

REPORT

JUNE 2024

Leveraging Artificial Intelligence for the Pursuit of Human Rights

DANIELA WILDI | MATTEO BRAIZAT | MYLAN EVRARD

GENEVA
GRADUATE
INSTITUTE

INSTITUT DE HAUTES
ÉTUDES INTERNATIONALES
ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE
OF INTERNATIONAL AND
DEVELOPMENT STUDIES

GENEVA
ACADEMY

Académie de droit international
humanitaire et de droits humains
Academy of International
Humanitarian Law and Human Rights

Leveraging Artificial Intelligence for the Pursuit of Human Rights

GENEVA GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES (IHEID)
Applied Research Project

RESEARCH TEAM

Daniela Wildi
Matteo Braizat
Mylan Evrard

ACADEMIC SUPERVISOR

Erica Moret

PARTNER ORGANIZATION

Geneva Academy of International Humanitarian Law and Human Rights
Bernard Duhaime
Erica Harper

POLICY BRIEF

June 2024
Geneva, Switzerland

Image Rights: 'Robot with digital red brain' by Peshkova, Adobe Stock.

Disclaimer: The views and opinions expressed in this report are those of the authors and do not necessarily represent those of the Geneva Academy of International Humanitarian Law and Human Rights and the Geneva Graduate Institute of International and Development Studies.

Acknowledgements

The authors of this report would like to express their gratitude to the people who provided their support throughout this entire research project and significantly contributed to the success of this project. We would like to thank the Geneva Academy for International Humanitarian Law and Human Rights, and in particular Erica Harper and Bernard Duhaime for their support over the course of the project. We would also like to thank the Geneva Graduate Institute (IHEID), and in particular Erica Moret for her academic and organizational advice, and mentorship during the process of this project. Additionally, we extend our sincere appreciation to Ansgar Koene, Cecilia Garcia Podoley, Danielle Ralic, Emmanuel Goffi, Gaelle Mogli, Giuliano Borter, Inês Gonçalves Ferreira, Jean Ng, Kolja Verhage, Marco Manca, Oana Ichim, Paola Gálvez Calligos, Richard Benjamins, Shiran Melamdovsky Somech, Shea Brown, Tatiana Caldas-Löttiger, Vahid Razavi, and Walid el Abed, whose invaluable insights and perspectives enriched this research project.

Executive Summary

This report aims to address a critical gap in policy discourse related to the intersection of Artificial Intelligence (AI) and human rights. While debates predominantly focus on the negative impacts of AI systems on human rights, this research takes a different approach. The project's primary goal was to identify potential avenues for leveraging AI for the pursuit of human rights. It concludes that, if developed and deployed responsibly, AI presents an opportunity for advancing human rights. A set of illustrative use cases highlights some of the applications in this frontier field. It also addresses limiting factors and risks to be managed and mitigated if the full potential is to be realized. Considering the objectives and outcome of this research, the report puts forth the following recommendations:

- 1. Developing a Data Strategy to Bolster Quality & Usability:** Stakeholders should prioritize refining datasets, focusing on improving data collection methods, incentivizing data sharing, and ensuring fair access to valuable information.
- 2. Implementing Mentorship & Training Programs to Expand the Pool of AI for Human Rights Talents:** Strategic investing in mentorship and training programs is key. These initiatives should aim to cultivate interdisciplinary skills and target underrepresented groups to diversify expertise in the field.
- 3. Strengthening Partnerships & Stakeholder Collaboration to Accelerate Positive Impact:** Strengthening partnerships and collaborative networks is imperative to fully harness the potential of AI in advancing human rights. Stakeholders should prioritize interdisciplinary collaborations and resource-sharing to drive meaningful progress in this crucial domain.

Stakeholders are encouraged to engage with the findings, and support both local and global dialogue aimed at shaping a more responsible use of AI for advancing human rights.

Main Findings

- Through desk research and analysis of ten AI for human rights use cases, the report has identified and characterized five domains, where these AI capabilities are especially pertinent: Human rights education, training, and awareness; monitoring; compliance; investigations; and governance. For example, in human rights education AI can empower girls' enrolment in school in under-resourced areas by utilizing machine learning models that leverage census and out-of-school data to efficiently target interventions, thus addressing critical issues of gender equality and education access. In other domains, AI can accelerate the monitoring of media coverage in death penalty cases worldwide or improve businesses' compliance with human rights standards through algorithmic auditing. In the field of human rights investigations, AI can, for instance, rapidly track and report child sexual abuse material from the Internet, accelerating prosecutions. Finally, it can

improve decision-making in human rights courts by facilitating legal analysis and case drafting, liberating more time and resources for both judges and lawyers.

- However, large-scale use of responsible AI for human rights involves certain risks that must be managed, and compromises may be necessary to ensure that the AI does not inadvertently harm those it is designed to help. Several risks include the generation of inaccurate outputs, the perpetuation of biases present in training data, the absence of 'explainability' in decision-making processes, the potential for widespread misinformation and malicious use, and the uneven distribution of resources and benefits. To mitigate these risks, further efforts are essential.
- Expanding the use of responsible AI usage for human rights will require overcoming some significant 'barriers', especially around data, AI talent shortages as well as 'last mile' hurdles associated with implementing AI solutions for human rights. Data issues include scarcity, accessibility, quality, and privacy concerns, especially in less-developed regions. Addressing these requires developing high-quality datasets and improving data sharing. AI talent shortages present another major 'barrier', necessitating investment in training and mentorship programs to build a diverse skilled workforce that combines technical expertise with domain-specific knowledge. Finally, the successful implementation of AI tools requires interdisciplinary collaboration and sufficient resources to integrate AI solutions effectively. Overcoming these 'barriers' necessitates comprehensive strategies, strong partnerships, and ongoing efforts to ensure AI technologies are responsibly deployed for human rights.
- An interdisciplinary approach, including stakeholders from both the private and public sectors, is essential to ensure that responsible AI can achieve its potential for human rights. AI technologies, while often developed in academic or corporate settings, need practical insights provided by frontline organizations and human rights experts to be effective. To maintain their effectiveness, these AI systems must also be regularly updated to keep pace with rapid technological developments. In addition, it is also important to plan for collaboration, for example with gatekeepers to data, at all stages of the AI development. Better collaboration will require initiatives to focus more on transparency and explainability of their AI systems, thus promoting understanding, trust, and acceptance among a wider group of stakeholders.
- There is a need for regulatory frameworks that prioritize human rights and accountability, establishing clear guidelines and standards to ensure the responsible development and use of AI, and advocating for a human-centered AI governance to create a balance between technological progress and the protection of human dignity and freedoms.

The responsible application of AI for human rights is an emerging topic and many research questions and issues remain unanswered. The use cases outlined in this report are still developing and do not cover all possibilities. It is anticipated to expand upon this foundation in the near future. Furthermore, data on technological innovations and their (potential) applications are currently incomplete. This report underlines the importance of further dialogue on the responsible application of AI for human rights and the need to scale up these efforts to unlock their full positive societal impact.

Table of Contents

ACKNOWLEDGEMENTS	i
EXECUTIVE SUMMARY	ii
Main Findings	ii
LIST OF FIGURES	v
LIST OF ACRONYMS & ABBREVIATIONS	vi
GLOSSARY	viii
INTRODUCTION.....	1
SECTION 1: AI & HUMAN RIGHTS - A REVIEW	2
Key Areas of Impact.....	2
Central Concerns	4
SECTION 2: METHODOLOGY	7
SECTION 3: RESPONSIBLE AI - USING AI FOR HUMAN RIGHTS	9
1. Human Rights Education, Training & Awareness.....	9
2. Human Rights Monitoring.....	11
3. Human Rights Compliance	12
4. Human Rights Investigations.....	14
5. Human Rights Governance.....	15
SECTION 4: MANAGING THE RISKS OF ADOPTING AI	17
Mitigation Strategies.....	18
SECTION 5: EXPANDING THE RESPONSIBLE USE OF AI.....	19
1. Developing a Data Strategy to Bolster Quality & Usability.....	20
2. Implementing Mentorship & Training Programs to Expand the Pool of AI for Human Rights Talent	21
3. Strengthening Partnerships & Stakeholder Collaboration to Accelerate Positive Impact	22
CONCLUSION.....	24
BIBLIOGRAPHY	xxxiv
ANNEX 1: INTERVIEW GUIDE.....	xlii

List of Figures

FIGURE 1: CLASSIFICATION OF AI SYSTEMS & COMMON EXAMPLES, ADAPTED FROM KHOSRAVI ET AL. (2023).....	6
FIGURE 2: EDUCATE GIRLS MACHINE LEARNING SYSTEM (ORIGINAL MODEL).....	10
FIGURE 3: EDUCATE GIRLS MACHINE LEARNING SYSTEM (NEW MODEL)	11
FIGURE 4: OBJECT DETECTION USING SATELLITE IMAGERY	14
TEXT BOX 1: HUMAN RIGHTS FOR AI PRACTITIONERS	5
TEXT BOX 2: AI FOR HUMAN RIGHTS PRACTITIONERS.....	5
TEXT BOX 3: REBUILDING THE EDUCATE GIRLS MACHINE LEARNING SYSTEM	10
TEXT BOX 4: INSIGHTS FROM STANFORD HAI'S INDEX REPORT 2024	21
TABLE 1: BARRIERS LIMITING THE USE OF AI FOR HUMAN RIGHTS.....	19

List of Acronyms & Abbreviations

AIDA	Artificial Intelligence and Data Act
AI	Artificial Intelligence
ASEAN	Association of Southeast Asian Nations
BCI	Brain-Computer Interface
CEDAW	Convention on the Elimination of All Forms of Discrimination against Women
CRPD	Convention on the Rights of Persons with Disabilities
CSAM	Child Sexual Abuse Material
DOJ	Department of Justice
ECHR	European Convention on Human Rights
ECtHR	European Court of Human Rights
ECPAT	End Child Prostitution, Child Pornography and Trafficking of Children for Sexual Purposes
EDPS	European Data Protection Supervisor
EEOC	Equal Employment Opportunity Commission
EU	European Union
FRT	Facial Recognition Technology
HRW	Human Rights Watch
ICCPR	International Covenant on Civil and Political Rights
ICERD	International Convention on the Elimination of All Forms of Racial Discrimination
IHRL	International Human Rights Law
ILO	International Labour Organization
IT	Information Technology
LLM	Large Language Model
MNO	Mobile Network Operator
ML	Machine Learning

NGO	Non-Governmental Organisation
NLP	Natural Language Processing
NSA	Non-State Actor
OCHA	Office for the Coordination of Humanitarian Affairs
OHCHR	Office of the United Nations High Commissioner for Human Rights
SDG	Sustainable Development Goal
SEC	Securities and Exchange Commission
SSI	Semi-Structured Interviews
TAN	Transnational Advocacy Network
UDHR	Universal Declaration of Human Rights
UN	United Nations
UNCHR	United Nations Commission on Human Rights
UNHRC	United Nations Human Rights Council
UNICRI	United Nations Interregional Crime and Justice Research Institute

Glossary

ARTIFICIAL INTELLIGENCE (AI): ‘the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.’¹

AUTOMATED DECISION-MAKING: describes situations where the user of the technology either does not have access to the source code or may not fully understand it due to having non- technical educational background.²

BIOMETRIC FACIAL RECOGNITION TECHNOLOGY: ‘entails automated differentiation of facial features in law enforcement [...] to solve crimes while integrating individual privacy, regulation, autonomy, security, and democratic accountability.’³

BRAIN-COMPUTER INTERFACES (BCIs) OR BRAIN-MACHINE INTERFACES (BMIs): surgically implemented chip inside the brain (invasive) or helmet (non-invasive) allowing for ‘bidirectional communication between the brain and either the outside world or another machine.’⁴

BLACK BOX: a system that produces results without the user being able to see or understand how it works.⁵

DEEP LEARNING: a ‘form of machine learning, the use of data to train a model to make predictions from new data.’⁶

DIGITAL CONSTITUTIONALISM: ‘a transposition of the authority of governments, their political expression, and constitutional values and rights into the digital environment.’⁷

EXPLAINABLE ARTIFICIAL INTELLIGENCE: ‘the ability of [a system] to summarize the events of the game/simulation, flag key events, and explain the behaviour of computer-controlled entities.’⁶

FACIAL RECOGNITION TECHNOLOGIES (FRT): ‘works by capturing an individual’s image and then identifying that person through analysing and mapping of those captured features comparing them to identified likenesses.’⁸

HASHING: a technology used in child protection to ‘convert a piece of known CSAM into a unique string of numbers through an algorithm, called a hash [acting] like a digital fingerprint for each piece of content.’⁹

HASH MATCHING: a system comparing hashes against known CSAM hash lists, searching for matches algorithmically without accessing users’ content (removed from the platform if match found).⁹

MACHINE LEARNING: ‘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 analyse and draw inferences from patterns in data’.¹

NEURO-RIGHTS: umbrella term describing ‘new human rights that essentially seek to protect the individual’s control over his or her mind’.¹⁰

NEUROTECHNOLOGIES: ‘the assembly of methods and instruments that enable a direct connection of technical components with the nervous system’.⁴




NEW HUMAN RIGHTS: ‘rights that, when first conceived, are not expressly recognised in any human rights treaty and are not in any other way recognised as rights in a legal sense’.¹⁰

SUPERVISION TECHNOLOGY (SupTech): machine-learning technology that lessens the burden of complying with or supervising compliance with regulatory requirements.¹¹

Introduction

Within the scientific landscape marked by the rapid evolution of artificial intelligence (AI), debate has emerged around its implications for the pursuit of human rights. As AI development continues to advance, it brings with it a range of challenges but also opportunities that can play an important role in the promotion of human rights. This report aims to enrich the emerging discussion on the relationship between these two fields, focusing on the potential positive impacts and applications of AI on human rights. Against this backdrop, the recently adopted UN General Assembly Resolution A/78/L.49 of 2024 backed by 120 Member States on the promotion of 'safe, secure and trustworthy' AI emphasizes the 'respect, protection and promotion of human rights in the design, development, deployment and the use of AI'.¹²

This report addresses three questions:

-  1. **Where can AI be leveraged for the pursuit of human rights, and what insights can be discerned?**
-  2. **What are the risks, challenges, and 'barriers' regarding the positive use of AI for human rights, and how can they be mitigated and managed?**
-  3. **How can different stakeholders expand the responsible use of AI for human rights?**

To illustrate the application of AI in the pursuit of human rights, the report is structured as follows. Section 1 reviews the intersection between AI and human rights, and how the discourse is perceived in the literature. Section 2 explains the methodological framework utilized in this research, which combines the examination of primary and secondary sources with the collection of semi-structured interviews with experts from various fields. Section 3 presents ten use cases derived from desk research and interviews that show the positive applications of AI in five identified human rights areas, notably human rights education, training, and awareness, monitoring, compliance, investigations, and governance. After acknowledging inherent risks related to AI and proposing some mitigation strategies in Section 4, the report concludes with recommendations for action and suggestions for further research to expand the responsible use of AI for human rights.

Section 1: AI & Human Rights - A Review

The International Human Rights framework is regarded as ‘the backbone of the freedom to live in dignity’ by the United Nations (UN).¹³ It comprises legal processes, institutions and norms, and requires accountability before the law, as well as compliance and enforcement. Despite recent legal innovations at the domestic level,^{14–18} AI regulations and norms remain scarce; there are no global institutions providing oversight, and it suffers from a lack of legitimacy, accountability and enforcement.¹⁹ An ontological clash thus arises from the interaction of the two domains, often resulting in a negative framing of AI concerning some specific areas of human rights, such as the protection of women,²⁰ state surveillance,^{3,6,8,21,22} data privacy,^{23,24} freedom of expression,^{25,26} or neuro-rights.^{10,27–29} Studies focusing on the ethical considerations of AI and human rights also tend to emphasize risks and challenges rather than opportunities.^{7,30–32} Such framing has an impact on the people and governing institutions, eventually influencing democratic voting, decision-making and the rule of law.³³ Although several studies and reports have already provided interesting case studies of positive applications of AI for human rights,^{1,2,34–38} they tend to shift their focus on broader ‘AI for good’ applications, namely applications that aim to improve the life of the people for their (social) good,³⁹ or achieving broader goals such as the Sustainable Development Goals (SDGs).^{39,40,42}

Key Areas of Impact

Against the background of the predominant negative discourse focusing on AI and human rights, this report aims at gathering concrete examples of positive applications of AI in enhancing not only human rights *per se*, but also the legal and governance processes required for their implementation and oversight. As there are still some limits to the implementation of AI (see Section 5), and based on the results of the data collection, the following thematical areas of human rights have been selected because of their likeliness to incorporate AI tools. They are also broad enough to be applicable to all human rights, thus facilitating analogies when it comes to positive applications of AI. This implies looking closer at the concepts of human rights education, training, and awareness, monitoring, compliance, investigations, and governance. Without being exhaustive, this list of key areas of impact of AI provides a comprehensive framing when it comes to human rights implementation.

EDUCATION, TRAINING & AWARENESS

Education here is conceived as both the education of human rights and the right to education.⁴¹ It is also understood as not only limited to scholarly education but also public awareness and professional training. AI’s potential primarily lies in contributing to teaching and research methodologies, as well as student assessments.⁴² Chatbots such as ChatGPT are already integrated into both teachers’ and students’ working habits and are most likely to further change education practices in the future. Similarly, the use of AI in targeted advertising raises both concerns on data privacy⁴³ and opportunities, as such technologies can also serve as tools for promoting awareness as regard to specific human rights violations. Finally, the intersection of AI and human rights education must also be acknowledged as subject areas, part of two different curriculums that do not usually overlap.

MONITORING

AI tools have the potential to bring real value-added by extensively assisting the tracking of human rights violations, as well as promoting policies to protect individual liberties.³¹ Indeed, human rights evaluation is particularly relevant in the context of the rapid advancement of information technology (IT) as its evolving legislation increasingly overlaps with human rights.³¹ However, monitoring large amounts of data using AI may raise privacy concerns, and what is acceptable or not regarding monitoring practices remains uncertain.⁸

COMPLIANCE

Whether due to the absence of political and corporate will or simply because of a lack of resources, there is a risk that states, and private entities do not comply with both domestic and international human rights law. Andrey. A. Tymoshenko (2022), Associate Professor at the University of the Prosecutor's Office of the Russian Federation, provides the example of the Russian Federation's potential non-compliance with the European Court of Human Rights (ECtHR) decision regarding the use of digital technologies.²¹ This should be investigated further, as regional courts such as ECtHR are pioneers in applying a pragmatic approach to the use of AI technology in national jurisdictions.²

INVESTIGATIONS

In certain instances, specific AI applications implementations, such as in biometric facial recognition technology^{3,8} and neurotechnology^{27,29}, may facilitate the investigation of human rights violations but also pose risks. For example, empirical studies have highlighted that the wireless communication standards of Brain-Computer Interfaces (BCIs) may expose subjects to interference risks from unauthorized external sources, especially if deployed in criminal investigations without adequate security measures.²⁷ Indeed, capturing neural signals for information about a subject's intentions, for example, can conflict with fundamental human rights principles, such as 'the right to privacy, the right against self-incrimination, the right to remain silent, and freedom of thought'.²⁷

GOVERNANCE

AI has potential to improve human rights governance by facilitating communication between states and non-state actors of the regime (understood here as 'a set of implicit or explicit principles, norms, rules and decision-making procedures around which actors' expectations converge in a given area of international relations'⁴⁴) or improve our understanding of the legal complexity that forms the international human rights framework.⁴⁵ But main stakeholders, such as human rights courts and NGOs, suffer from a limited access to AI-driven technologies, and important questions arise concerning who is legitimate to do what and who to hold accountable when AI is used for legal analysis or decision-making.²

Central Concerns

LEGITIMACY

The question of legitimacy, accountability and enforcement remains central throughout all the above-described processes, as they are strongly challenged by the use of AI in human rights contexts. Indeed, legitimacy concerns in human rights may arise when the actions of public authorities, such as the police or an administration, are questioned by society. This is highly relevant in the case of facial recognition systems and algorithmic surveillance. Therefore, legitimacy can only be achieved through 'consistent lawful practices and societal expectations'.³ Legitimacy may also refer to the right of a particular legislation to rule cases related to human rights. However, in some cases this may raise an issue if the only legitimate court is of constitutional or national authority on a matter related to the use of surveillance technologies by the state.²¹

ACCOUNTABILITY

Accountability provides the obligation to explain, justify, and take responsibility for actions.⁸ This issue is particularly relevant in the case of a given government's use of digital technologies. Tesson et al. (2022) take the example of the Canadian Artificial Intelligence and Data Act (AIDA)⁴⁶ and its potential applicability to only the private sector, therefore exempting the government and distancing the country (as well as other relevant stakeholders, e.g. non-state actors) from global accountability measures.²² The use of AI tools such as facial recognition technologies can also integrate democratic accountability, essential in human rights rulemaking.³ Additionally, the issue of accountability may arise when a human rights court like the ECtHR uses automated/algorithmic decision-making to process cases. The court must then be clear on how it is held accountable for such a decision.²

ENFORCEMENT

Enforcement is usually ensured by national authorities first,¹⁴ but can also be achieved through regional or international institutions such as the ECtHR.⁶ Almeida et al. (2022) highlight the issue of enforcement when the very law enforcement officers are using AI-related tools on the population such as facial recognition technology (FRT);⁸ while Bacalu (2022) gathered literature on how law enforcement authorities can leverage digital technologies that have social impact.³ In the context of AI used by both private and public entities, researchers particularly highlight the need to enforce privacy laws.^{8,47}

Interlinkages between AI and human rights often present as many opportunities as challenges. The low level of regulation regarding AI is effectively mirrored in the current literature, which tends to focus more on the risks associated with such applications. One notable element is that both academic and policy papers usually focus on a particular sector (be it governments, NGOs or tech companies), thus framing a siloed discourse that is not fully representative of the potential AI can have for enhancing human rights. By presenting concrete use cases within five broad categories, this report attempts to demonstrate the richness of the current applications and interdisciplinary as well as intersectoral collaborations.

TEXT BOX 1: HUMAN RIGHTS FOR AI PRACTITIONERS

Human rights are ‘rights that belong to an individual or a group of individuals simply for being human, or as a consequence of inherent human vulnerability’.⁴⁸ They include initially ‘the right to life and liberty, freedom from slavery and torture, freedom of opinion and expression, the right to work and education’, among others.¹³ Although some of them were already considered into domestic constitutions up until the Second World War, they obtained international recognition through the 1948 Universal Declaration of Human Rights (UDHR), and later with the two 1966 International Covenants: the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR). There are currently nine binding international human rights instruments in force (supplemented by protocols), such as the 1979 Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) or the 2006 Convention on the Rights of Persons with Disabilities (CRPD), that vary in terms of states recognition.⁴⁹ Additionally, few regional human rights agreements exist, such as the 1950 European Convention on Human Rights (ECHR) or the 1986 African Charter on Human and People’s Rights, which allow for more tailored human rights frameworks. They are also increasingly recognised in the UN plan of actions and other policy frameworks (often in their preamble), such as the 2015 Sendai Framework for Disaster Risk Reduction and Paris Agreement on Climate Change.

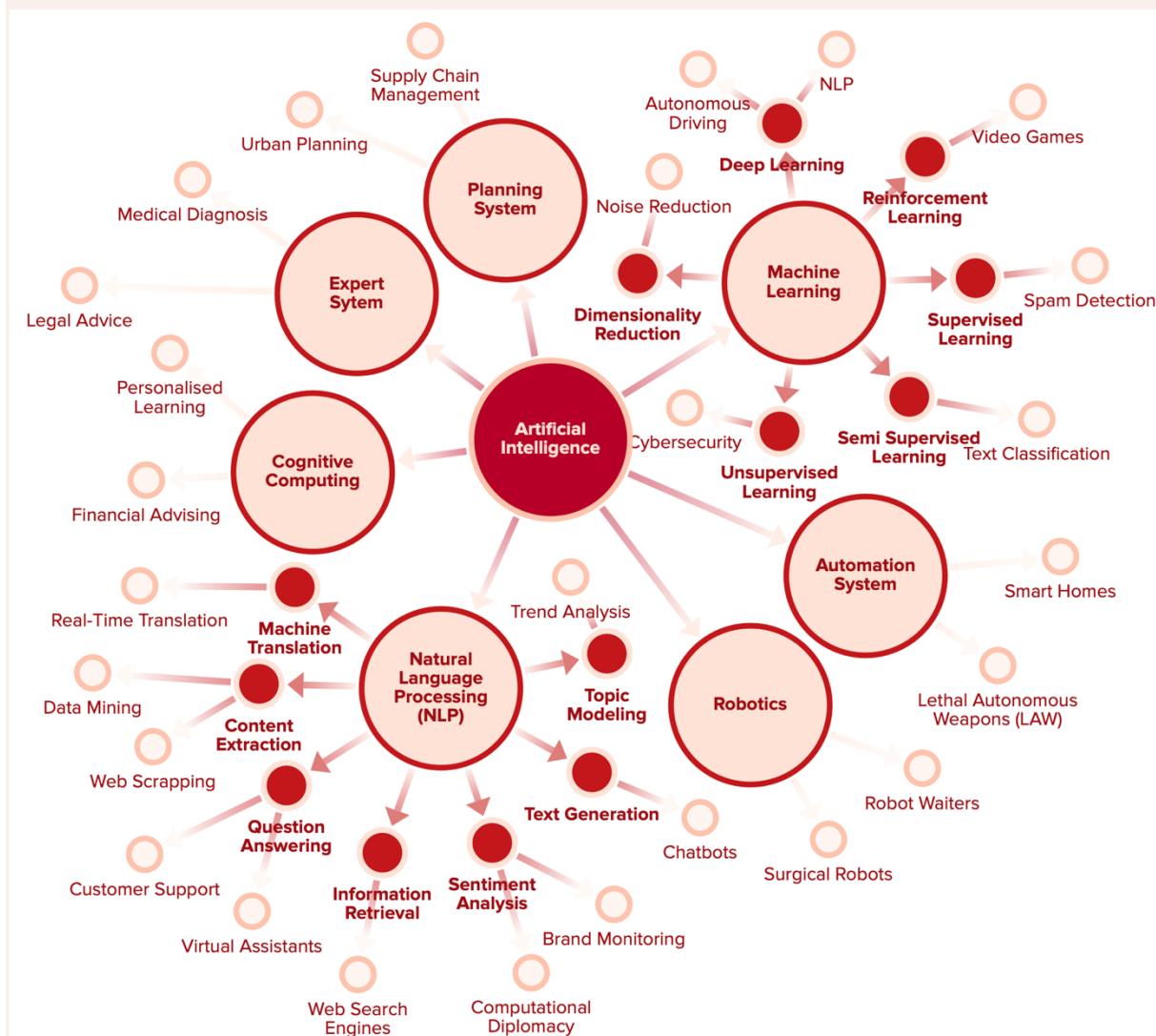
Made of 47 Member States, the United Nations Human Rights Council (UNHRC), which replaced the United Nations Commission on Human Rights (UNCHR) in 2006, is the main intergovernmental body responsible for investigating allegations of human rights violations worldwide through annual sessions, working groups and the use of Special Rapporteurs. It works closely with the Office of the UN High Commissioner for Human Rights (OHCHR) which serves as its Secretariat and provides assistance to governments in implementing international human rights standards. Other stakeholders such as businesses and non-state actors also have incentives to voluntarily incorporate these standards, such as the 1988 International Labour Organization (ILO) Declaration on Fundamental Principles and Rights at Work, or the 2011 United Nations Guiding Principles on Business and Human Rights.

With the rise of new environmental and technological issues in public debates, human rights now go beyond their traditional scope by also encompassing environmental rights, such as the right to a healthy environment, recently recognized almost universally in the UN General Assembly Resolution 76/L.75,⁵⁰ or the newly exercised digital rights, that can apply to data protection and privacy, digital identity, surveillance technologies, or online violence, among others. Finally, the introduction of AI and neuro-technologies such as BCIs also raise legal and ethical issues requiring what researchers refer to as ‘new human rights’, namely rights that are not expressly recognized as such in a legal sense, like the right to mental privacy and mental integrity, the right to psychological continuity and to cognitive liberty.¹⁰ It is most likely that human rights will continue to evolve as we create new means to express them.

TEXT BOX 2: AI FOR HUMAN RIGHTS PRACTITIONERS

The discipline of AI is considered to have started with the introduction in 1950 of the Turing test by computer scientists Alan Turing, which aimed to determine whether a machine could exhibit human behaviour. Although it features the word ‘intelligence’, current AI systems do not possess cognitive abilities as such and use logic-based interpretations of the world to make inferences or solve tasks.⁵¹ Indeed, AI originally builds on traditional machine learning models, namely algorithms requiring human intervention to be developed and maintained, such as web search or original chatbots. However, these models remained limited in their ability to process raw data (e.g., image, text, sound, etc.).⁵² Deep-learning methods were then introduced to transform these raw data into suitable representations that the algorithm could detect through different levels of abstraction called neural networks.⁵³ Using labelled data (supervised learning) and unlabelled data (unsupervised learning), algorithms are trained to detect objects with near-human accuracy.⁵³ With the rise of machine learning, similar techniques such as word-embedding or sequence-to-sequence models called large language models (LLMs) have been used to understand and generate human language, performing various natural language processing (NLPs) tasks such as question answering, sentiment analysis, topic classification or translation.⁵⁴ OpenAI’s ChatGPT (Generative Pre-trained Transformer) is one example of an LLM using NLP.⁵⁵ Other AI systems range from cognitive computing, which stimulates human thought processes or crunch enormous amount of data,⁵⁶ to automation systems, which operate machines with minimal human intervention⁵⁷, and expert systems, which use knowledge and inference procedures to simulate the judgement or behaviour of a human.⁵⁸ Figure 1 represents these different AI systems according to current terminologies,⁵⁵ as well as a few examples (smaller circles on the outskirts of the figure).

FIGURE 1: CLASSIFICATION OF AI SYSTEMS & COMMON EXAMPLES, ADAPTED FROM KHOSRAVI ET AL. (2023)



Indeed, because AI is ‘not a homogenous concept’,⁵⁹ it remains constantly in a state of flux, with overlapping categories and terminologies that might change in the future.⁵² In view of AI’s rapid advancements, researchers already differentiate between three broad categories of AI based on their capabilities: (i) Artificial Narrow AI (or ‘weak AI’), which includes more or less all AI systems used so far, (ii) General AI (or ‘strong AI’), that is only conceptual so far and would entail abilities to learn and perform without any human intervention, and (iii) Super AI, a purely theoretical system that would possess learning and cognitive abilities superior to those of humans, and would have emotions, needs and desires of their own.⁵² If based on current functionalities, then AI can be divided into four types: (i) Reactive Machine AI, working on available data without any memory, (ii) Limited Memory AI, operating on a specific task with temporary memory, (iii) Theory of Mind AI, that could understand thoughts and emotions, simulating human-like relationships, and (iv) Self-Aware AI, corresponding to a Super AI.⁵² Because of its broad scope of applicability, one of the biggest issue of AI remains the question of responsibility, namely which moral agent is to be held responsible for the AI’s output. Only ethical, explainable, transparent and accountable AI systems allow humans to take full responsibility for the actions of ‘artificial moral agents’.⁶⁰ This largely remains a legal and political question that requires the establishment of common standards across stakeholders (from individuals to governments) to make sure that such conditions are respected.⁵¹

Section 2: Methodology

This section explains and justifies the methodological approach adopted in the research project.

RESEARCH DESIGN

In order to achieve the project's objective, this research employed a qualitative method approach, involving the examination of primary and secondary sources alongside conducting subject matter expert interviews. Primary sources included binding international conventions, all stemming from the 1949 Universal Declaration of Human Rights, as well as regulatory frameworks and international soft law instruments related to data protection and AI ethics. Secondary sources included academic literature as well as civil society reports and policy papers.

DATA COLLECTION & MANAGEMENT

Primary data was collected through semi-structured interviews (SSI) with a total of 18 subject matter experts, held between 19.03.2024 and 25.04.2024, encompassing public officials, human rights and technology sector representatives, UN-affiliated experts, civil society activists, and academic scholars. Interviewees were primarily identified through professional networking platforms and by reviewing participants at industry events. This diverse range of expertise allowed for a multi-stakeholder representation. Utilizing an SSI guide, the questions addressed the positive applications of AI in the human rights context, associated risks, challenges, and ethical considerations, mitigation strategies, diverse roles of different stakeholders, strategies for interdisciplinary collaboration, and future directions and recommendations in leveraging AI for the pursuit of human rights. An open-ended approach allowed for gathering interviewee's additional perspectives (see Annex 1: Interview Guide).

ANALYSIS OF DATA

The analysis of data for this research project adopted a multifaceted qualitative approach, integrating insights derived from desk research and SSI with subject matter experts. The thematic analysis of secondary sources identified recurring patterns, emerging themes, commonalities, and perspectives in existing discourse.⁶¹ In analyzing the SSI data, a dual approach was employed using thematic analysis for overarching insights, and content analysis to provide case studies that illustrate key stakeholder perspectives. To enhance the reliability and validity of the findings, the interview data was cross-referenced with secondary sources. The synthesized insights from this comprehensive data analysis formed the basis of this report.⁶²

LIMITATIONS

The following research limitations were encountered. The first challenge revolves around reliance on key experts for SSI, potentially introducing bias as a limitation in representing perspectives due to time constraints during the data collection process. The execution of the research depended on who can be engaged for the interview, potentially overlooking diverse stakeholder opinions. The second challenge pertains to the rapidly evolving nature of AI technology. Data concerning technological innovations and their positive applications are

currently incomplete. The report acknowledges that the outlined use cases are evolving and do not encompass all potential scenarios. However, this report is intended to serve as a baseline for leveraging AI in the pursuit of human rights, providing a foundation for further research. Stakeholders are encouraged to engage with the findings. Simultaneously, potential changes in the policy and regulatory landscape, during and after the research, could impact the relevance and applicability of proposed recommendations. To mitigate this, the report established a forward-looking perspective providing a basis for recommendations adaptable to evolving technological and regulatory landscapes.

ETHICAL CONSIDERATIONS

Throughout the research process, adherence to the IHEID (Geneva Graduate Institute for International and Development Studies) ethical guidelines was maintained to guarantee participants' confidentiality and uphold data privacy. Necessary approvals and informed consent were obtained from participants before conducting the interviews. An information sheet providing project details accompanied the consent process.

Section 3: Responsible AI - Using AI for Human Rights

Responsible AI deployment varies based on the domain, capabilities, barriers, and risk profiles of specific use cases. To demonstrate the wide range of areas where responsible AI technologies can be applied for human rights, this section presents ten illustrative use cases. Although these are only a set of examples, they highlight the diversity of applications, the different capabilities that could be utilized, and the positive impact across five human rights areas chosen for this report: 1. Education, Awareness & Training; 2. Monitoring; 3. Compliance; 4. Investigations, and 5. Governance. The presented use cases are in no means to be considered exhaustive and continue to develop alongside AI's capabilities.

1. Human Rights Education, Training & Awareness

The UN Committee on Economic, Social and Cultural Rights, in General Comment No. 13 on the Right to Education, states, 'Education is both a human right in itself and an indispensable means of realizing other human rights'.⁴¹ Human rights education involves diverse educational efforts, such as training programs, informational campaigns, and learning activities to promote universal understanding, respect, and observance of all human rights and fundamental freedoms. In doing so, it plays an important role in preventing human rights violations and abuses.⁶³ In educational and training settings, AI has been used in certain contexts for tasks such as enhancing teaching and research methodologies, student assessments, personalizing learning experiences, and improving decision-making through predictive analytics and adaptive systems. Thus, AI technologies influence how we teach, assess, and acquire knowledge in different fields, including human rights, thereby also raising awareness.⁴²

USE CASE 1.1: GENERATIVE AI TO STRENGTHEN ONLINE SUPPORT FOR VICTIMS OF DOMESTIC VIOLENCE THROUGH ANIMATED, ACCESSIBLE, MULTILINGUAL CHATBOT

Led by the Swiss non-profit organization Spring ACT with support from D-ID (AI Video Generator) and Microsoft, the animated Sophia.chat project was launched in December 2021 and uses Generative AI to enhance global online support for victims of domestic violence.⁶⁴ According to UN Women, an estimated 736 million women – nearly one in three globally – have experienced violence in their lifetime, while approximately 48,000 women were killed by their intimate partners or family members worldwide in 2022.⁶⁵ Through D-ID's text-to-video technology, the chatbot is animated with a photorealistic avatar, enhancing relatability and accessibility for users seeking assistance. Additionally, Microsoft Azure AI's text-to-speech capabilities and Azure AI cloud infrastructure enable the chatbot to operate in multiple languages (English, French, Spanish, Russian, Arabic, Swahili, and Mandarin). More than 15,000 individuals have used this service since its launch in 2021.⁶⁶ This initiative aims to raise awareness about domestic violence, to empower victims to assert their rights, and to educate them about available support services in a safe and confidential manner.⁶⁷

USE CASE 1.2: EMPOWERING GIRL'S ENROLLMENT IN SCHOOL

Educate Girls, a Mumbai-based non-profit organization, is dedicated to addressing educational inequality in rural and under-resourced areas of India.⁶⁸ The organization has adopted a machine learning model that leverages census data, which is manually cleaned and

updated where necessary, and district-level out-of-school data. This innovative approach enables Educate Girls staff to reach a greater number of prospective students faster and target interventions more accurately. Before this model was developed, the member staff had to travel from village to village to collect the required data which was then compiled and analyzed manually to pinpoint areas most in need. In India, approximately four million girls aged six to 14 are not enrolled in school, making gender equality and access to education critical issues. According to Safeena Husain, Founder and Board Member of Educate Girls, employing AI for precision targeting allows Educate Girls to achieve in five years what would have otherwise taken 45 years.⁶⁹ Since 2018, the organization has targeted to enroll over 1.6 million out-of-school girls – around 40% of all such girls in India – into grades one through ten.⁷⁰ Currently, UNICEF is supporting the Educate Girls’ program in Rajasthan and Madhya. Such initiatives address gender equality and education access, combined with its advocacy and community engagement efforts.

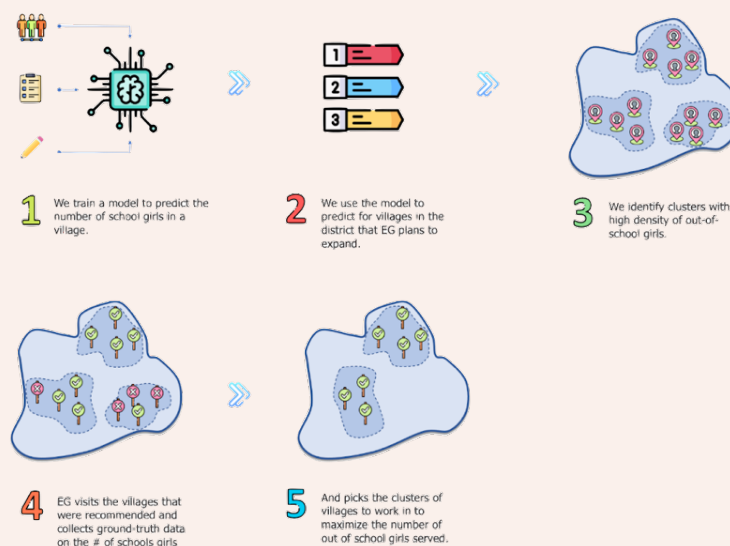
TEXT BOX 3: REBUILDING THE EDUCATE GIRLS MACHINE LEARNING SYSTEM

Educate Girls recently improved their machine learning model to better identify villages with high numbers of out-of-school girls. This update aimed to incorporate recent feedback and data to enhance the model’s performance.

ORIGINAL MODEL

The original model, developed in 2018-2019, used census data from 2011, the 2017-2018 DISE surveys, and Annual Statutes of Education Report (ASER)⁷¹ district-level data. This approach improved efficiency by 50-100%, helping Educate Girls reach more girls within the same budget. While this allowed the organization to accelerate the scaling of their programs, the model struggled with predictions in regions far from its training set.⁷²

FIGURE 2: EDUCATE GIRLS MACHINE LEARNING SYSTEM (ORIGINAL MODEL)



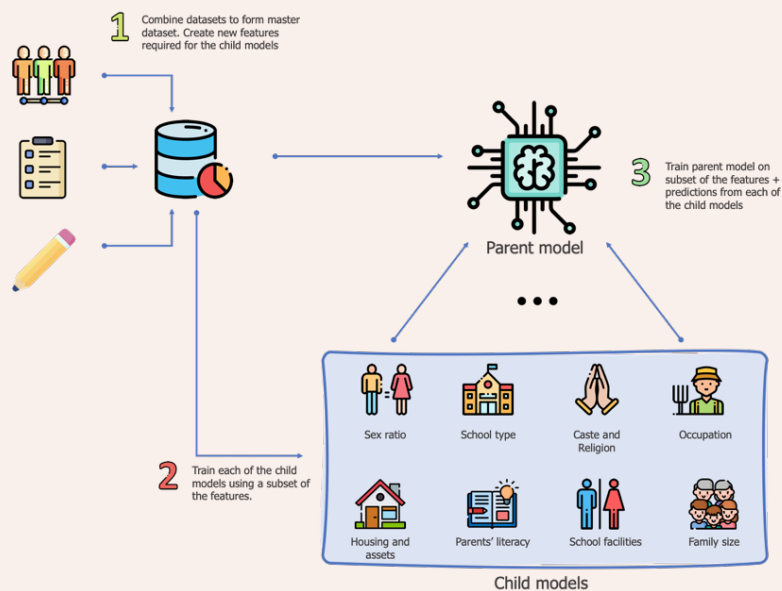
This diagram has been designed using resources from Flaticon.com

Source: Ravinutala S. Rebuilding the Educate Girls machine learning model. IDinsight. 2019; published online April 29. <https://www.idinsight.org/article/rebuilding-the-educate-girls-machine-learning-model/> (accessed May 20, 2024)

NEED FOR A NEW MODEL

The need for a new model arose from these limitations and the desire to incorporate new insights and data. Educate Girls and IDinsight identified several key factors influencing girls' education, such as parents' literacy, caste, occupation, and school facilities. The new model's architecture leverages these causal theories, employing a two-tier system: 'child models' that focus on specific theories and a 'parent model' that aggregates their predictions. This assembling approach improved stability and performance, reducing cross-validation error by nearly 20%.⁷²

FIGURE 3: EDUCATE GIRLS MACHINE LEARNING SYSTEM (NEW MODEL)



Source: Ravinutala S. Rebuilding the Educate Girls machine learning model. IDinsight. 2019; published online April 29. <https://www.idinsight.org/article/rebuilding-the-educate-girls-machine-learning-model/> (accessed May 20, 2024)

IMPORTANCE OF CONTINUOUS IMPROVEMENT IN AI MODELS

This development highlights the need for constant reevaluation and improvement of AI models. As the example demonstrates, incorporating new data and insights is crucial for maintaining model accuracy and effectiveness. Constant updates ensure that the model adapts to changing conditions, thereby maintaining transparency and trust. Educate Girl's approach not only optimized their current efforts but also sets a precedent for the responsible use of AI for human rights.

2. Human Rights Monitoring

Monitoring is a mechanism for improving the assessment and protection of human rights. It involves gathering information, analyzing it and drawing up recommendations to guide states in their assessment of the human rights situation, and thus enhance their duty to respect and realize these rights.⁷³ According to Gauthier de Beco, Professor of Law at the University of Leeds Beckett, this fosters an environment that promotes transparency and accountability, as it raises public awareness and encourages them to voice their claims in the face of abuses. The development of technologies such as AI in this process improves the efficiency and accuracy of monitoring.⁷⁴

USE CASE 2.1: MEDIA SURVEILLANCE AND TRACING OF CAPITAL PUNISHMENT

The creation of a new tool to facilitate the monitoring of information concerning death penalty cases was launched in 2018 by Amnesty International, in collaboration with Element AI, a company that provides AI services and solutions to help organizations harness the power of AI. Previously, volunteers were tasked with scouring the web for articles of interest and then entering the essential information themselves into a database. According to Amnesty, the introduction of AI has marked a turning point in their approach, using advanced algorithms such as NLP to identify and categorize large quantities of data, enabling the automation of the monitoring of media coverage on death penalty cases. With an accuracy rate of 79%, the tool increases efficiency in identifying relevant articles, offering the possibility of faster intervention and thus reinforcing the credibility of Amnesty International's reports and their actions based on this information.¹ Although the NGO plans to improve this tool, as it still requires human supervision, it reduces the tasks of volunteers so that they can concentrate on higher value-added activities, such as implementing action strategies and advocating for policy changes. Finally, this technology broadens Amnesty's monitoring capacity, enabling them to track human rights violations more exhaustively.

USA CASE 2.2: AI-DRIVEN MOBILITY DATA FOR HUMAN RIGHTS-CENTRIC PANDEMIC RESPONSE

During the COVID-19 pandemic, the European Commission collaborated with 17 mobile network operators (MNOs) to develop two tools to help health authorities better understand and manage the spread of the virus. The first technology developed is based on the association of MNO mobility data with official virus statistics, to highlight correspondences between travel patterns and case numbers at various levels (country, region, province). This tool has not only been used to represent vaccination status in member states, but also by the Ministry of Health in Spain to monitor travel flows and assess the effect of mobility restrictions on the country's economic situation. The second technology that has been developed is the Scenario Analysis Toolbox, which enabled public health services in several European countries to model cases of infection, adjusting various parameters to improve their response to the pandemic, particularly in terms of the speed of interventions.⁷⁵ These tools have been used to predict the spread of the virus and anticipate outbreaks, enabling health services to prepare for the implementation of containment measures.⁷⁶ Data protection follows confidentiality standards, thanks to the implementation of anonymization and aggregation processes being certified by the European Data Protection Supervisor (EDPS). According to Richard Benjamins (CEO of OdiselA), these technologies have improved public health measures in relation to the socio-economic aspect, notably by estimating the impact of mobility restrictions and promoting international cooperation to combat the virus. Thus, AI-driven mobility data can serve to enhance human rights monitoring by providing precise, real-time insights into population movements and public health trends to ensure equitable resource allocation, but also improve accountability in decision making.

3. Human Rights Compliance

With the proliferation of international human rights standards, voluntary goals and due diligence, compliance with human rights is no longer the sole responsibility of states. Private and other non-state actors also need to ensure that the way they operate respect human rights norms and domestic legislations. In an increasingly fragmented and data-driven legal

landscape, maintaining a high degree of compliance with human rights standards is becoming more complex for both the public and private sectors.⁴⁵ When trained on the relevant data, AI models can check in real-time the level of compliance of specific corporate practices or help governmental agencies to detect foreign corruption and fraud in real-time while saving time for administrators and regulators.

USE CASE 3.1: ALGORITHMIC AI-AUDITING

Companies now often resort to algorithms and other AI-driven technologies to boost their productivity, facilitate recruitment processes and better target consumers.⁷⁷ However, whether it comes from their developers, the data they are fed with, or simply their users, algorithms and algorithmic AI often bring their share of biases, challenging the principles of equity and non-discrimination with regard to the targets (consumers, job candidates, employees and so on).⁴³ This raises the issue of responsibility of algorithms in decision-making, from their design to their final application. Additionally, it may also be difficult for firms using such algorithms to fully comply with the emerging fragmented legal landscape regulating AI.⁷⁸ Auditing companies such as BABL AI are using both human and AI abilities to assess the conformity of firms' algorithms and other AI-driven technologies with current regulations (such as the newly adopted EU AI Act, or the U.S. Equal Employment Opportunity commission (EEOC)) and to limit biases linked to discriminatory practices. As algorithms are becoming ever more complex, the use of AI in tracking biases and malpractices is likely to increase in the coming years. This evolution should not be seen as the result of companies' lack of will in complying with human rights standards in their day-to-day operations, but rather as a lack of financial and managerial resources that AI can help to fill efficiently if properly administered. Auditing companies are thus likely to play an important role in the future in enhancing businesses' compliance with human rights and ethical requirements.

USE CASE 3.2: REAL-TIME COMPLIANCE REPORTING AND PREDICTIVE ANALYTICS

With the rise of data-driven decision-making, both governments and companies now have the possibility to better track and report corruption, human rights violations or financial frauds using real-time and predictive AI-powered compliance tools.⁷⁹ Indeed, the U.S. Securities and Exchange Commission (SEC), for example, is currently leveraging supervisory technologies (SupTech) to detect potential insider trading and inaccuracies in financial reporting. The SEC has already produced six enforcement actions involving fraud charges and other significant penalties,¹¹ at the time of writing. Regulators, such as the ones working at the U.S. Department of Justice (DOJ), are also using such tools to detect wrongdoings involving foreign corruption or corporate misconducts, such as the DOJ Criminal Division's Evaluation of Corporate Compliance Programs.¹¹ Additionally, AI can also be used to analyze vast amounts of historical data, witness testimonies, social media posts, emails, complaints or transactional data, in order to detect trends, patterns or potential anomalies.⁷⁹ NLP techniques can be used to quickly filter informative content in the context of communication and highlight compliance breaches. With the right controls, such technologies can help investigators in public and private spheres to track regulatory changes, better collect and anonymize whistleblowers' complaints and update the risk profiles of different companies.¹¹ Depending on the data available, similar techniques can also be used by NGOs or the civil society to identify, in real-time, governmental violations of human rights and even predict them, enabling preventative measures.⁷⁹

4. Human Rights Investigations

Investigations play a central role in uncovering and resolving human rights abuses. They consist of objective examinations of specific human rights violations, with the aim of establishing facts, acknowledging responsibility and ensuring justice for victims.⁸⁰ By contributing to a culture of truth and reparation, these investigations help to consolidate the principles of the rule of law. The use of AI in these investigations opens up new possibilities for optimizing the search for, and analysis of, evidence, boosting efficiency during the various stages of this process.⁸¹

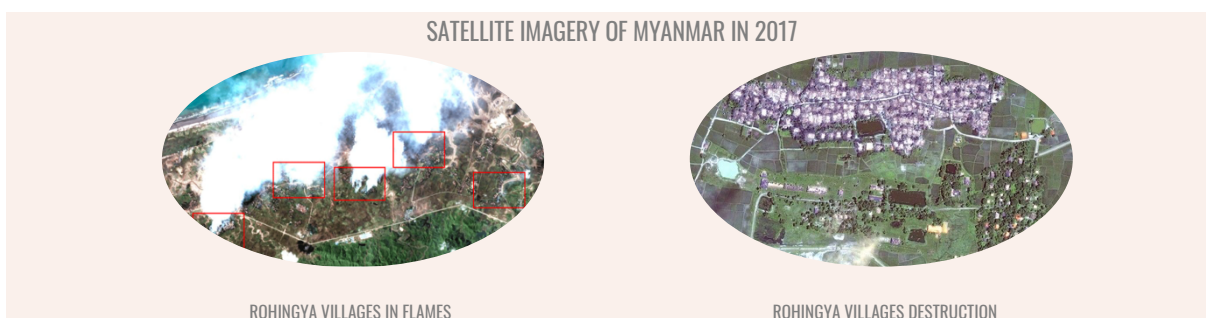
‘SUPPORTIVE POTENTIAL OF AI IN HUMAN RIGHTS IS IN FIGHTING CRIMES OR SUPPORTING LAW ENFORCEMENT OFFICIALS IN THEIR MISSION OF DETECTING, PREVENTING, AND INVESTIGATING CRIMES.’

Inês Gonçalves Ferreira (Associate Programme Management Officer at UNICRI), April 2024.

USE CASE 4.1: AI-ASSISTED SATELLITE IMAGERY & DRONE DATA

Human Rights Watch (HRW) has integrated AI and other advanced technologies into its human rights investigations, with the aim of improving its methods. Previously, HRW's field investigations included interviews, crime scene visits and analysis of court and hospital records. However, due to the many challenges encountered in inaccessible or dangerous regions that limited the efficiency of investigations, HRW collaborated with Element AI to develop tools to improve their effectiveness. One of the key technologies is remote sensing, which enables conflict zones to be investigated in real time, even when they are difficult to access, using satellite and drone data. According to HRW, this tool proved efficient during the 2017 ethnic conflicts in Myanmar, when they were able to quickly detect burning Rohingya villages using thermal data from environmental satellites. AI has enhanced the examination of large quantities of data to detect anomalies that a human cannot perceive, enabling the extraction of digital evidence confirming testimonies and the extent of ethnic violence. In addition, to better understand the extent of the damage caused, HRW uses photogrammetry and 3D modeling to generate precise spatial representations of conflict areas. Finally, thanks to on-site AI such as the NVIDIA DGX Station, an AI-focused computing workstation designed for data scientists and researchers, the US-registered organization can analyze sensitive data ethically without compromising confidentiality. These technologies enable HRW to improve its effectiveness in promoting accountability and justice in the face of human rights violations.⁸²

FIGURE 4: OBJECT DETECTION USING SATELLITE IMAGERY^{82,83}



USE CASE 4.2: AI'S CONTRIBUTION TO SAFEGUARDING CHILDREN'S RIGHTS

The fight against child trafficking and prostitution is a crucial issue, against which major organizations such as End Child Prostitution, Child Pornography and Trafficking of Children for Sexual Purposes (ECPAT International) play an essential role as torchbearers, working to defend children's rights.⁸⁴ This fight has taken on a whole new dimension with the arrival of technologies that can help protect children. For example, Thorn, an US-registered NGO formerly known as the DNA Foundation, has helped technology companies in the fight against the spread of child sexual abuse material (CSAM), by creating a human-centred AI called Safer. This tool enables known and unknown CSAM to be detected, examined and reported, through advanced techniques such as hashing, matching and AI/ML predictive classifier. Safer is notably used by companies such as VSCO to secure their online platform, having enabled them to report 35,000 CSAM files to the authorities in three years. Furthermore, the impact of this technology has been significant since 2019, with over five million CSAM files notified for deletion from the Internet.⁸⁵ According to Thorn, this AI speeds up investigations by identifying suspect files much more rapidly, supports investigations with evidence in prosecutions, and enables investigators to prioritize unknown CSAM cases in order to rescue victims more promptly.

5. Human Rights Governance

Although it mainly draws its legitimacy from the UN, and especially the UNHRC, the global human rights regime remains relatively fragmented in terms of institutions and has experienced profound changes in the last two decades with the proliferation of non-state actors (NSAs), advocacy networks and a complex landscape of standards.⁴⁵ Additionally, despite their efficiency, human rights courts are saturated with cases and lack the necessary resources to function optimally.² Maintaining cohesion, coherence and collaboration between the various components of the regime (institutional or normative) is of primary importance to ensure a proper protection of fundamental human rights at the global level. Whether it is in court, at world summits or online, the use of AI can result in highly positive outcomes for the whole regime by accelerating legal processes and strengthening collaboration between relevant actors.

USE CASE 5.1 CASE DRAFTING & LEGAL ANALYSIS

AI can be used to draft and select cases, as well as to analyze and summarize extensive lists of legal documents, and thus liberating a substantive amount of time for the Court's personnel.⁸⁶ Human Rights courts such as the European Court of Human Rights are suffering from an important backlog of cases recently.² A former president of the Court allegedly said that, if done responsibly,⁸⁷ the use of information technology (IT), and more specifically AI, could improve the situation by helping to manage the Court's docket.² Additionally, it could be used to analyze patterns among cases, draw insights from precedent ones and identify relevant criteria that could better inform decision-making processes. Without necessarily implying automated decision-making, providing judges with detailed information in a short period of time might be a real value-added for the Court. Currently, such technologies are mainly considered for use at the court level, but an actual application would require a full ordinance from the private sector regarding such software and accurate accountability. Similar technologies are already used by lawyers to advise their clients and to decide whether to

litigate or not, or by researchers to better understand what influenced a judgment outcome.² Indeed, not only AI-powered legal analysis can accelerate legal processes, but it can also serve to better inform policymakers, further strengthening cross-sectoral collaboration within the human rights regime.⁸⁸

USE CASE 5.2 ENHANCING NGO COLLABORATION

NGOs and other non-state actors, such as transnational advocacy networks (TANs), are playing an increasingly important role in human rights governance today by monitoring human rights violations in specific locations, conducting field investigations and bypassing formal channels in relaying important information.⁴⁵ However, NSAs are also struggling from a lack of resources, training and time. By connecting many NGOs worldwide and teaching them to use AI for their benefit, the nonprofit initiative ConnectAID helps NSAs to better communicate and connect with other stakeholders and to collect more funds using LLMs such as ChatGPT.⁸⁹ Indeed, NGOs do not necessarily have the time and resources to extensively communicate on their work, nor their achievements. This may eventually undermine their impact, fragilize human rights protection and further fragment the whole regime. By understanding and using AI to generate online content, NGOs can better target and sensibelize their audience by adjusting the language, selecting the rights hashtags, finding the facts or sources that will be the most relevant or simply publishing extensively to increase their visibility.⁸⁹ It is nonetheless important to maintain a certain level of authenticity when using algorithmic models. Connect AID claims that by ensuring the proper use of AI, it helps NGOs in the human rights and development sectors to collaborate more effectively. This collaboration starts online and continues in practice at world summits, rather than operating separately in silos.

Section 4: Managing the Risks of Adopting AI

Risks and challenges are inherent to the use of AI and need to be managed to harness AI's potential. Ethics surrounding the responsible use of AI have also spurred a growing body of research, along with the development of guidelines and frameworks to ensure ethical practices. However, as one interviewee noted, the complexities of applying a uniform ethical standard across different cultures and contexts highlight the need to consider cultural specificities in ethical considerations.⁵⁹ Various challenges include the production of inaccurate outputs, the perpetuation of biases inherent in training data, the lack of 'explainability' (the ability to identify the features or data sets that lead to particular decisions or predictions), the potential for large-scale misinformation and malicious use, and the uneven distribution of resources and benefits. As noted by various subject matter experts, AI tools and techniques can be misappropriated for harmful uses, despite being designed for positive purposes. The following points propose areas of improvement to limit such harmful uses.

ENSURING FAIRNESS & REDUCING BIAS

Algorithmic systems can reflect biases from their developers to the data sets used for their training. Data that is compromised by bias and discrimination can be detrimental as rights violations are more concentrated among minorities and vulnerable populations such as children, women, or disabled people.^{1,20,22} Employing algorithmic systems in public sector settings to evaluate an individual's eligibility for social services like humanitarian aid may lead to further discriminate effects based on factors such as socio-economic status, geographic locations, and other analyzed data points.²² Additionally, when used in the rulemaking process of human rights law through automated decision-making or predictive policing (e.g., Use Case 3.2), AI can introduce bias and discrimination that further undermine legitimacy and perpetuate inequalities.²

ENHANCING PRIVACY, DATA PROTECTION & CYBERSECURITY MEASURES

Projects and initiatives leveraging AI for human rights often require the access, collection and analysis of vast amounts of personal data, such as health or financial data of vulnerable populations (e.g., Use Cases 1.2, 2.2, and 4.1). While organizations are aware of the potential harm from data system breaches, resource constraints may impede their ability to employ the latest cybersecurity measures.³⁵

PREVENTING MANIPULATION & MALICIOUS USE

Manipulation and malicious use encompass the ability of AI to generate and disseminate fake news, scams, hate speech, and actions through social media or other channels of communication with unmatched speed and ease.²⁶ In 2022, a UN report revealed instances where misinformation was exploited to fuel hatred against marginalized communities and prevent civilian access to humanitarian routes.⁹⁰ A huge issue also concerns AI's ability to inform and influence people's conduct by profiling, identifying and tracking them.²⁶



IMPROVING EXPLAINABILITY OF DECISIONS MADE BY COMPLEX AI MODELS

Many AI systems operate as ‘black boxes’, making it difficult to identify the data or logic used to arrive at certain decisions. This is especially relevant for generative AI solutions (e.g., Use Case 1.1), which can eventually produce inaccurate or harmful responses.³⁹ Explaining the results from large and complex AI models in a way that is understandable to humans is essential for the application of AI in human rights. Explainable AI models have several advantages, especially for non-profit organizations, because they enhance transparency, build trust with different stakeholders such as data providers, and ensure accountability for model outcomes.⁸⁴

While the responsible use of AI for human rights is large, some AI-specific challenges will need to be overcome. The types of risks and their magnitude differ considerably from case to case. Richard Benjamins (CEO of OdiselA) notes that ‘early detection and mitigation of risks are more manageable and cost-effective during development than after deployment’.⁷⁶ In an evaluation of how diverse consequences are being addressed within a human rights framework, it became evident that the influence of AI on human rights is not uniformly distributed in society. In certain instances, specific AI implementations may enhance the enjoyment of human rights for some, while concurrently diminishing it for others.³⁴

Mitigation Strategies

Examining these risks and potential harms becomes useful for assessing mitigation strategies and appropriate methods for developing and deploying AI for human rights. For instance, Data Science for Social Good, an initiative launched at the University of Chicago in 2013 that uses data science to create positive social impact, has developed bias detection tools that enable developers to audit science systems for bias and equity.⁹¹ Additionally, the Allen Institute for AI has recently introduced a platform for comparing large text data sets to measure the prevalence of toxic, low-quality, duplicate, or personally identifiable information used in training various LLMs.⁹² Furthermore, some of the use cases presented in the report (e.g., Use Case 2.1 and 4.2) require access to data of vulnerable populations that are at risk of cybersecurity threats. Several organizations have formulated data privacy guidelines, tool lists and custom security frameworks tailored for non-profit organizations with limited resources.⁹³ Furthermore, Disha, a UN-led multi-partner initiative, unites academic, AI ethics hubs, data suppliers as well as foundations and technology firms to construct responsible AI solutions. Among its initial projects is a collaboration with a telecom company to develop a tool assisting multiple NGOs in Asia with disaster resilience and recovery efforts using mobile data. Interdisciplinary collaboration among different stakeholders can enable the deployment of enhanced, uniform and joint AI products for the benefit of human rights.⁹⁴

While the above-mentioned examples represent some possible mitigation strategies, it is important to note that this constitutes only a limited list useful for risk mitigation and further efforts are necessary.

Section 5: Expanding the Responsible Use of AI

Entities that aim to deploy AI for human rights can face many challenges in regard to expansion. Based on desk research and interviews with subject matter experts, the following ‘barriers’ were identified that could stand in the way of expanding responsible use of AI for human rights.

TABLE 1: BARRIERS LIMITING THE USE OF AI FOR HUMAN RIGHTS

DATA CHALLENGES	AI TALENT SHORTAGES	IMPLEMENTATION HURDLES
<ul style="list-style-type: none"> • Availability: Scarcity of relevant datasets, particularly in less-developed regions and among vulnerable demographics. • Accessibility: High costs and proprietary restrictions limit access to rich datasets for organizations with budget constraints. • Quality: Existing data often has issues such as missing entries, poor organization, and requires extensive cleaning. • Privacy: Privacy concerns, especially regarding sensitive human rights data, restrict data collection and sharing. • Language Diversity: Challenges in curating data across different languages, impacting the inclusivity of AI solutions. 	<ul style="list-style-type: none"> • Expertise Distribution: AI talent is unevenly distributed globally, with significant shortages in low- and middle-income countries. • Competition: Non-profits and governments struggle to attract AI talent due to competition with the private sector offering higher salaries. • Interdisciplinary Skills: Difficulty finding professionals who possess both AI skills and domain-specific knowledge (e.g., law, regulation, local cultures). • Training Programs: Lack of sufficient training programs that focus on the intersection of AI and human rights. • Retention: Challenges in retaining skilled AI professionals in non-profit and governmental sectors. 	<ul style="list-style-type: none"> • Integration: Difficulty integrating AI systems with existing human rights workflows and infrastructure. • Cost: High costs associated with AI development, implementation, and maintenance hinder accessibility for many organizations. • Training Needs: Ongoing training required to understand AI model results is time- and resource-intensive. • Infrastructure Limitations: Organizations may face infrastructure constraints that limit the effective deployment of AI solutions. • Ethical Concerns: Ensuring AI solutions align with ethical standards and do not inadvertently harm vulnerable populations.

The recurring themes impeding AI-driven progress target data challenges, AI talent shortages as well as final obstacles associated with implementing AI solutions for human rights. Therefore, this report proposes three interconnected, overarching strategies to address the identified ‘barriers. These approaches can be adopted by mission-driven organisations, governments, foundations, academia, developers, businesses, human rights experts, and other stakeholders to support the use of AI for human rights.

1. Developing a Data Strategy to Bolster Quality & Usability

Data is crucial for the success of AI initiatives, both within the sector and beyond. AI systems rely on data patterns and information to learn and make predictions. Therefore, the quality and quantity of data significantly influence the accuracy and effectiveness of AI models.

DATA CHALLENGES: AVAILABILITY, ACCESSIBILITY & QUALITY

Relevant datasets for human rights-related topics are challenging to create or curate due to factors like data scarcity in less-developed regions and from more vulnerable demographic groups (which are also the intended beneficiaries of many human rights initiatives), language diversity, and privacy concerns. Existing rich datasets may be privately owned or expensive, hindering access for organizations with budget constraints. In addition, accessible data may suffer from quality issues like missing entries or poor organization, which necessitates resource-intensive data cleaning processes.³⁹

'THERE IS A PROBLEM OF GETTING ACCESS TO THE DATA AND BUILDING A SUSTAINABLE ECOSYSTEM AROUND THE DATA FOR CREATING VALUE. IT IS A WELL-UNDERSTOOD PROBLEM.'

Richard Benjamins (CEO of OdiselIA), March 2024.



HOW TO ADDRESS THIS BARRIER: DATA COLLECTION, STRATEGY & STANDARDS

The reliance of AI on high-quality data indicates that it can no longer be considered merely an add-on. Stakeholders should therefore enhance data-driven AI for human rights applications by developing high-quality datasets, which will improve scalability, reduce risks, and boost output quality. This may require prioritizing investment in data collection before AI development. Supporting data collection in resource-poor areas and for marginalized populations is crucial, as is creating incentives for organizations with rich datasets to share them. When faced with limited resources, organizations may explore creative solutions, leverage alternative technologies, or unconventional methods to gather relevant data. The Data for Development Network (D4D.net) is an alliance of Global South-based research organizations working on enhancing data collection capacities, particularly in resource-poor settings, and fosters collaboration to ensure data is used effectively for development purposes and in support of marginalized communities.⁹⁵ Initiatives like Flowminder exemplify efforts to ethically utilize anonymized telecommunication data to understand the dynamics of human mobility in low-income countries without compromising privacy or exploiting vulnerable populations.⁹⁶

An effective data strategy is integral to tech companies and should become an essential part for AI initiatives as it helps plan systematically and avoid pitfalls in data-dependent initiatives. For example, Grammarly and Casetext, which are writing and legal research forecasting tools respectively, rely on an ongoing feedback loop of data to continuously improve their performance.

In future endeavours, data strategy could then be used as a standard requirement for future projects. AI initiatives can explore leveraging pre-existing curated datasets, necessitating standardized data management practices to enable their utilization.⁹⁷ There are potential advantages in establishing collaborative data standards, especially involving projects focused on specific thematic domains. This collaborative approach would ensure fair access to data for all innovation teams involved. For example, the Humanitarian Data Exchange managed by the UN Office for the Coordination of Humanitarian Affairs (OCHA) is an open platform for sharing data about crises and human rights abuses. It provides access to more than 20,000 curated datasets from various humanitarian organizations.⁹⁸ Another example is the World Bank Global Data Facility, a global funding tool that supports long-term improvements in data systems and data capital in low- and middle-income countries to enhance lives and protect the planet.⁹⁹

2. Implementing Mentorship & Training Programs to Expand the Pool of AI for Human Rights Talent

Creating new AI technology poses significant challenges, as it demands both technical expertise and a deep understanding of the human rights field or the specific regions being targeted. Therefore, it is crucial to invest in skilled AI talent.

AI TALENT SHORTAGES

While the supply of AI talents has increased globally, it remains unevenly distributed, with limited access for governments, social enterprises, and non-profit organizations, especially in low-and-middle income countries.¹⁰⁰ Competition from the private sector further complicates talent acquisition, and the combination of AI skills with domain-specific knowledge such as on law and regulation, specific organizational contexts, or local culture poses additional challenges.³⁹

TEXT BOX 4: INSIGHTS FROM STANFORD HAI'S INDEX REPORT 2024

The AI Index Report for 2024 by Stanford HAI indicates a notable rise in the acquisition of AI skills from 2015 to 2023. Germany, India, and the U.S. had the highest rates of AI skill penetration, with the U.S. leading by a significant margin, 2.2 higher than in the rest of the world.



HOW TO ADDRESS THIS BARRIER: TRAINING & CAPACITY BUILDING

Expanding the pool of AI talent for human rights entails implementing initiatives, 'either in-house if AI is to become a core part of [their] value proposition, or through strategic partnership, if not'. In the short term, academic institutions and tech companies possessing technical expertise could allocate their talent to support low-resource organizations. In the long term, investments in training programs or relevant educational degrees can foster growth.

For example, the AI for Good Innovation Factory is an UN-led pitching platform to help start-ups grow and scale their innovative AI-powered and SDG-driven solutions by providing

mentoring and connecting them to different stakeholders.¹⁰¹ Large companies, such as Google and Microsoft have used their data science expertise by either lending or seconding talent to other organizations, or by allocating time for their employees to contribute to AI-powered initiatives.¹⁰² As AI talent needs to have a deep understanding of the human right field, educational initiatives should expand their talent in interdisciplinary fields. Furthermore, OpenAI Scholars provided stipends and mentorship to individuals from underrepresented groups to study deep learning and open-source projects.¹⁰³ Similarly, the Google DeepMind scholarship program extends its reach to universities globally, including partner institutions across the African continent.¹⁰⁴

3. Strengthening Partnerships & Stakeholder Collaboration to Accelerate Positive Impact

Interdisciplinary collaboration can address needs of mission-driven organizations and mitigate numerous challenges. While researchers or tech companies may develop AI applications, their impact relies on the adoption by frontline organizations. Partnerships can facilitate this connection and accelerate the positive impact to the front lines.

IMPLEMENTATION HURDLES

Successful implementation of AI tools requires frontline workers to be receptive to AI-powered solutions. Therefore, organizations must adapt their procedures to integrate new methods of operation and overcome infrastructure limitations. Addressing these issues may require substantial funding. They may face additional barriers such as the need for ongoing training to understand the results produced by AI models, which can be time- and resource-intensive. Concerns about AI-related risks (as discussed above), particularly those affecting vulnerable populations, could also discourage organizations from adapting AI solutions. Moreover, an expert stressed the importance of developing practical and accessible technologies, noting that ‘overly sophisticated monitoring systems that are not feasible in [for example] typical healthcare settings’ can hinder broad implementation as well as equitable access.¹⁰⁵ Additionally, experts cited the cost of implementation as a significant barrier to AI adoption. Thus, it is important to plan for collaboration, for example with gatekeepers to data, at all stages of the AI development. Better collaboration will require initiatives to focus more on transparency and explainability of their AI systems, thus promoting understanding, trust, and acceptance among a wider group of stakeholders.

‘EFFECTIVE IMPLEMENTATION REQUIRES A MULTI-DISCIPLINARY APPROACH, SUFFICIENT AI LITERACY, AND ACCESS TO NECESSARY RESOURCES, ESPECIALLY FOR COMMUNITIES AND SECTORS THAT ARE LESS ALFUEENT.’

Danielle Ralie (CEO & Founder of Ancora.ai), April 2024.

HOW TO ADDRESS THIS ‘BARRIER’: **INTERDISCIPLINARY COLLABORATION EFFORTS**

Interdisciplinary collaboration among various stakeholders can fulfil needs of mission-driven organizations and alleviate the challenges discussed above. This also requires collaboration across multiple levels - international bodies like the UN, regional organizations like the Association of Southeast Asian Nations (ASEAN), and national governments.¹⁰⁶ To

ensure AI technologies are effective, they must incorporate the practical insights from frontline organizations and human rights experts, even though they are frequently developed within academic or corporate settings. Collaborations can expedite the deployment of responsible AI for human rights to areas where they are most needed. These partnerships can also facilitate the sharing and development of critical resources like data, technological infrastructure, and applications. This synergy not only accelerates the deployment of innovative solutions but also fosters trust across various sectors, ultimately strengthening the human rights framework.^{78,107}

The Schwab Foundation's Global Alliance for Social Entrepreneurship launched an initiative on AI for Social Innovation, co-initiated by Microsoft and supported by EY. This collaboration between social innovators, impact investors, academic, ecosystem actors and technology leaders – such as Microsoft, SAP, Salesforce and Verizon – aims to mobilize support for social innovators to adopt AI for positive impact.¹⁰⁸ This initiative, as well as the previously mentioned UN-led Disha project, enable the strengthening of partnerships and stakeholder collaboration to accelerate positive impact.

Conclusion

The progress of AI and its related technologies – from machine learning to NLP, computer vision, and predictive analytics – has coincided with a growing number of successful AI for human rights deployment. Expanding their use for addressing different human rights areas will require interdisciplinary collaboration to ensure access to talent, data solutions, and AI applications and models.

Today's AI initiatives for human rights already enhance human rights education, training and awareness through personalized learning experiences and broader access to resources; aid in monitoring by analyzing vast data to detect abuse patterns and issue real-time alerts; ensure compliance through automated reporting and regulation adherence; assist investigations by processing evidence efficiently and revealing critical insights and streamline governance decision-making processes. However, leveraging responsible AI for human rights at scale entails certain risks that need careful management. These include the potential for inaccurate outputs, reinforcing existing biases in training data, lack of transparency in AI decisions, the spread of misinformation, malicious use, and unequal access to AI benefits. Addressing these challenges requires ongoing efforts to mitigate risks and ensure the AI is beneficial and fair for all.

Using responsible AI for human rights is a collaborative effort. To achieve optimal outcomes and create positive impacts, businesses, human rights experts, academic institutions, civil society, and governments must work together, breaking down barriers, combining processing power, and sharing expertise, knowledge, and relevant data.

Bibliography

- 1 Dulka A. The Use of Artificial Intelligence in International Human Rights Law. *Stanford Technology Law Review* 2023; **26**.
- 2 Molbæk-Steensig H. AI at the European Court of Human Rights : technological improvement or leaving justice by the wayside? *Ordine internazionale e diritti umani* 2022; **5**: 1254–67.
- 3 Bacalu F. Biometric Facial Recognition Technology, Law Enforcement Algorithmic Automation, and Data-driven Predictive Policing Systems in Human Rights Protections and Abuses. *Review of Contemporary Philosophy* 2022; : 38–54.
- 4 Ramcharan B, Brett R, Clark AM, Parker P. The Protection Roles of Human Rights NGOs: Essays in Honour of Adrien-Claude Zoller. Brill | Nijhoff, 2023 DOI:10.1163/9789004516786.
- 5 Black box. 2024; published online June 12. <https://dictionary.cambridge.org/dictionary/english/black-box> (accessed June 13, 2024).
- 6 Kosta E. Algorithmic state surveillance: Challenging the notion of agency in human rights. *Regulation & Governance* 2022; **16**: 212–24.
- 7 Lăpădat RA. Digital constitutionalism: a perspective over the increasing role of the private actors in securing the exercise of human rights. *Perspectives of Law and Public Administration* 2022; **11**: 157–64.
- 8 Almeida D, Shmarko K, Lomas E. The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. *AI Ethics* 2022; **2**: 377–87.
- 9 How Hashing and Matching Can Help Prevent Revictimization - Thorn. <https://www.thorn.org/blog/hashing-detect-child-sex-abuse-imagery/> (accessed June 13, 2024).
- 10 Hertz N. Neurorights – Do we Need New Human Rights? A Reconsideration of the Right to Freedom of Thought. *Neuroethics* 2023; **16**: 5.
- 11 Bandy AB, Reitmeier EA, Suárez MC, Diaz S. The US Government Is Using AI To Detect Potential Wrongdoing, and Companies Should Too | Insights | Skadden, Arps, Slate, Meagher & Flom LLP. Skadden. <https://www.skadden.com/insights/publications/2024/03/insights-special-edition/the-us-government-is-using-ai> (accessed May 20, 2024).
- 12 United Nations General Assembly. (2024). Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development (A/78/L.49).

- 13 Human Rights. United Nations and the Rule of Law. <https://www.un.org/ruleoflaw/thematic-areas/human-rights/> (accessed May 20, 2024).
- 14 The Toronto Declaration. The Toronto Declaration. 2018. <https://www.torontodeclaration.org/declaration-text/english/> (accessed Oct 26, 2023).
- 15 H.R.6216 - 116th Congress (2019-2020): National Artificial Intelligence Initiative Act of 2020 | Congress.gov | Library of Congress. <https://www.congress.gov/bill/116th-congress/house-bill/6216> (accessed May 20, 2024).
- 16 S.2551 - 117th Congress (2021-2022): AI Training Act | Congress.gov | Library of Congress. <https://www.congress.gov/bill/117th-congress/senate-bill/2551> (accessed May 20, 2024).
- 17 Government Bill (House of Commons) C-27 (44-1) - First Reading - Digital Charter Implementation Act, 2022 - Parliament of Canada. <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading> (accessed May 20, 2024).
- 18 Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). 2021 <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206> (accessed Dec 25, 2023).
- 19 Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization (UNESCO). <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics> (accessed May 20, 2024).
- 20 Abashidze AKh, Goncharenko OK. Protection of Women from Violence and Domestic Violence in the Context of Digitalization. In: Inshakova AO, Frolova EE, eds. *Smart Technologies for the Digitisation of Industry: Entrepreneurial Environment*. Singapore: Springer Singapore, 2022: 179–86.
- 21 Tymoshenko AA. Binding Decisions of the European Court of Human Rights on the Transformation of Russian Legislation in the Era of Digitalization. *Russian Journal of Legal Studies* 2022; **9**: 9–14.
- 22 Tesson Christelle, Stevens Yuan, Malik Momin M., Solomun Sonja, Dwivedi Supriya, Andrey Sam. AI Oversight, Accountability and Protecting Human Rights: Comments on Canada’s Proposed Artificial Intelligence and Data Act. *Cybersecure Policy Exchange, Center for Information Technology Policy, Centre for Media, Technology, and Democracy* 2022. <https://www.cybersecurepolicy.ca/aida>.
- 23 Mantelero A, Esposito MS. An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review* 2021; **41**: 105561.
- 24 Rachovitsa A, Johann N. The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case. *Human Rights Law Review* 2022; **22**. DOI:10.1093/hrlr/ngac010.
- 25 Mrčela M, Vuletić I. Rethinking the Privilege Against Self-Incrimination in Terms of Emerging Neurotechnologies: Comparing the European and United States Perspective. *192023*; **3**: 207–23.

- 26 Mubangizi JC. A human-rights based approach to the use and regulation of artificial intelligence - an african perspective. *Journal of Southwest Jiaotong University*2022; **57**: 551–61.
- 27 Tobi ES, Liam C, Meyers DG. Neurotechnology: Redesigning the brain-computer interface technology for criminal procedure purposes. 2022.
- 28 van Slobbe M. Freedom of thinking and neurotechnology: ensuring free thought in the age of brain-computer interfaces. 2022.
- 29 Baselga-Garriga C, Rodriguez P, Yuste R. Neuro Rights: A Human Rights Solution to Ethical Issues of Neurotechnologies. In: López-Silva P, Valera L, eds. *Protecting the Mind*. Cham: Springer International Publishing, 2022: 157–61.
- 30 Sartor G. Artificial intelligence and human rights: Between law and ethics. *Maastricht Journal of European and Comparative Law* 2010; **27**: 705–19.
- 31 Agarwal A. Information Technology vis-a-vis Human Rights: An Analytical and Legal Approach. *Int'l JI Mgmt & Human* 2022; **5 Issue 2**: [xiii]-122.
- 32 Ashok M, Madan R, Joha A, Sivarajah U. Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management* 2022; **62**: 102433.
- 33 Entman R. Framing: Toward Clarification of A Fractured Paradigm. *The Journal of Communication* 1993; **43**: 51–8.
- 34 Raso F, Hilligoss H, Krishnamurthy V, Bavitz C, Kim LY. Artificial Intelligence & Human Rights: Opportunities & Risks. *SSRN Journal*/2018. DOI:10.2139/ssrn.3259344.
- 35 Ienca M. Democratizing cognitive technology: a proactive approach. *Ethics Inf Technol*2019; **21**: 267–80.
- 36 Das R. Artificial Intelligence: Advantages and Disadvantages from the Perspective of Human Rights in India. *Int'l JI Mgmt & Human* 2022; **5**: [xcii]-985.
- 37 Mpinga EK, Bukonda NK, Qailouli S, Chastonay P. Artificial Intelligence and Human Rights: Are There Signs of an Emerging Discipline? A Systematic Review. *JMDH*2022; **15**: 235–46.
- 38 Gabriel S, Han JX, Liu E, *et al*. Advancing Equality: Harnessing Generative AI to Combat Systemic Racism. *An MIT Exploration of Generative AI*2024. DOI:10.21428/e4baedd9.7dc53bbf.
- 39 Bankhwal M, Bisht A, Chui M, Roberts R, van Heteren A. AI for social good: Improving lives and protecting the planet. McKinsey Digital, 2024.
- 40 AI for Good. AI for Good. <https://aiforgood.itu.int/> (accessed May 20, 2024).

- 41 United Nations Committee on Economic, Social and Cultural Rights, General (ECOSOC). Comment No. 13 on the Right to Education (Art. 13 of the Covenant), UN Doc. E/C.12/1999/10 (1999).
- 42 Special Rapporteur on the Right to Education. Call for contributions: artificial intelligence in education and its human rights - based use at the service of the advancement of the right to education. The Office of the High Commissioner for Human Rights (OHCHR). <https://www.ohchr.org/en/calls-for-input/2024/call-contributions-artificial-intelligence-education-and-its-human-rights> (accessed May 20, 2024).
- 43 Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona Spain: ACM, 2020: 469–81.
- 44 Krasner SD (ed.). *International Regimes*, Cornell University Press. 1983.
- 45 Pegram T. Global human rights governance and orchestration: National human rights institutions as intermediaries. *European Journal of International Relations* 2015; **21**: 595–620.
- 46 Artificial Intelligence and Data Act. 2023; published online Sept 27. <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act> (accessed June 13, 2024).
- 47 United Nations High Commissioner for Human Rights. The right to privacy in the digital age. 2021.
- 48 Human rights. Britannica. 2024; published online May 24. <https://www.britannica.com/topic/human-rights> (accessed June 12, 2024).
- 49 The Core International Human Rights Instruments and their monitoring bodies. Office of the United Nations High Commissioner for Human Rights (OHCHR). <https://www.ohchr.org/en/core-international-human-rights-instruments-and-their-monitoring-bodies> (accessed June 11, 2024).
- 50 United Nations General Assembly, Session No. 76 on Promotion and protection of human rights: human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms, Agenda item 74 (b), UN Doc. (A/76/L.75, 2022).
- 51 DiMatteo LA, Poncibò C, Cannarsa M, editors. *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*. Cambridge: Cambridge University Press, 2022 DOI:10.1017/9781009072168.
- 52 Types of Artificial Intelligence. <https://www.ibm.com/think/topics/artificial-intelligence-types> (accessed June 12, 2024).
- 53 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436–44.

- 54 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011; **18**: 544–51.
- 55 Khosravi T, Sudani Z, Oladnabi M. To what extent does ChatGPT understand genetics? *Innovations in Education and Teaching International* 2023; : 1–10.
- 56 Gupta S, Kar AK, Baabdullah A, Al-Khowaiter WAA. Big data with cognitive computing: A review for the future. *International Journal of Information Management* 2018; **42**: 78–89.
- 57 What Is Automation? | IBM. 2021; published online July 27. <https://www.ibm.com/topics/automation> (accessed June 13, 2024).
- 58 What Is an Expert System? | Definition from TechTarget. Enterprise AI. <https://www.techtarget.com/searchenterpriseai/definition/expert-system> (accessed June 13, 2024).
- 59 Interview with Emmanuel Goffi (Director of Studies at Human Technology Foundation), April 2024.
- 60 Voeneke S, Kellmeyer P, Mueller O, Burgard W. The Cambridge Handbook of Responsible Artificial Intelligence.
- 61 Guest G, MacQueen KM, Namey EE. Applied Thematic Analysis. SAGE Publications, 2011.
- 62 Flick U, Kardoff E von, Steinke I. A Companion to Qualitative Research. SAGE, 2004.
- 63 United Nations General Assembly (UNGA). United Nations Declaration on Human Rights Education and Training (Art. 2), UN Doc. A/RES/66/137, 2011).
- 64 Shiran Mlamdovsky Somech. Animated Sophia.chat. Generative AI for Good. <https://www.generativeaiforgood.com/animated-sophia-chat> (accessed June 10, 2024).
- 65 United Nations Office on Drugs and Crime. Gender-related Killings of Women and Girls (Femicide/feminicide): Global Estimates of Female Intimate Partner/family-related Homicides in 2022. United Nations, 2023 DOI:10.18356/9789213587072.
- 66 Switzerland M. D-ID's Generative AI to Power Online Chatbot for Victims of Domestic Violence. Microsoft Switzerland News Center. 2023. <https://news.microsoft.com/de-ch/2023/03/08/d-ids-generative-ai-to-power-online-chatbot-for-victims-of-domestic-violence/> (accessed May 20, 2024).
- 67 Interview with Shiran Melamdovsky Somech (Founder of Generative AI for Good), April 2024. .
- 68 Husain S, Vaishnav M. Educate Girls NGO. Educate Girls NGO. <https://www.educategirls.ngo/> (accessed June 10, 2024).

- 69 Naidu E. Using AI technology, innovative Educate Girls wins WISE Prize for Education. Inside Education - Inspiring Minds. 2023; published online Dec 18. <https://insideeducation.co.za/using-ai-technology-innovative-educate-girls-wins-wise-prize-for-education/> (accessed June 10, 2024).
- 70 Sampson R, McManus J, Brockman B, Ravinutala S. Educate Girls: improving learning outcomes for millions of children in India. IDinsight. <https://www.idinsight.org/project/educate-girls-improving-learning-outcomes-for-millions-of-marginalized-children-in-india/> (accessed May 20, 2024).
- 71 ASER: Annual Status of Education Report. <https://asercentre.org/> (accessed June 12, 2024).
- 72 Ravinutala S. Rebuilding the Educate Girls machine learning model. IDinsight. 2019. <https://www.idinsight.org/article/rebuilding-the-educate-girls-machine-learning-model/> (accessed May 20, 2024).
- 73 OHCHR. Chapter 02-Basic Principles of Human Rights Monitoring. In: Manual on Human Rights Monitoring. 2001. <https://www.ohchr.org/sites/default/files/Documents/Publications/Chapter02-MHRM.pdf> (accessed May 20, 2024).
- 74 de Beco G. Human Rights Monitoring Mechanisms of the Council of Europe. Abingdon: Routledge., 2012.
- 75 Benjamins R, Vos J, Verhulst S. Mobile Big Data in the fight against COVID-19. *Data & Policy* 2022; **4**. DOI:10.1017/dap.2021.39.
- 76 Interview with Richard Benjamins (Co-founder and CEO of OdiselA), March 2024.
- 77 Martin K. Ethical Implications and Accountability of Algorithms. *J Bus Ethics* 2019; **160**: 835–50.
- 78 Interview with Shea Brown (Founder and CEO of BABL AI), April 2024.
- 79 Interview with Jean Ng (Founder of JHN Studio), April 2024.
- 80 Uganda Human Rights Commission. Human Rights Investigators' Handbook. 2014.
- 81 Interview with Cecilia Garcia Podoley (Lawyer & Board Member at Ethics & Compliance Switzerland (ECS)), April 2024.
- 82 Salian I. AI in the Sky Aids Feet on the Ground Spotting Human Rights Violations. NVIDIA Blog. 2019; published online April 4. <https://blogs.nvidia.com/blog/human-rights-watch-ai-gtc/> (accessed May 20, 2024).
- 83 Burma: New Satellite Images Confirm Mass Destruction | Human Rights Watch. 2017. <https://www.hrw.org/news/2017/10/17/burma-new-satellite-images-confirm-mass-destruction> (accessed June 13, 2024).

- 84 Interview with Tatiana Caldas-Löttiger (Founder & CEO, International WoMenX In Business For Ethical AI (IWIB4AI)), April 2024. .
- 85 How Thorn Makes the Internet Safer. Thorn. <https://www.thorn.org/blog/how-thorn-makes-the-internet-safer-helps-stop-the-cycle-of-abuse/> (accessed May 20, 2024).
- 86 Interview with Oana Ichim (Professor at the Geneva Graduate Institute), April 2024. .
- 87 AI is creating fake legal cases and making its way into real courtrooms, with disastrous results. UNSW Sites. <https://www.unsw.edu.au/newsroom/news/2024/03/AI-creating-fake-legal-cases-disastrous-results> (accessed June 11, 2024).
- 88 Interview with Giuliano Borter (Senior AI Policy Officer at Center for AI & Digital Policy), April 2024. .
- 89 Interview with Gaelle Mogli (Founder & President of ConnectAID), March 2024. .
- 90 Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Office of the High Commissioner for Human Rights, United Nations. 'Disinformation and Freedom of Opinion and Expression during Armed Conflicts.' (Report A/77/288, 2022). .
- 91 Center for Data Science and Public Policy. The bias and fairness audit toolkit for machine learning. Aequitas Bias & Fairness Audit. <http://aequitas.dssg.io> (accessed May 20, 2024).
- 92 Elazar Y, Bhagia A, Magnusson I, *et al.* What's In My Big Data? 2023. DOI:10.48550/ARXIV.2310.20707.
- 93 Online Privacy for Nonprofits: A Guide to Better Practices. Electronic Frontier Foundation. 2022; published online Aug 19. <https://www.eff.org/fr/node/106965> (accessed May 20, 2024).
- 94 Disha - Unlocking data and AI solutions for social impact. UN Global Pulse. <https://disha.unglobalpulse.org/> (accessed May 20, 2024).
- 95 Data for Development Research Hub. D4D.net Data for Development. <https://www.d4d.net/> (accessed May 20, 2024).
- 96 Innovative solutions from mobile data analytics. Flowminder. <https://www.flowminder.org> (accessed May 20, 2024).
- 97 Sykes P. AI & Humanitarianism - Keeping the human in humanitarianism. 2023.
- 98 Humanitarian Data Exchange (HDX). OCHA Services. <https://data.humdata.org/> (accessed May 20, 2024).
- 99 Global Data Facility. The World Bank. <https://www.worldbank.org/en/programs/global-data-facility/about> (accessed May 20, 2024).

- 100 Artificial Intelligence Index Report 2024. Stanford University Human-Centered Artificial Intelligence, 2024 https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf (accessed May 20, 2024).
- 101 AI for Good Innovation Factory. <https://aiforgood.itu.int/about-ai-for-good/innovation-factory/> (accessed May 20, 2024).
- 102 Our Work. Google.org. <https://www.google.org/intl/de/our-work/> (accessed May 20, 2024).
- 103 OpenAI Scholars. <https://openai.com/index/openai-scholars/> (accessed May 20, 2024).
- 104 About: Education. Google DeepMind. <https://deepmind.google/about/education/#> (accessed May 20, 2024).
- 105 Interview with Danielle Ralic (CEO & Founder of Ancora.ai), April 2024. .
- 106 Interview with Ansgar Koene (EY Global AI Ethics & Regulatory Leader), April 2024. .
- 107 Interview with Walid el Abed (CEO & Founder of Global Data Excellence), April 2024. .
- 108 World Economic Forum. AI for Social Innovation. Schwab Foundation’s Global Alliance for Social Entrepreneurship. <https://initiatives.weforum.org/global-alliance-for-social-entrepreneurship/ai-for-social-innovation> (accessed May 20, 2024).

Annex 1: Interview Guide

1. INTRODUCTION

2. POSITIVE APPLICATIONS OF AI IN THE HUMAN RIGHTS CONTEXT

- **Positive Application:** Could you share insights into your projects that demonstrate the positive use of AI technology, specifically highlighting its potential to enhance and broaden the enjoyment of human rights? This includes its application within the fields of human rights education, governance, compliance and monitoring, and investigations.
- *In the absence of specific examples:* How do you envision such initiatives being applied within a human rights framework, particularly in the field of human rights education, governance compliance and monitoring, and investigations.

3. RISKS, CHALLENGES & ETHICAL CONSIDERATIONS

- **Risks and Challenges:** Regarding the opportunities and/or positive applications at hand, are there any associated risks, challenges and ethical considerations?
- **Possible Mitigation Strategies:** If so, how can they be efficiently managed and mitigated?

4. THE DIVERSE ROLES OF GOVERNMENTS, PRIVATE & HUMAN SECTOR, INTERNATIONAL ORGANIZATIONS, ACADEMIA & CIVIL SOCIETY

- **Particular Role of your Sector:** How do you define your contribution and position within the intersection of AI and human rights?
- **Roles of Other Stakeholders and Collaboration:** What roles do you believe other stakeholders should undertake to effectively harness AI for the advancement of human rights? And how can your specific sector collaborate with them to maximize the potential of AI in promoting human rights? Can you give concrete examples?
- **Strengthening Interdisciplinary Engagement:** How can we enhance interdisciplinary collaboration in various aspects of AI and human rights, including technology development, deployment, governance, regulation, and oversight?

5. OUTLOOK: FUTURE DIRECTIONS & RECOMMENDATIONS

- **Envisioning the Future of AI and Human Rights:** How do you foresee the future of AI technologies positively impacting human rights, and are there any emerging trends or developments that you find particularly promising or concerning?
- **Role of your Sector in Shaping Perceptions:** How does your sector contribute to shifting the prevailing negative narrative about AI's impact on human rights towards a more positive outlook, while ensuring risks (such as transparency and accountability, etc.) in AI systems deployed for human rights purposes?
- **Framework Recommendations:** In light of the current discourse and challenges, what specific framework recommendations would you propose to guide the responsible deployment of AI within your sector to prioritize and safeguard human rights?

6. CLOSING THOUGHTS & ADDITIONAL RESOURCES

- Do you have any final thoughts or key points you would like to share with us? Would you like to share any resources, articles, publications related to our discussion today?