists and politicians to design a large enough dataset of tweets.[136] They then had over 6,500 volunteers analyze over 280,000 unique tweets to tag them for abusive or problematic content, including asking them to tag for "misogynistic, homophobic or racist abuse or other types of violence."[137] Once all were tagged, three experts took a random sample of 1,000 tweets to assess and validate the work of the crowdsourced labeling.[138] They then expanded the findings, and Element AI was able to design an abuse analysis "Troll Patrol" report,[139] concluding in their report that 7.1% of the tweets directed at these women (or 1.1 million tweets a year) fall into the "problematic" or "abusive" category.[140]

The AI was also able to extrapolate findings based on demographic markers. It found that, when compared to white women, women of color were 34% more likely to experience abusive or problematic tweets[141] and, specifically, Black women were 84% more likely.[142] There was little distinction by political affiliation—both liberal and conservative women were targeted. In

---

[135] *Troll Patrol Findings: Using Crowdsourcing, Data Science, & Machine Learning to Measure Violence and Abuse Against Women on Twitter*, AMNESTY INT'L, https://perma.cc/76ND-GJLE; *see also Toxic Twitter—A Toxic Place for Women*: *Chapter 1*, AMNESTY INT'L, https://perma.cc/7GGU-WAAV.

[136] *Id.*

[137] See *id.* for methodology ("The volunteers were shown an anonymized tweet mentioning one of the women in our study, then were asked simple questions about whether the tweets were abusive or problematic, and if so, whether they revealed misogynistic, homophobic or racist abuse, or other types of violence. Each tweet was analyzed by multiple people. The volunteers were given a tutorial and definitions and examples of abusive and problematic content, as well as an online forum where they could discuss the tweets with each other and with Amnesty International's researchers.").

[138] *Id.*

[139] *Id.*

[140] *Id.*

[141] *Id.*

[142] *Id.* While approximately 6.7% of tweets received by white women were abusive or problematic. *Id.* By contrast, black women received 60% more problematic tweets and 84% more abusive tweets (and of the abusive tweets, black women received 70% more racist tweets than white women). *Id.*

evaluating performance of the tool, Amnesty and Element AI found that the model was comparable to that of a regular digital volunteer, but fell short when compared to experts: "the AI was able to correctly identify 2 in every 14 tweets as abusive or problematic in comparison to experts who identified 1 in every 14 tweets as abusive or problematic."[143] More broadly, it has been widely documented that the use of AI in content moderation has caused disparate harmful impacts to vulnerable populations.[144] One example is content moderation of Arabic languages on Facebook and Instagram. In 2022, a member of Meta's Oversight Board, Rachel Wolbers, acknowledged the harms (and embedded biases) in content moderation: "we have specifically called out [Meta] to produce a lot more transparency around their content moderation and Arabic and potential biases that the company may be building into their content moderation Arabic."[145]

Even with its shortcomings, Amnesty developed a list of recommendations both for Twitter and for states based on the findings from the study. Recommendations to Twitter include: (1) publish meaningful data on how they handle violence and abuse on their platform; (2) make reporting easier and more transparent, "ensuring that decisions to restrict content are consistent with international human rights law and standards, are transparent, and allow for effective appeal"; (3) clarify how reports of abuse are dealt with and moderators deployed; and (4) improve security, privacy, and other safety risks or features.[146] Recommendations to states include: (1) adopt legislation to combat abuse of women online; (2) invest in programs to better educate and train law enforcement on the issue; (3) educate the public about abuse online and promote gender equality more broadly; and (4) invest in publicly available services or programs for women who have experienced abuse online.[147]

Amnesty specifically uses international human rights law and the obligations of states and private actors as a way to push for accountability and

---

[143] *See id.* ("While it is far from perfect, the model has advanced the state of the art compared to existing models and on some metrics, achieves results comparable to our digital volunteers at predicting abuse.").

[144] *See, e.g.*, Eugenia Siapera, *AI Content Moderation, Racism and (de)Coloniality*, 4 INT'L J. BULLYING PREVENTION 55 (2022), https://perma.cc/SE9V-BCE6; Kyle Wiggers, *How Bias Creeps Into the AI Designed to Detect Toxicity*, VENTUREBEAT (Dec. 9, 2021), https://perma.cc/VE66-ULC5; *There Isn't Enough Moderation in Arabic and Non-English Languages, Meta Oversight Board's Head of Global Engagement Tells Forum in Dubai*, ARAB NEWS (May 17, 2022), https://perma.cc/8L2H-2S8Z.

[145] *There Isn't Enough Moderation*, *supra* note 144.

[146] *Toxic Twitter—The Solution*: *Chapter 8*, AMNESTY INT'L, https://perma.cc/BZW2-9TWR.

[147] *Id.*

policy change based on the information that the AI and machine learning tools were able to identify and report on.

More broadly, this case study provides a salient example of how AI might be used to ascertain the status of women within a specific domain, providing detailed insight into rights enjoyment. For example, tracking abuse of women on Twitter serves as a reflection of that country's public discourse vis-a-vis women's rights. Harms related to bias are discussed in greater length in Part IV.

G. Case Study: AI as a Tool for Expanding Language Access: Translation Services and Multilingual Chatbots

Though not directly related to monitoring and reporting on rights, this case study is an example of the potential AI has to improve access to language and translation services in real time, which is of central importance to international human rights. AI-based speech-to-text[148] and translation services can be used in a variety of ways in the human rights system. These services have the potential to "greatly increase the scale of processing audio, video, and text-based foreign language information."[149]

One example of this potential is the Norwegian Refugee Council's use of chatbots to assist Venezuelan migrants in Colombia with learning their rights according to current immigration policies and laws.[150] The chatbots use AI and machine learning technologies to engage with migrants and incorporate real-time translation services. Another example is the development and deployment of an application called "Dr. Tania" that helps Indonesian farmers identify and treat crop disease.[151] The smart phone app allows users to interact with an AI driven chatbot that uses deep learning AI as every new photo is added to increase its accuracy of diagnoses. A final example is Masakhane, a project dedicated to strengthening natural language processing of native African languages.[152] To be effective, the AI needs to be fed enough training data to produce accurate results. One of the main challenges when it comes to African

---

[148] LITTMAN ET AL., *supra* note 76, at 12, 34.

[149] HOROWITZ, *supra* note 78.

[150] Leila Toplic, *AI in the Humanitarian Sector*, NETHOPE (Oct. 6, 2020), https://perma.cc/R9Z5-S652; *see also* David Felipe Garcia Herrera, *Four Things You Should Know About Venezuelans in Colombia*, NORWEGIAN REFUGEE COUNCIL (June 15, 2021), https://perma.cc/UDE4-TSQP.

[151] Leander Jones, *Dr Tania: An Indonesian AI Chatbot Helps Farmers Identify and Treat Crop Disease*, RESET (Aug. 6, 2020), https://perma.cc/K2E2-AGSQ.

[152] Kate Cashman, *Masakhane: Using AI to Bring African Languages Into the Global Conversation*, RESET (June 2, 2020), https://perma.cc/2QSQ-K7T8.

languages is that language data has been lacking, scattered, or not publicly available. More than one hundred researchers from across the continent are collaborating to crowdsource data and develop the algorithm.

In an international context, the capacity of AI technologies to translate in real time and to bridge language barriers is salient to rights monitoring and has the potential of being incredibly impactful. For example, this type of AI could be deployed to conduct real-time AI led interviews of victims of rights abuses in multiple languages or used to compile and make sense of data from multiple languages.

H. AI as a Tool for Mitigating Trauma Exposure and Mental Health Consequences in Human Rights Workers

Each of these case studies also raises interesting questions about sensitive data and the use of AI to shield human rights workers from traumatic information by shifting some of the burden to machine learning. If AI can take a first pass at processing traumatizing information, this would reduce the amount of traumatizing information human rights workers would have to encounter.

Data sensitivity is a major issue across human rights monitoring and reporting since it includes frequent and significant exposure to major human rights violations. Collecting, analyzing, and synthesizing information about human rights violations inevitably includes interacting with highly graphic and sensitive material. For example, it often incorporates data, images, and narratives related to torture, sexual violence, killings, starvation, property destruction, trafficking, extreme violence, and incidents involving women and children. Processing and handling this type of information can be traumatizing for human rights workers who are exposed to violence and trauma.[153] One of the additional benefits of having AI technologies processing this type of data is that it can shield human rights officials from extensive exposure and desensitization to this traumatizing information.

A 2018 study conducted by professors at Columbia and NYU found that there are strong correlations between human rights workers' exposure to trauma and the likelihood that they will develop symptoms of PTSD, depression,

---

[153] Margaret Satterhwaite, *Evidence of Trauma: The Impact of Human Rights Work on Advocates*, OPEN GLOBAL RTS. (Apr. 7, 2017), https://perma.cc/SQE4-6S2F.

and other mental health consequences like burnout.[154] The study adopts a multidisciplinary approach with academics and practitioners from both the human rights space and field of psychology.[155] They found that the human rights workers in the study had significant exposure to trauma (both in terms of the tasks undertaken and their own direct exposure as victims).[156] With regard to exposure and frequency, 89.3% conducted interviews with witnesses, 78.9% witnessed violations of basic needs, 63.3% visited sites of violations, 34.4% witnessed violence, roughly 20% experienced threats that they would be taken hostage, beaten, or assaulted or were arrested by the government, and 6.4% were actually taken hostage, beaten, or assaulted.[157] All but two of the tasks that were assessed as part of the study (litigation and providing medical care) were correlated with PTSD severity.

Reducing human rights workers' exposure to these types of traumatizing information may have benefits. A more complicated question is what types of information AI can and should process and what information requires a human being (for example, would a victim be harmed by having to do an intake interview with an AI bot rather than a human being?). This will be discussed at length in Part IV.

I. Patterns of Positive Use Cases

Now that we have examined case studies and articulated beneficial uses of AI, we can take a step back and examine the patterns that have emerged. Broadly, AI can be used in the following ways in the context of international human rights law:

- To boost efficiencies and productivity by making data collection, processing, and analysis faster and more effective by weeding through and make sense of massive amounts of data and information in an instant;
- To compile information and draft reports;
- To better forecast and predict trends which can support organizations and states in strategic planning efforts or resource allocation;

---

[154] Knuckey et al., *Trauma, Depression, and Burnout in the Human Rights Field: Identifying Barriers and Pathways to Resilient Advocacy*, 49 COLUM. HUM. RTS. L. REV. 267 (2018), https://perma.cc/355G-4VF3.
[155] *Id.* at 270.
[156] *Id.* at 303-04.
[157] *Id.* at 322.

- To expand organizations or states' capacity to make sense of information and data by providing a deeper understanding of data across rights and across demographic markers;
- To allow for better tracking of information over time;
- To disaggregate data to better understand how rights enjoyment (or violations) disproportionately impact differing communities, groups, and subsets of populations;
- To solve duplication issues, to scour through massive amounts of open-source data, to determine gaps in data;
- To validate and authenticate reports and data by comparing data sets against one another.

Ultimately, AI can be deployed to better monitor and report on rights both by civil society organizations and by states immediately and over time. In addition, these reports can be used in adversarial ways to hold governments accountable when they violate rights.

Even still, with each deployment and use of AI, ethical issues arise, limitations exist, and there is a risk of additional harm being introduced or exacerbated.

III. LIMITATIONS AND RISKS

There is a great deal of fear when it comes to AI. In September 2021, U.N. Human High Commissioner for Human Rights, Michelle Bachelet, gave an impassioned speech to the Council of Europe's Committee on Legal Affairs and Human Rights urging states to stop the use of AI until appropriate safeguards can be put into place to prevent human rights violations.[158] She said, "[w]e cannot afford to continue playing catch-up regarding AI—allowing its use with limited or no boundaries or oversight and dealing with the almost inevitable human rights consequences after the fact. The power of AI to serve people is undeniable, but so is AI's ability to feed human rights violations at an enormous scale with virtually no visibility. Action is needed now to put human rights guardrails on the use of AI, for the good of all of us."[159]

The risk of misusing AI technologies in ways that violate basic human rights is not unique to the private sector. In fact, from a practical and policy

---

[158] Michael Dziedzic, *Urgent Action Needed Over Artificial Intelligence Risks to Human Rights*, UN NEWS (Sept. 15, 2021), https://perma.cc/58GF-U4SF.
[159] *Id.*

standpoint, there is potentially an even larger risk of misuse by states when states are positioned as both the regulator and the user of these technologies. States are the primary violators of human rights, making it critical that any system that allows (and encourages) states to use AI and machine learning technologies is on high alert for potential risks associated with their use. There is always a risk that states will end up turning around and using the same technology to further human rights violations or evade responsibility and accountability.

As discussed in Part I, many legal scholars have evaluated the risks and harms of AI broadly. AI, like all new technology, invariably implicates new legal questions and raises major human rights concerns. As noted previously, scholars have looked at, "lack of algorithmic transparency; cybersecurity vulnerabilities; unfairness, bias and discrimination; lack of contestability; legal personhood issues; intellectual property issues; adverse effects on workers; privacy and data protection issues; liability for damage; and lack of accountability for harms."[160] And while these concerns about AI's impact on human rights exist within and are relevant to the international human rights space, this Note focuses more specifically on identifying the limitations and harms of AI's use in the human rights system—largely in the context of monitoring and reporting on rights.

This section proceeds in the following way: it begins by identifying and exploring the limitations of AI's use in human rights monitoring and reporting systems, largely focusing on data completeness, issues around accuracy, data analysis and decision-making, and AI's relationship to in-the-field groundwork. It then shifts to looking at unique harms that may arise when AI is used in the international human rights law space—an arena where the consequences of these harms are heightened because of the risk of perpetuating and furthering human rights abuses.

A. The Data Problem: Exploring the Limitations of AI in the International Human Rights Space

The most obvious limitation with AI and machine learning in the human rights space is access to data and data completeness. Often, the most limiting factor for the use AI is not the creation of the AI tool or its algorithm, but is

[160] Rodrigues, *supra* note 68.

instead not having enough data or not having the right data.[161] AI relies on having a great deal of data to "learn" from—algorithms require massive quantities of data in order to turn out accurate results and provide meaningful outputs.[162] This is increasingly true if the desired output is more complicated.[163] If the data has not been collected or does not exist, the AI cannot learn.[164] If the data is incomplete, the AI may learn things incorrectly.

This problem may be especially salient in the human rights monitoring and reporting context where nonstate and state actors alike may not have ready access to the type of data that would be necessary to teach the AI. In both cases, there are major limitations related to resource constraints. The paradox is that while AI would eventually ease some of those resource constraints should it be used to streamline processes and make monitoring and reporting easier, the resource constraints that currently exist may prevent AI from being developed and deployed in the first place. This is especially true for state actors who have less incentive to monitor rights closely.

The international human rights system relies on data primarily from three sources: international institutions (like the World Bank, IMF, etc.), international human rights NGOs (like Amnesty International and Human Rights Watch), and state actors who self-report.[165] There are challenges with data completeness and accuracy across the board.

In many ways, international human rights organizations have been the primary source of information about rights violations.[166] However, they have been critiqued as "problematically fragmented, hierarchical, non-collaborative, and excessively shaped by organizational self-interest" ultimately resulting in a lack of data sharing.[167] This is compounded by their limited access to data collected by the state or limitations related to accessing data sources. Many of the case studies demonstrate how the organizations can try and solve this problem – one way has been appealing to the public, crowdsourcing information and data. Another has involved using new types of technologies to

---

[161] LITTMAN ET AL., *supra* note 76, at 9.
[162] Matthew Steward, *The Limitations of Machine Learning*, TOWARD DATA SCI. (July 29, 2019), https://perma.cc/9UNN-ZHMS.
[163] *Id.*
[164] McKendrick, *supra* note 61; *see also* DOMINGOS, *supra* note 59.
[165] *See generally,* OFF. OF THE U.N. HIGH COMM'R, MANUAL ON HUMAN RIGHTS MONITORING ch. 13 (2011), https://perma.cc/M6CG-73L5.
[166] Philip Alston & Colin Gillespie, *Global Human Rights Monitoring, New Technologies, and the Politics of Information*, 23 EUROPEAN J. INT'L L. 1089 (2012).
[167] *Id.*

access data sources, such as the use of sensors and satellite imagery, rather than relying on data from state actors. In addition, civil society organizations may be able to use open-source information on the internet to collect more data. An example of this is the use of open-source data from 18 sources used in the DRC's displacement forecasts. Even in that example, though, they highlight their own "data problem," noting that the "data on forced displacement depends wholly on the numbers from UNHCR and IDMC."[168] While these are highly reliable sources, they may still leave out some that have been displaced in 2021.[169] They note that the data can be imbalanced, meaning that for certain geographies or for certain indicators, more or less data may be available and thus more or less prone to inaccuracy. By way of example, they note that data is more readily available for labor statistics than for governance and violence. They also find that there are often delays in obtaining data from institutional providers, and that some target variables may just be missing altogether.[170] They came up with their own methodology for addressing the data gaps, including cross validating over a longer period of time.[171]

As part of their obligation to the nine international human rights treaties, states are required to regularly monitor and report on their progress.[172] There would be an assumption then that data collection is taking place and could serve as a foundation for AI development. Each of the treaties above also envisage that the state undertakes the regular monitoring of rights, collecting massive amounts of data and information that makes its way into the reports.[173] However, for a state with limited resources, it is not only burdensome to compile and submit the reports, but it may be exceptionally challenging to regularly collect and publish data. As a result, data is often incomplete, inconsistent, and unreliable.[174] As demonstrated, the monitoring and reporting requirements that states undertake when they sign on to international treaties are extensive—and as a result, the fragmented system has been described as "inadequate, ineffective, and [in] crisis."[175] There are a variety of reasons data may be incomplete or may not exist. For example, it

---

[168] FEBRUARY 2022 FORECAST, *supra* note 100, at 50.
[169] *Id.* at 50-51.
[170] *Id.* at 51.
[171] *Id.*
[172] OFF. OF THE U.N. HIGH COMM'R FOR HUM. RTS., *supra* note 35, at 21-22.
[173] *Id.*
[174] Lindsay Ferris & Noel Isama, *Making the Case for Open Human Rights Data*, SUNLIGHT FOUND. (Dec. 21, 2015, 4:25 PM), https://perma.cc/MG8F-N9CP.
[175] Creamer & Simmons, *supra* note 43, at 18.

could be because the state hasn't collected it because data collection mechanisms do not exist or it could be because the data is hard to capture and quantify (for example, while extrajudicial killings might be easy to quantify, something like freedom of movement within a state may be more challenging).[176]

In the context of using AI as a tool for monitoring and reporting, the data paradox again presents itself. It is the same states who lack the resources to adequately monitor and report that could benefit the most from tools like AI that would make monitoring and reporting more feasible and meaningful.[177] It is no surprise that "[g]ross domestic product per capita is strongly associated with the likelihood of reporting, suggesting that wealthier states are better able to bear the costs of compiling legislation, collecting data, and studying outcomes."[178] Disproportionate impact on states that are already resource constrained or do not have robust internal reporting machinery in place results in many states either submitting the same report to each treaty body without treaty specific data.[179] This has obvious implications for the realization of human rights broadly, but also creates a barrier for states to adopt AI technologies. Once a state gets behind on reporting, it is increasingly challenging to catch up or track progress over time because of missing or inadequate data from previous cycles.[180] If the treaty body can't access the data in real time, they may not be able to make appropriate recommendations or ensure that the state is meeting its obligations.[181]

Therefore, instead of thinking of AI as a full solution, it should instead be thought of as one tool that states can use to supplement and bolster their existing strategies and methodologies. It is important that mechanisms and systems be put in place to ensure data and outcomes are validated and authenticated.[182]

---

[176] Ferris & Isama, *supra* note 174.

[177] Creamer & Simmons, *supra* note 43, at 16.

[178] *Id.* at 19.

[179] INT'L COUNCIL ON HUM. RTS. POL'Y, HUMAN RIGHTS STANDARDS: LEARNING FROM EXPERIENCE (2006), https://perma.cc/UDZ9-D7SE.

[180] *Id.*

[181] *Id.*

[182] *See* Noel Isama, *The Importance of Accurate, Verified Human Rights Data*, SUNLIGHT FOUND. (Mar. 14, 2016, 4:03 PM), https://perma.cc/2RXG-WVEB ("National statistics agencies must be trusted sources for the kinds of data that is crucial for more effective monitoring of human rights. And not only should the data be trusted, but how governments reach their conclusions is equally important: Without accompanying open data with transparent methodologies —

B. Decision-Making and Data Analysis

Current AI technologies are limited in their ability to evaluate and make sense of qualitative information that helps explain or contextualize the realities of human rights violations on the ground. AI might be able to tell us what is happening, but it can't explain why. Cultural relativism is the idea that human rights must take into account cultural differences—and its principles are embedded into many of the international human rights law treaties. International human rights law specifically allows for flexibility for different states to approach and ensure rights in ways that account for culture nuances. AI technologies might be limited in its ability to pick up on these nuances or take into account cultural differences. There may be context and narrative necessary to understand the context in which human rights are being abused and to identify the need for solutions. In addition, as discussed previously, there are many sensitivities that come with doing human rights work and interacting with victims of human rights abuses. Therefore, there may be instances where it is not appropriate to use AI in place of human beings.

It may be helpful to illustrate the line between what AI can and cannot do through a concrete example. In 2010, the U.N. Secretary General directed the OHCHR to lead a mapping exercise to document and quantify human rights violations in the Democratic Republic of Congo (DRC) over a ten-year period (1993-2003).[183] It was done in partnership with the Congolese authorities[184] and civil society with the goal of enabling victims to obtain justice by uncovering human rights violations and supporting the Congolese government in coming up with different ways to achieve justice, redress harms, and make reforms.[185] The mapping exercise took ten months and utilized twenty full-time human rights officers.[186]

---

even absent clear reasons to suspect manipulation of national statistics — true legitimacy is impossible, especially when statistics differ from prevailing public sentiment.").

[183] OFF. OF THE U.N. HIGH COMM'R FOR HUM. RTS., REPORT OF THE MAPPING EXERCISE DOCUMENTING THE MOST SERIOUS VIOLATIONS OF HUMAN RIGHTS AND INTERNATIONAL HUMANITARIAN LAW COMMITTED WITHIN THE TERRITORY OF THE DEMOCRATIC REPUBLIC OF THE CONGO BETWEEN MARCH 1993 AND JUNE 2003 (2010), https://perma.cc/8JY6-2EW8.

[184] *Id.* at 3-4 ("The Congolese Government has expressed its support for the Mapping Exercise on a number of occasions, notably in the statement delivered by the Minister for Human Rights at the Special session of the Human Rights Council on the human rights situation in the East of the DRC in November 2008 and in various meetings between the Chief of the Mapping Exercise and the Justice and Human Rights Ministers.").

[185] *Id.* at 3.

[186] *Id.*

Based on the information collected from varying sources (data, surveys, witness interviews, and consultation with field experts), they created and populated a database listing each grave violation of international humanitarian law.[187] Each included: "a description of the violation(s), their nature and location in time and space, the victim(s) and their approximate number and the—often armed—group(s) to which the perpetrators belong(ed)."[188] Because of major time and resource constraints, the mapping exercise focused on collecting only basic information and reported only the most grave human rights violations.[189] Once that phase of the project was complete, an in-field team was deployed to the DRC for six-months to verify and corroborate the information that was obtained in the initial phase, and to report on previously undocumented violations (for example, they had identified gaps in the data and focused their in-field operations on areas where information was missing or lacking).[190] Phase 3 focused on cross checking the data against other independent sources that documented the same violations in order to validate and authenticate the information.[191]

The initial phase of the project exemplifies the type of reporting that could have greatly benefited from the use of AI technologies. AI is well equipped to analyze vast amounts of data and information and index accordingly. This type of technology is similar to the types of AI technologies used in the Darfur[192] and Displacement[193] case studies where machine learning was used to collect information, index based on a number of indicators. Like the Death Penalty case study,[194] the AI could then populate a database with the specified information listed above. Rather than having a person go back and identify the major gaps in data ahead of Phase 2, the AI/ML technology could accomplish the same objective in a more streamlined, efficient, and much faster way. Similarly, the final phase focused on cross-checking and authenticating information, which is also something that could be supplemented with AI technologies.

This example also highlights an important limitation of the AI technology. While existing AI technologies are capable of accomplishing most, if not all, of

---

[187] *Id.* at 36-37.
[188] *Id.* at 36.
[189] *Id.* at 35-36.
[190] *Id.* at 36.
[191] *Id.* at 37, 44.
[192] Cornebise et al., *supra* note 80.
[193] 2021 FORECAST, *supra* note 99; FEBRUARY 2022 FORECAST, *supra* note 100; JULY 2022 FORECAST, *supra* note 100.
[194] AMNESTY INT'L & ELEMENT AI, *supra* note 114.

the tasks in Phase 1, AI is less equipped to support the types of methodologies deployed in Phase 2, which relied more extensively on in-person field research and manual, non-traditional data collection. Instead of thinking of AI as a full solution, it should instead be thought of as one tool that states can use in addition to existing strategies and methodologies. Here, existing AI technologies are capable of making Phase 1 more efficient (since it is capable of dealing with massive amounts of data at once) and requiring less staff resource time (perhaps only needing staff to check the data after). Ideally, use of AI in the first phase would enable states to reallocate their resources to the types of monitoring and reporting AI is not yet capable of.

Generally speaking, states could use this AI-driven methodology (collecting, synthesizing, indexing, categorizing, and aggregating data) to meet many of their international monitoring and reporting requirements—both common requirements and treaty-specific requirements. The common core requirements largely consist of taking large amounts of national, regional, and local data on a variety of different rights issues and aggregating them to report on overall implementation, enjoyment, and violation of rights. Many of the treaty requirements follow suit. Reporting related to discrimination also requires that data be disaggregated to understand whether and to what extent states are ensuring or violating rights with distinction to race, color, sex, language, religion, political or other opinion, national or social origin, etc. In other words, how are certain demographic subsets of the population disproportionately impacted by rights violations? Existing AI technologies already have these capabilities and, as the case studies indicated, are operating at high levels of accuracy. Ideally, states would use this capability for the rights where there is little need to ascertain or interpret the data,[195] and as a result, could focus their limited resources on rights that require interpretative judgment in reporting.[196]

---

[195] For example, under Article 24, the right to health and health services and the right to freedom from discrimination, the Convention on the Rights of the Child requires states to provide data disaggregated by age, gender, race, minority membership, religion, rural/urban, etc. It includes rates of infant mortality, proportion of children with low birth weight, proportion of children born in hospitals, etc. For the most part, this is the type of data that AI could more easily collect, synthesize and disaggregate because there is little need for interpretive judgment. G.A. Res 44/25, Convention on the Rights of the Child, art. 24 (Nov. 20, 1989).

[196] By contrast, Article 5 of the Convention on the Elimination of All Forms of Discrimination Against Women requires states to take "all appropriate measures . . . to modify the social and cultural patterns of conduct of men and women with the view to achieving the

C. Identifying the Potential Harms of AI and the Risk of Perpetuating and Furthering Human Rights Abuses

1. Repurposing of Data to Facilitate Harms

There is a larger risk of misuse by states when states are positioned as both the regulator and the user of these technologies. There is always a risk that states will end up turning around and using the same technology to further human rights violations or evade responsibility and accountability. This is especially true where there is government instability—one group in power may use data collected during a previous regime to perpetuate harm once they come into power. Take, for example, the Rwandan genocide. The Hutu government targeted Tutsis based on their minority status. An AI technology that tracks information related to ethnicity status could have been used by the Hutu government as a tool to target and kill civilians. States are the primary violators of human rights, and as such their use of AI and machine learning technologies should put human rights groups on high alert of potential risks associated with its use.

Collecting information about human rights violations might open up already vulnerable communities to future harms. For example, the Darfur village destruction case study highlights the need to mitigate risk and build protections into the AI in order to protect vulnerable groups.[197] This issue is specific to international human rights monitoring and reporting, but not specific to AI. It is salient regardless of the type of data collection or reporting. However, because AI expands the scope of information so greatly, it does lead to heightened risks. In that case, tracking and potentially divulging the precise location of vulnerable villages would open those communities up to potential targeting by insurgent groups.

Another example is in the context of migration flows. In September 2022, Statewatch along with more than a dozen other organizations and leaders, penned a letter to the EU ITFlows Consortium urging them to stop the use of EUMigraTool, an AI tool designed to track migration flow in Europe.[198] They

---

elimination of prejudices . . . which are based on the idea of inferiority or superiority . . . or non-stereotyped roles for men and women." This is an example of the type of right that relies far less on simple facts or data and requires more substantive interpretation and contextual understanding. G.A. Res. 34/180, Convention on the Elimination of All Forms of Discrimination Against Women, art. 5 (Dec. 18, 1979).

[197] Cornebise et. al., *supra* note 80.

[198] *Using AI Tools to Predict Migration Flows Will Lead to Human Rights Abuses and Must Cease*, STATEWATCH (Sept. 27, 2022), https://perma.cc/8CET-K5Q5.

flagged the "potential, and probable" misuse of the tool, particularly by authorities at border crossings, to perpetuate harms against migrants.[199] In the letter, they highlight an urgent, serious risk associated with the current use of the AI tool (and AI more broadly) in the migrant context: "The project offers a techno-solutionist answer to migration responses without addressing the structurally oppressive dimension of EU migration policies . . . . By engaging in this project, the Consortium will legitimize the idea that migration is a problem and that it can be fixed via technical solutions, discharging institutions of their responsibilities regarding the deterioration of fundamental rights at the EU borders and within EU member states."[200]

2. Bias

In general, AI relies on data that is either generated by humans or is the product of a system that was designed by humans. As a result, whatever biases those humans had when they generated the data or created the system, make their way into the AI. At the same time, AI machine learning continues to learn based on the information that is provided to it. Therefore, AI continues to amplify biases. This is aptly summarized by scholars surveying bias in the field of AI: "Algorithms are part of existing (biased) institutions and structures, but they may also amplify or introduce bias as they favor those phenomena and aspects of human behavior that are easily quantifiable over those which are hard or even impossible to measure."[201] Biased data and AI systems pose major risks to human rights, and already the impacts of bias in AI are being seen in various studies[202] are and discussed below. Challenges associated with bias are amplified in the international human rights system, where accurate data and models are absolutely necessary[203] and where these types of biases would be detrimental because rights violations are more concentrated among minorities.

---

[199] *Id.*

[200] *Id.*

[201] Ntousti et al., *Bias in Data-Driven Artificial Intelligence Systems – An Introductory Survey*, 10 WIREs DATA MINING & KNOWLEDGE DISCOVERY 1356 (Feb. 3, 2020), https://perma.cc/ZM63-ZS7W.

[202] *See, e.g.*, Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. ON MACH. LEARNING RSCH. 1 (2018), https://perma.cc/QV9C-8AJU; Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://perma.cc/RS4S-3BWV.

[203] Meg Satterthwaite, *Human Rights Data Used the Wrong Way Can Be Misleading*, OPEN DEMOCRACY (Sept. 1, 2016), https://perma.cc/7WXX-X6GE.

As the U.N. High Commissioner for Human Rights Michelle Bachelet stated, "the risk of discrimination linked to AI-driven decisions—decisions that can change, define or damage human lives—is all too real."[204] International human rights monitoring and reporting relies on the detailed data in the reports to make key recommendations related to judicial and policy changes. Central to every treaty mechanism is the idea that rights should be realized without any distinction in relation to race, gender, national origin, language, religion, etc. Both the core reporting requirements and the detailed treaty requirements specifically request disaggregated information. For the realization of these rights, it is essential states understand how laws, policy and programs impact different groups in order to cure any discrimination. Data that is compromised by bias and discrimination can be detrimental to the realization of rights. For AI to be a viable and effective tool for monitoring and reporting of rights, AI technologies will need to focus on identifying and mitigating bias in the data.

A salient example of this is the Dutch government's use of an unregulated AI algorithm that tax authorities created and deployed in order to detect fraud in childcare benefit applications.[205] The design of the AI tool was fraught with racial and ethnically discriminatory bias. As a result, tens of thousands of families were incorrectly accused of tax fraud.[206] Many of these parents and families were from low-income households and/or are members of ethnic minorities.[207] In designing the tool, tax authorities incorporated nationality status as a risk factor with Dutch citizens receiving lower risk-scores and non-Dutch nationals receiving higher risk-scores.[208] The algorithm learned that non-nationality status was higher risk and reproduced this bias over and over again resulting in a continuous loop of discrimination and bias.[209] They note that there was "no meaningful human oversight"[210] to check or fix the algorithm as non-Dutch nationals continued to be incorrectly flagged as more likely to be fraudulent than their Dutch national counterparts.[211] This example illustrates how a design flaw can result in additional harms—particularly where basic

---

[204] Al Jazeera Staff, *UN Call for Moratorium on AI that Threatens Human Rights*, AL JAZEERA (Sept. 15, 2021), https://perma.cc/H3K2-VPJG.
[205] *Dutch Childcare Benefits Scandal an Urgent Wake-up Call to Ban Racist Algorithms*, AMNESTY INT'L (Oct. 25, 2021), https://perma.cc/ZSL5-XH2E.
[206] *Id.*
[207] *Id.*
[208] *Id.*
[209] *Id.*
[210] *Id.*
[211] *Id.*

rights are at stake. This is relevant to the international human rights monitoring and reporting system where states are being asked to disaggregate data on a number of demographic markers to report on discrimination. Should those markers be used incorrectly (here, the use of nationality should not have been flagged as an indicator for fraud), then that has the potential to have damning consequences in reporting on rights enjoyment or violations. For example, many of the international human rights law treaties hold different standards of obligations for citizens and non-citizens within their jurisdiction. AI incorrectly identifying individuals or communities as non-citizens in reporting could prevent bodies from holding states accountable to fulfill their obligations.

One of the reasons it went on for so many years unchecked is because the AI was a "black box" system—in other words, there was no accountability or oversight. As a result, Amnesty International, who uncovered the issue, developed a list of action items for the Dutch government and governments using these types of tools in other contexts.[212] This included two that are particularly important to the international human rights monitoring context: (1) "implementing a mandatory and binding human rights impact assessment before the use of such systems" and (2) "establish[ing] effective monitoring and oversight mechanisms for algorithmic systems in the public sector."[213] Another best practice is to allow mechanisms for impacted communities to complain, since they tend to see these kinds of adverse impacts first. In the human rights reporting system, these types of complaints can supplement state or NGO reports.

Another example is the use of AI in criminal justice sentencing using facial recognition technology. Many facial recognition tools have been found to be biased towards darker skinned individuals with significantly higher error rates for black individuals and people of color.[214] For example, one AI tool used for law enforcement purposes had an error rate of just 0.8% for white men, but a much, much higher error rate of 34.7% for black women.[215] Criminal sentencing is another area where racial bias seeps into AI-based decision-making. A 2019 study found that COMPAS, an AI tool being used to predict and evaluate risk in criminal sentencing has been found to flag black defendants as having a higher propensity for future crimes at almost double the rate of white defendants

---

[212] *Id.*

[213] *Id.*

[214] Morgan Livingston, *Preventing Racial Bias in Federal AI*, 16 J. SCI. POL'Y & GOVERNANCE 2 (May 2020).

[215] *Id.* (citing Buolamwini & Gebru, *supra* note 202).

(45% compared to 24% respectively).[216] This tool is currently being used in close to a dozen US states.[217]

Much like the Dutch example, the algorithm has bias baked in from the start, continues to learn from that bias, and creates a loop. It is easy to see how this type of example—the use of biased facial recognition technology—could be replicated in many other human rights monitoring settings, particularly those using machine visioning where the system uses analogue-to-digital conversion to capture and analyze visual information. Oftentimes, the bias comes from what images are used to train the algorithm. The use of satellite images and drone images to determine rights violations is an area that may be vulnerable to these same types of biases, especially if the technology is built off of systems that have different error rates for different races, genders, or ethnic groups. Imagine if an AI algorithm was trained to identify destruction in a desert and then was replicated and used to identify destruction in a forest. This raises issues of both bias and accuracy. In the same vein, imagine if an AI algorithm that has major performance failures in identifying black faces as opposed to white faces was deployed and used to track harms in a majority black nation. Bias seeps into initial algorithm development and only worsens over time if not corrected.

Issues around data bias in algorithmic learning are of particular salience in the context of human rights monitoring where those at highest risk of rights violations are often from the same communities that algorithms are biased against—women, ethnic minorities, and children.

3. Privacy and Data Protection

There is an inherent connection between the use of AI technologies and the right to privacy. Much literature focuses on the risks that AI poses in terms of digital surveillance and data privacy. Even the most law-abiding, human-rights-respecting, cautious states inevitably risk compromising their citizens' rights by having massive amounts of sensitive biometric or person-specific data centralized in one location or on one AI platform. Systems can be breached, data can be compromised, and sensitive information can make its way into the wrong hands.

---

[216] *Id.*; Angwin et al., *supra* note 202.
[217] Livingston, *supra* note 214.

In the context of the international human rights monitoring and reporting system, this becomes particularly harmful where state actors hold data in unstable regimes. There is a risk that the information and data collected with the initial goal of helping advance rights could then fall into the wrong hands should a regime change take place. It is important to weigh this risk as part of a risk assessment in using AI technologies.

A concrete example of this is the United States leaving behind sensitive data when withdrawing troops from Afghanistan in 2021, risking the possibility of Afghani Taliban members obtaining the data. During the war, the U.S. military collected and maintained key biometric data on Afghani citizens, such as fingerprints, iris scans, and facial images.[218] The technology was widely used, building an extensive biometric identification system.[219] The purpose of the technology was to allow U.S. soldiers to confirm whether or not someone was an ally and identify and track threats.[220] When the United States withdrew its troops, it failed to collect all of the devices—many were abandoned and left behind.[221] While the United States maintained that the data is not at risk and that the military took prudent steps to ensure it would not fall into the hands of the Taliban, human rights advocates called on international actors to intervene, warning that the Taliban might be able to hijack and use the data to identify and target individuals who worked with opposing forces.[222] On August 17, 2021, the Taliban seized the U.S. military biometric devices.[223] Thirty-six civil society organizations promptly signed a letter denouncing the use of digitized, searchable databases because of the risk and expressing their concerns that mandatory and centralized collection of extensive data was "always dangerous."[224] While there may be some salient arguments for the collection of biometric data, the risks posed by having personalized information stored in centralized technology banks far outweigh any potential benefits. The opportunity for misuse is too great and the consequences are too severe.

---

[218] Drew Harwell, *Ukraine is Scanning Faces of Dead Russians, then Contacting Their Mothers*, WASH. POST (Apr. 15, 2022), https://perma.cc/JKR6-6KPQ.

[219] FBI, *Mission Afghanistan: Biometrics* (Apr. 29, 2011), https://perma.cc/6ADQ-Y9PQ.

[220] April Glaser, *A U.S.-Built Biometric System Sparks Concerns for Aghans*, NBC NEWS (Aug. 31, 2021), https://perma.cc/GG9H-KB6T; *see also* ACCESS NOW ET AL., *Civil Society Calls on International Actors in Afghanistan to Secure Digital Identity and Biometric Data Immediately*, WHYID (Aug. 25, 2021), https://perma.cc/U2DS-Z8HM.

[221] *Id.*

[222] *Id.*

[223] Ken Klippenstein & Sara Sirota, *The Taliban Have Seized U.S. Military Biometric Devices*, INTERCEPT (Aug. 17, 2021), https://perma.cc/WD69-PLYX.

[224] ACCESS NOW ET AL., *supra* note 220.

An additional harm within this context is the anonymization of data by AI and the ease of re-identification, where anonymized data can be traced back to specific people's identities if consolidated with data from other sources.[225] To illustrate this point, the author uses the example of a computer science graduate student who was able to re-identify anonymized data of state employee hospital visits after the insurance company experienced a data breach. By matching the records with voter files, she was able to identify and piece together the intimate details of the then Massachusetts governor's hospital stay, diagnosis and prescriptions. She points out that technologies to re-identify data have advanced considerably since then—the re-identification of anonymized data is now far more advanced and quicker to deploy. AI promises that information collected as the machine learns will be anonymized.[226] There is a paradox between data privacy and data utility that creates a challenging dynamic in the context of AI and human rights. The more you strip down the data, the safer it is for the individual or group, but the less useful it is to the state or NGO trying to utilize the data. For example, to combat discrimination, all international human rights law treaties require states to provide detailed information about rights enjoyment disaggregated by demographic markers that help to track and report critical data related to discrimination. On one hand, disaggregated data risks violating vulnerable communities' data privacy because of the level of detail provided in the information; one such solution is to strip the data to provide privacy protections. On the other, the more detailed the data is, the more easily it can be used for human rights bodies to hold states accountable for discriminatory behavior. Though it is more of a big data concern, automation in this space introduces potential harms that are relevant to AI and the need to balance competing values in assessing whether AI should be deployed and to what extent.

IV. PROPOSED FRAMEWORK

In order for AI to be a viable, effective tool to support the realization of human rights largely depends on whether the proper constraints and regulations can be adopted to ensure the AI technology is used in morally sound ways.

---

[225] DANIELLE KEATS CITRON, THE FIGHT FOR PRIVACY: PROTECTING DIGNITY, IDENTITY, AND LOVE IN THE DIGITAL AGE 15-17 (2022).
[226] *Id.*

Based on the information gleaned from the case studies and exploration of limits and harms, this Note proposes a framework for assessing impact of AI in the international human rights monitoring and reporting context. Using a cost benefit analysis, practitioners can use this framework to determine within the human rights space: (1) areas where AI can be used with little concern; (2) areas where AI can be used with appropriate measures of caution and constraint; (3) high risk areas where the major costs (and limited constraints) might still be outweighed by the potential benefits; and (4) areas where the use of AI is never appropriate.

As indicated by the case studies, there are many examples of positive use cases of AI in international human rights law. With these major benefits and positive applications in mind, there are clear examples of issue areas where civil society or state actors should not be afraid to use AI in those cases. In fact, AI can provide major benefits. Even still, across the board, any deployment of new technologies that deal with human rights should be taken with appropriate caution and guided by a risk assessment. Understanding where to focus the use of AI can improve efficiency. And there are some areas that, at this point, are not appropriate for use of AI—where the risk of harm so greatly outweighs any benefits.

A. Factors for Evaluating Impacts

1. Evaluate the Actors Involved in Creating and Deploying the AI

It is critical to evaluate who is involved in both the creation of AI technology and the deployment or use of the AI and its outputs. Broadly, the greater the power of the user over those whose rights are at risk, the more potential there is for harm, and the more caution that should be taken when adopting its use.

Assessing the type of actor and their level of credibility can help in this analysis. This might include looking at whether the developer and user is a civil society organization, a regional or international organization, or an individual state. For example, a civil society organization (like Amnesty International or Human Right Watch) exerts less control over the parties whose rights are being evaluated. Because they maintain less power, the risks are consequently lower and there may be less need for external regulation by a third party. As a result, the use of AI may be more acceptable. Other considerations might include organizational reputation and credibility, relationship to the government they are accessing, and experience and expertise. By contrast, individual states

exercise direct control and jurisdiction over parties whose rights are at risk. Because they maintain higher levels of power over these parties, the risk of abuse or misuse is automatically higher and there may be more need for third party observation or regulation. Additional considerations might include government stability, regime type (democracy may weigh in favor of use, authoritarian regime may weigh against its use), and ongoing humanitarian crisis, war or other risk of instability. Regional and international organizations likely fall somewhere in between civil society use and state use. Considerations identified for each should be evaluated in relation to these types of organizations.

Another factor for consideration is what types of third-party private companies are being brought on to assist with the development of these technologies. Have they partnered with civil society or state actors in the past? Do they stand to benefit financially or politically by being involved? Do they have a reputation for being a legitimate and trustworthy business, including whether they have processes and procedures in place for protection of technology and data?

2. Evaluate How the AI Is Being Used Both in the Near Term and Identify Any Future Uses

As the case studies demonstrate, there are a broad range of ways in which AI can be used—understanding how AI will be used is critical for evaluating the benefits and harms that may accompany its use. Generally speaking, the simpler a task is, the less likely it is to be harmful. For example, simple data compilation or data analysis may carry very little risk. The question then becomes whether there are controls in place to ensure accuracy, completeness, etc. Tasks or uses that require some degree of lived human experience or some specific cultural knowledge or understanding may be riskier; they may only be appropriate with a proper level of constraint to ensure errors are not introduced and harms are not perpetuated.

In circumstances where there is little other way to obtain or evaluate information or data, the benefits of AI use may outweigh the costs or risks associated with its use. For example, instances where there is a complete lack of resources or situations where humans do not have access to the on-the-ground data may weigh in favor of using AI. Two examples where relevant questions came up were the Darfur village destruction and the use of thermal

data to monitor ethnic violence in Myanmar. In both instances, civil society organizations had little ability to collect information on the ground.

Areas where we should be most concerned about AI's use are in instances where technology is being developed and deployed to make decisions for humans. This includes instances where AI is being used to make determinations of people's fate or in the context of value judgments or the development of new policies. Likewise, there are distinct situations where a human element cannot be replicated or where the value of having a human doing the work of the AI is more beneficial. This might include interactions with victims, instances that require relationship building, or complicated qualitative assessments that require nuanced information. Here, it may not be appropriate to use AI if it can be avoided.

In analyzing this factor, it is also necessary to ask what is lost by having AI conduct the task in lieu of a person. Again, the ChatGPT example is helpful in thinking through the type of questions that should be asked to do this analysis.

3. Evaluate How the AI Will Be Designed and Developed

AI is only as effective as it is carefully developed. As discussed previously, AI learns from massive data inputs and the data itself impacts the ability of the AI to be effective, free from error, bias or other harms. In assessing how the AI is being designed and developed, relevant questions related to data include:

- Does enough data exist to adequately teach the algorithm?
- If there is not enough data, is there a way to obtain data? This might include crowdsourcing like in the Decode Darfur or Amnesty Twitter examples.
- Is the data complete and accurate?
- Is data being derived from several sources? Are the sources reliable?
- Is the data or AI algorithm vulnerable to biases? Is it possible to counteract or mitigate these biases?

Each question should be thoughtfully answered and accounted for in any cost-benefit analysis.

It may also be salient to think about the type of tool being developed, as different types of AI necessitate different considerations and present different harms. For example, deep learning requires additional scrutiny with regard to data inputs; is there enough data to support neural networks? Similarly, natural language processing may require very specific types of data that cannot be easily obtained or created. An example of this is Masakhane, the project

dedicated to strengthening natural language processing of native African languages where there are massive gaps in information.

Finally, any analysis should take into account what resources are required to build out the AI and whether the financial and human investment are worthwhile. This includes assessing what resources it would take to build and train the AI and whether those resources are available. If the AI requires extensive resources and those resources are largely unavailable, it might not be a worthwhile endeavor. If the AI requires only limited resources, but there are no to little resources available, it may still be worthwhile. If the AI requires only limited resources and resources are available, it may be worthwhile. If building the AI requires extensive resources and the resources are available, it may be worthwhile so long as it does not take resources from other worthwhile endeavors. In doing this analysis, it is also worthwhile to think through how the AI might be used in the long term—can it be used in other ways or on other projects? This may weigh in favor of investing resources in its development.

4. To What Extent Individual Rights Are at Risk

In the context of human rights, it is extremely important to evaluate the extent to which individual rights are at risk when AI is used. Part of this analysis necessarily requires one to evaluate where the data is coming from and how it is being protected. As discussed previously, data privacy is closely connected to AI's use. In the Displacement Forecasting example, data came from reputable sources such as the U.N. International Displacement Monitoring Organization, the IMF, and the World Bank, and was not specific to any one person. By contrast, the DRC Mapping Exercise explicitly collected information about perpetrators and victims, including detailed information about their locations, their organizational associations, and more. The former is probably less risky insofar as the information is reliable, is largely publicly available, and there are constraints in place to protect victims. The latter is riskier because the information includes private details about vulnerable communities.

In evaluating these risks, it is also helpful to think through what type of information and data is being included and evaluated. Ask: Does the AI either rely on or output data regarding vulnerable groups (i.e., ethnic minorities, women, children, etc.)? For example, demographic data may be necessary but also may increase risks of harm (i.e., ethnic targeting). Similarly, information that includes several data points may be more easily identifiable and thus carry more potential for harm. The same is true if information is specific to victims as

opposed to more generalized. One way to mitigate these risks is data anonymization. Another involves data protection measures. If data can adequately be anonymized, it may be less risky and weigh in favor of its use. The opposite is true if it cannot be anonymized. Another consideration is whether the AI deals with data regarding rights now or over time. If it includes current monitoring, are individuals or communities impacted currently at risk?

5. Evaluate the Potential for Additional Harm

One of the most critical questions for consideration is whether the AI can be repurposed to effectuate additional harms. This also requires a revisiting of Factor 1 and thoughtful analysis of considerations like regime type, government stability, who has access to data and whether protections are in place.

An additional factor for consideration is the potential harm to victims. For example, would it be harmful to victims to have AI conduct interviews or testimony in lieu of a human rights worker? At the same time, are there benefits to human rights workers? This relates back to the issue of mental health impacts on human rights workers, where PTSD and depression are prevalent.

6. Evaluate What Mechanisms Exist to Ensure That the AI Is Being Used Properly

One way to mitigate harms outlined throughout is to adopt mechanisms to ensure that AI is being used properly. First, is there an appropriate internal validation process for the AI? Where these types of checks exist, there is less risk and it is more likely that AI can be deployed responsibly. Where these types of internal validation checks do not exist, the risk of harm is higher. Second, can AI results be validated by a third party? For example, if AI uses open-source data that can be validated by third party watchdog organizations or civil society groups, there may be less risk. Where AI relies on closed data sources and no external validation mechanisms exist, the likelihood of external data validation is depleted and there is a higher risk of harm. Another consideration is whether the data AI is learning from comes from reliable sources like international institutions.

One of the major draws of AI is that it continues to learn and perfect over time. As a result, it is necessary that there be supervision or a check on the algorithm over time to ensure that it is continuing to work effectively and not introduce or multiply existing errors. Ask, is the algorithm supervised? Is it

continuously checked for error and corrected accordingly? If so, risk is reduced. If not, the risk is heightened. For example, if AI outputs are checked by a third party such as a treaty body checking reports from states against shadow reports from third party civil society organizations, the risk is lower.

7. Cost-Benefit Analysis

The next step is to weigh these factors against one another to determine whether AI should be designed and deployed. Where there are relatively low risks across the board, AI can likely be deployed with relative ease and organizations and states should not be afraid to deploy technologies. In those instances, it is likely a worthwhile endeavor so long as the resources are there to see it through.

Where the AI may not be effective because of limitations, it should not be deployed unless limitations can be overcome. There is a high risk of error in these instances. These are places where while there may be great potential for AI's use, significant limitations to the capabilities of AI prevent it from being a worthwhile endeavor. In those instances, it may not be worth the initial resource pull to set up the AI if the AI won't be able to produce enough of an output to make it worthwhile.

Where the risks are high, but the benefits are also high, additional analysis may be needed to determine whether AI should be deployed. In these instances, constraints can be put into place (such as validation mechanisms or data protection policies) to mitigate risks.

Where the risks are high and the benefits are low, it is likely not appropriate to use AI. This includes circumstances where communities are put at additional risk, where there is potential for data to be used to effectuate harms, or where it is being deployed by an unstable actor with no external checking mechanisms.

8. Applying the Framework

Consider the example of the famine in Yemen, one of the most troubling human rights crises of our time,[227] to illustrate briefly how an analysis under this framework could be undertaken. Let us hypothesize that a human rights organization in the region is looking to submit a report to the United Nations related to famine as a result of the civil war in Yemen. The organization is

[227] *Yemen Crisis*, UNICEF, https://perma.cc/M3MQ-M6MY.

contemplating using AI in three ways: (1) using remote sensing and satellite imaging to see where food and aid are located in connection with where villages and populations are located; (2) comparing this information to open-source data to disaggregate and make sense of the information along demographic markers; and (3) using ChatGPT to compile a report.

Here, the first factor of evaluating the actors involved in the creation and deployment of the AI include raising questions related to the legitimacy of the organization, the third-party technology company building the AI, and any other actors involved in its use. For factor two, evaluating how the AI is being used both in the near term and the identification of any future uses, includes a detailed look at how the technology is being used. Here, there are multiple uses, and each should be assessed both separately and together to determine whether issues come up. Factor three, evaluating how the AI will be designed and developed, necessitates a deep dive into the ways in which the algorithm is being developed, how the AI is being taught, whether it is able to correct for issues that come up, etc. A separate analysis would need to take place for each of the three uses—with red flags raised across each (Is there missing or incomplete data? Where is the data coming from? Is there potential for bias? Are data privacy concerns raised?). Similar questions are raised under factor five where additional harms may be introduced. Here, does detailing data related to the locations of certain groups open those groups up to new harms from the government or insurgency groups? The answer is likely yes. Finally, factor six looks at the potential for mechanisms to mitigate these harms. An example in this hypothetical is having third parties to verify the algorithm and data; instituting guidelines around sharing the information beyond the organization; and/or limiting the amount of specified data or instituting ways to anonymize it in the reports.

The organization should weigh these factors against one another to determine whether to use the AI, in what ways, and to what extent.

CONCLUSION

New AI technologies are being introduced every day, and current uses of AI continue to expand into the field of human rights. It is easy to be pessimistic about or fear the ways in which AI continues to change the world around us. But, as this Note explores, there are also exciting and promising benefits of AI in this space. The question becomes: how can AI be used as a tool to ensure rights and to hold governments accountable when they fail to fulfill their

obligations or when they commit human rights violations? While AI technologies disrupt our understanding of law and ethics, they also provide promising opportunities for the advancement of rights. To ensure this happens, an examination of case studies of current uses and an exploration of risks and harms can help us to evaluate when and how AI should be used.

This Note and its proposed evaluation framework can serve as a tool for various actors in this space. Traditional, static regulator frameworks, which rely on government regulation and enforcement, are ill-equipped to manage varying and rapidly growing AI technologies. This type of pacing problem is especially challenging at the international level, where there is not necessarily an overarching regulating or enforcement body equipped to respond to AI's roll-out. A framework like the one proposed in this Note can serve as a foundation for self-regulation by both state and private actors. For actors who are using AI with the goal of advancing human rights, like civil society organizations, this framework provides a sort of instruction manual for potential users in evaluating the use of these types of technologies. Ideally, states, civil society organizations, and third-party developers will be prompted to evaluate AI's limitations and harms, resource investments, and the types of protections necessary to ensure rights prior to its use and on an ongoing basis as AI is used. Even where actors do not themselves evaluate AI's use; this Note's framework can be used as a tool to hold states accountable by providing civil society actors a new avenue or set of soft law standards.

After thoughtful analysis of the use of AI, and with the proper legal and regulatory constraints in place, the potential for positive enhancement of human rights because of the use of AI is encouraging.