THE HUMAN RIGHTS PLATFORM

GENEVA

CSS ETH Zürich

WORKING PAPERS

AI Decoded:
Key Concepts and Applications of Artificial Intelligence for Human Rights and SDG Monitoring

January 2025

Milica Mirković and Jennifer Victoria Scurrell

This publication has been externally peer reviewed

A project of the: GENEVA ACADEMY | Académie de droit international humanitaire et de droits humains / Academy of International Humanitarian Law and Human Rights

The Geneva Academy, a Joint Centre of GENEVA GRADUATE INSTITUTE | INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT / GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES | UNIVERSITÉ DE GENÈVE FACULTY OF LAW

TABLE OF CONTENTS

INTRODUCTION

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have increasingly been viewed as opportunities to develop solutions that address global challenges and impact various policy fields. The success of AI in commercial applications and its positive influence in areas such as healthcare, access to information, or governance, has demonstrated its potential as a force for social good.[1] Numerous studies highlight the benefits of targeted AI and ML applications for social good yet emphasize the need to balance innovation with responsible use that addresses associated risks.

This paper aims to bridge the knowledge gap between technical and policy-oriented domains by providing basic definitions of AI and ML and offering a concise yet comprehensive overview of AI's use for social good, focusing on how it has contributed and can continue to contribute to tracking SDG progress and monitoring human rights effectively, ethically, and responsibly.

Following an "AI Glossary", which aims to introduce in simple terms key concepts related to AI and ML to non-expert readers, the paper first explores the most relevant features of AI and ML, such as the development of ML models, their main categories, how they function, and their relevance to various stages of data processing. It then shifts to the application of AI for social good, examining how AI can improve monitoring systems through its analytical capabilities.

The paper highlights AI's contributions to SDG tracking, such as improving data availability and supporting effective, informed policymaking through timely insights and predictive analytics, illustrating these contributions with concrete examples. It then discusses how AI can support human rights monitoring by outlining its key advantages and explaining how these features contribute to improved reporting and accountability.

The final section provides a critical discussion of the main risks and ethical considerations surrounding the use of AI for monitoring purposes and beyond, including challenges related to bias, privacy, and transparency, as well as broader environmental, economic, and socio-cultural concerns. It also suggests approaches to AI that enable its responsible and ethical use in support of social good.

1 Nenad Tomašev et al., "AI for Social Good: Unlocking the Opportunity for Positive Impact," Nature Communications 11, no. 1 (2020).

AI GLOSSARY

AI Alignment: the aim of designing AI systems whose objectives and behaviors align with human values and societal goals (Braithwaite, 2024)

Algorithm: An algorithm is a set of rules or instructions that enable computers to solve a problem or perform a task. At its core, it is a sequence of steps that transform an input, ranging from simple data like numbers or text to complex inputs like images, into an output. (Cormen et al., 2022) This output could be a solution, a prediction, or any result the algorithm is designed to achieve. There are three basic categories of algorithms:

1. Linear Sequence Algorithms: Follow a series of steps in order, like a recipe, where each step depends on the previous one.
2. Conditional Algorithms: Use "if/then" decisions to determine actions based on specific conditions.
3. Looping Algorithms: Repeat steps until a condition is met or a set number of repetitions is completed (Scribbr, 2023).

For example, search algorithms help the computer find a particular item in a dataset, while sorting algorithms organize elements of a dataset in a specific order. (Datacamp, 2023)

ChatGPT: an LLM chat-bot created by OpenAI that identifies linguistic patterns in large text datasets and generates new, clear and logical textual responses to specific user prompts, allowing for human-like conversations. Various chat-bots from other providers such as Meta, Anthropic, and Inflection exist. (OpenAI 2022)

Computer Vision: a ML technique that handles image and video data. Here, the algorithm enables the detection of patterns in large amounts of image and video content, allowing AI to track motion, identify faces, and extract features from it. (Zou 2023)

Deep Learning: Deep learning is a branch of ML that uses multiple layers of neural networks to automatically learn complex patterns in data. This approach enables tasks like image and speech recognition with minimal human intervention. (Goodfellow, Bengio, and Courville 2016)

Explainable AI (XAI): the ability of AI systems to deliver clear explanations for their actions and decisions, aimed at making their behavior understandable to humans by clarifying the mechanisms behind their decision-making processes. (EDPS 2023)

Generative AI (GenAI): a type of AI that not only recognizes complex patterns in data (like images, text, and audio) but also creates entirely new content based on this understanding, such as text, images, audio, or video, in response to user prompts. (Bail 2024)

Large Language Models (LLMs): a type of GenAI that is created specifically to generate new textual content. LLMs work on large datasets through deep learning, which allows them to understand text automatically and to generate new content based on user prompts. (cloudflare 2024)

Natural Language Processing (NLP): it involves algorithms that process vast amounts of text, allowing for the statistical interpretation of human language. NLP summarizes information, infers meaning, and generates new content based on the input it receives. (Stryker and Holdsworth 2024)

Neural Networks: a ML model inspired by the structure and supposed function of the human brain. It has layers of artificial neurons (nodes) that connect and communicate, each having its own rules for passing information to the next layer. (IBM) Neural networks "learn" by using data to improve over time, and once trained, they can quickly classify and recognize complex information. (IBM)

Responsible AI: principles that guide the design and use of AI, considering its societal impact. The aim of these principles is to align AI with legal standards and ethical values and integrate them into AI applications to minimize its intended and unintended risks. (IBM 2024)

AI AND ML: KEY CONCEPTS, DEFINITIONS, AND CORE FEATURES

AI is designed to mimic certain aspects of human intelligence. The technology encompasses a range of capabilities that allow computers to process information similarly to how humans do.[2] AI systems can perform tasks that typically require human capabilities, such as the interpretation of information, learning from experience, decision-making, problem-solving, and the generation of new outputs. Unlike traditional software that follows pre-programmed rules such as in expert systems[3], modern AI models typically "learn" to perform these tasks by analyzing vast amounts of training data.[4]

AI operates through various techniques, primarily using algorithms to process data and solve complex problems. One of the subfields of AI is ML, where algorithms enable computers to learn from data and make predictions or decisions without being explicitly programmed for every specific task.[5] In simple terms, ML analyzes large amounts of data and learns to identify patterns (e.g., spot trends in rights violations), detect anomalies in data that signal potential issues (e.g., discover forest changes and track deforestation), and estimate outcomes based on the data it has analyzed (e.g., predict movements of displaced people).[6] ML allows for scaled-up and fast classification and clustering of data, offering approximations that can support evidence-based decision-making in human rights and sustainable development.

The design of an ML model typically involves two main phases: training and inference. Training consists of providing the machine with a large dataset and fine-tuning its algorithm so it can learn from the data.[7] During training, the model identifies similarities and correlations in data, allowing it to recognize patterns or anomalies when faced with new data.

During inference, a trained machine applies its acquired knowledge to make predictions or decisions on new, unseen data.[8] In this phase the model focuses on applying what it has learned and becomes capable of spotting trends and categorizing new data. This process enables the automation of tasks that would typically need human capabilities, although humans still play a crucial role in guiding the machine by selecting important features in the data, adjusting the algorithm, and providing labeled examples to help it learn accurately.

ML is typically divided into three main categories:[9]

Supervised learning: The model is trained on labeled data, meaning each input comes with the respective correct output, allowing the machine to learn by comparing predictions to those outputs and improving over time;

Unsupervised learning: The model works with unlabeled data with no correct answers provided, and searches for patterns within it to then generate clusters and groupings. This method helps uncover structure within the data;

2 Nuria Oliver, "Artificial Intelligence for Social Good," ELLIS Alicante Foundation, n.d., https://ellisalicante.org/artificial-intelligence-social-good.

3 Example: Early credit card fraud detection systems, where the system monitors transactions, including details like amount, location, and frequency. The system might flag and block the transaction based on the following example rules: a) IF "transaction amount > $10,000" AND "location = unusual" THEN "flag for review" b) IF "multiple transactions in a short time" AND "merchant type = high risk" THEN "temporarily block card" c) IF "transaction outside home country" AND "no travel notification" THEN "send alert to user".

4 "AI vs Traditional Programming - What's the Difference?" GeeksforGeeks, August 21, 2024, https://www.geeksforgeeks.org/what-is-artificial-intelligence-ai-and-how-does-it-differ-from-traditional-programming/

5 Reinaldo Padilha França et al., "An Overview of Deep Learning in Big Data, Image, and Signal Processing in the Modern Digital Age," Trends in Deep Learning Methodologies, 2021, 63–87

6 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law," Stanford Law School (2023). https://law.stanford.edu/publications/the-use-of-artificial-intelligence-in-international-human-rights-law/.

7 David Berrio, "Ai Inference VS Training vs Fine Tuning: What's the Difference?" HatchWorks, September 6, 2024, https://hatchworks.com/blog/gen-ai/ai-inference-training-and-fine-tuning/.

8 Ibid.

9 Reinaldo Padilha França et al., "An Overview of Deep Learning in Big Data, Image, and Signal Processing in the Modern Digital Age".

Reinforcement learning: The model learns by trial and error, taking actions to maximize rewards. It's often used in applications like games, robotics, and navigation.

ML represents a fundamental AI technique to gather data, analyze and learn from it, and predict certain outcomes. Advanced methods like deep learning, natural language processing, and generative AI build on this approach. While these methods vary in how they learn, they all rely on data-driven learning rather than fixed software rules like in expert systems.[10] This implies that the quality of data available to machines plays a crucial role in their learning process.

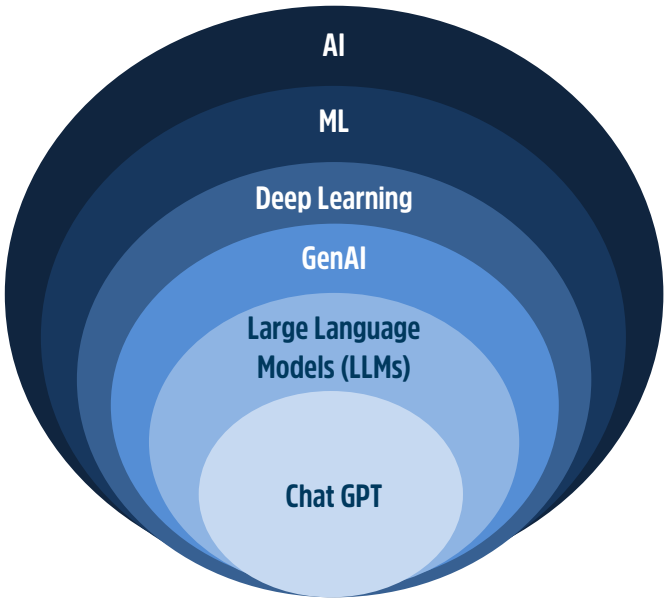

Diagram illustrating the connections between various AI systems

AI's main advantage is its ability to automate tasks and process large volumes of data much faster than humans.[11] Focusing on the key stages of data processing — data collection, publication, uptake, and impact — AI offers important contributions at each step:

Data Collection: AI can gather relevant information quickly and accurately in a scaled-up manner;[12]

Data Publication: AI facilitates real-time sharing of data, often using advanced visualizations to present complex information clearly;[13]

Data Uptake: Through automated analysis, AI generates insights that can guide decision-makers;

Impact Tracking: AI can monitor the effects of policies and decisions based on data analysis.[14]

By streamlining these processes, AI reduces the time needed for data analysis and minimizes human labor as well as error.[15] It provides scalability, by handling increasing data volumes without requiring additional resources.[16] The algorithms improve as they analyze more data, leading to better insights and more accurate predictions.[17] Moreover, AI can work with both structured data (easily organized and searchable) and unstructured data (more complex and harder to categorize).[18]

It is important to remember that AI systems are part of a larger ecosystem that includes human users. Therefore, their implementation should prioritize human needs and consider the significant effects they may have on people's lives.[19] Additionally, the

10 Ibid.

11 A Beduschi, "Human Rights and the Governance of Artificial Intelligence," Geneva Academy, March 2, 2020.

12 "The Data Value Chain: Moving from Production to Impact," Open Data Watch, n.d., https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact/.

13 Ibid.

14 Ibid.

15 The Upwork Team, "AI in Data Analysis: Basics, Examples, and Applied Uses,"

16 Domenico Zipoli, "The Human Rights Data Revolution", Geneva Academy of International Humanitarian Law and Human Rights (March 2024), https://geneva-academy.ch/research/publications/detail/763-briefing-n0-23-the-human-rights-data-revolution.

17 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law".

18 Ibid.

19 Mark O Riedl, "Human-Centered Artificial Intelligence and Machine Learning," ArXiv, Cornell University (2019).

outcomes produced by AI must be explainable and understandable, even for those without technical expertise.

In the following sections, we will explore the potential societal impacts of AI, including its risks and implications for sustainable development and human rights.

AI FOR SOCIAL GOOD

The purpose of AI is twofold: to employ machines for practical tasks, often using approaches that differ from how human minds operate, and, for scientific aims, to explore and address questions about humans through AI models.[20] These capabilities make AI a relevant tool for analyzing complex global conditions, identifying vulnerable regions and social groups in need of support, and forecasting trends within that context.[21] AI has the potential to contribute to addressing social challenges, and its tools and methods are being applied to various issues related to human and environmental well-being. The AI for Social Good (AI4SG) initiative, for example, aims to apply AI to achieve outcomes that support societal needs.[22]

AI can be utilized to promote sustainable development in an inclusive way, as emphasized by the AI4SG movement.[23] This initiative brings together stakeholders from diverse sectors to leverage AI for advancing the Sustainable Development Goals (SDGs), with a focus on maximizing AI's positive impact across these goals.[24] Numerous projects within the movement demonstrate that AI is being applied within every SDG, and experts suggest that AI can contribute to achieving at least 134 specific targets across the SDGs.[25]

Beyond supporting SDGs, AI plays a role in promoting human rights. For instance, AI improves equitable access to education by providing personalized learning experiences and helping students in remote or disadvantaged communities access high-quality educational resources.[26] In healthcare, it has proven to enhance the accurate and rapid analysis of clinical images and enable the early diagnosis of diseases such as breast cancer, which can facilitate prompt intervention and reduce costs associated with late-stage treatment.[27] In environmental protection, AI models monitor deforestation, track and reduce methane emissions, provide real-time air quality insights, and help shape policies to combat pollution, thereby contributing to the realization of the right to a healthy environment.[28]

A landmark UN General Assembly resolution promoting "safe, secure, and trustworthy" AI systems recognizes AI's potential to accelerate progress toward the SDGs and emphasizes that its use should be directed toward protecting and promoting human rights.[29] Various UN agencies are reflecting this commitment by using AI in their sustainable development projects; for example, the UN Refugee Agency (UNHCR) uses AI to monitor and track displaced populations, while the World Food Program (WFP) leverages machine learning tools for real-time hunger predictions across 94 countries.[30] To further enhance AI's impact in this sense, stronger collaboration among different stakeholders is needed, fostering partnerships

20 Margaret A Boden, Artificial Intelligence: A Very Short Introduction (Oxford University Press, 2018).

21 Nuria Oliver, "Artificial Intelligence for Social Good".

22 Dilek Fraisl et al., "AI through the Lens of Official Statistics and the Sustainable Development Goals: The Benefits and Risks"

23 Nenad Tomašev et al., "AI for Social Good: Unlocking the Opportunity for Positive Impact,".

24 Ibid.

25 Ricardo Vinuesa et al., "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals," Nature Communications 11, no. 1 (January 13, 2020).

26 UNESCO, "Global Education Monitoring Report 2023: Technology in Education: A Tool on Whose Terms," (July 26, 2023).

27 Margaret Chustecki, "Benefits and Risks of AI in Health Care: Narrative Review," Interactive Journal of Medical Research 13 (2024)2024.

28 UNEP, "How Artificial Intelligence Is Helping Tackle Environmental Challenges," UNEP (2022), https://www.unep.org/news-and-stories/story/how-artificial-intelligence-helping-tackle-environmental-challenges.

29 United Nations, "General Assembly Adopts Landmark Resolution on Artificial Intelligence | UN News," news.un.org, March 21, 2024, https://news.un.org/en/story/2024/03/1147831.

30 Dilek Fraisl et al., "AI through the Lens of Official Statistics and the Sustainable Development Goals: The Benefits and Risks".

and knowledge-sharing between field experts, policymakers, and AI researchers.

The following sections explore the use of AI for monitoring purposes. The first section focuses on how AI can enhance the monitoring of SDG progress by addressing data availability issues and improving all stages of data processing, thereby enabling stakeholders to engage in informed, timely policymaking. The second section discusses how AI contributes to human rights monitoring and highlights its key advantages in this context. The final section examines the potential risks associated with applying AI for human rights monitoring and emphasizes the importance of ensuring its use remains responsible and ethical.

ENHANCING SDG MONITORING THROUGH AI

The advancements AI has brought to data collection and analysis play an important role in tracking progress toward implementing the SDGs and achieving their targets. Effective implementation requires tools that enable policymakers to conduct reliable, objective situational analyses to inform policy decisions and evaluate their impact on progress toward SDG targets. The accuracy of these analyses and impact assessments depends heavily on the availability of high-quality data and the ability to process it effectively, making data collection and processing critical to achieving the SDGs.

In many countries, the availability of data is a significant barrier to effectively track progress in the implementation of SDGs. This challenge is particularly pronounced in developing nations, which often lack access to quality data in critical policy areas relevant to these goals, preventing decision-makers from obtaining the accurate information needed to inform their policymaking and advance the SDGs.[31] The indicators for monitoring progress in achieving SDGs frequently exceed these countries' technical and financial capacities. Experts indicate that 72% of nations require external assistance to measure these indicators and evaluate progress in related policy areas.[32] Additionally, these countries typically rely on traditional data sources, such as censuses and surveys, which are infrequently conducted and fail to provide the reliable, up-to-date data necessary for informed, efficient decision-making and implementation of SDGs.[33]

Open data platforms play an important role in addressing these data gaps. In this context, alongside databases maintained by governments, international organizations, NGOs, and scientific agencies, private corporations can also make significant contributions by offering open data services, helping to close data gaps across countries.[34] However, it is important to recognize the potential risks of overreliance on private entities, as it could enable them to dominate the data landscape and create unequal power dynamics, allowing corporate interests to influence public policymaking.[35]

Another approach to improving data availability and supporting SDG indicators is leveraging non-traditional data sources, such as Earth observation data, encompassing but not limited to satellite-based remote sensing and drone imagery, or citizen science initiatives.[36] When combined with these diverse data sources, AI tools and techniques can help bridge data gaps faced by different countries and address challenges related to data availability in monitoring the SDGs. For instance, in Ghana AI has utilized drone imagery and citizen science to identify hotspots of litter and plastic accumulation along the coastline, automatically analyzing data on the locations and types of plastics detected.[37] This process proved more efficient than traditional data-gathering methods, improving data

31 Mehrbakhsh Nilashi et al., "Critical Data Challenges in Measuring the Performance of Sustainable Development Goals: Solutions and the Role of Big Data Analytics," Harvard Data Science Review 5, no. 3 (July 27, 2023).

32 Ciarán O'Brien, "Big Data and A.I. for the SDGs: Private Corporation Involvement in SDG Data-driven Development, Policy and Decision-making," May 2022, https://sdgs.un.org/sites/default/files/2022-05/2.3.1-28-OBrien%20-big%20data%20and%20AI.pdf.

33 Dilek Fraisl et al., "AI through the Lens of Official Statistics and the Sustainable Development Goals: The Benefits and Risks".

34 Ciarán O'Brien, "Big Data and A.I. for the SDGs: Private Corporation Involvement in SDG Data-driven Development, Policy and Decision-making".

35 Ibid.

36 Steffen Fritz et al., "Citizen Science and the United Nations Sustainable Development Goals," Nature Sustainability 2, no. 10 (October 9, 2019): 922–30

37 Dilek Fraisl et al., Feasibility study on marine litterdetection and reporting in Ghana, UN SDSN Trends, 2023, https://www.sdsntrends.org/citizen-science-project?locale=en.

availability and granularity. It provided NGOs and other stakeholders with relevant data, enabling more targeted cleanup efforts and contributing to data collection for SDG reporting. Specifically, these AI-supported efforts addressed SDG target 14.1, focused on preventing and reducing marine pollution of all kinds, including plastic debris.[38]

In Colombia, AI and satellite imagery were integrated with traditional, census data to track poverty at a more detailed level and improve data availability on marginalized communities.[39] An AI algorithm analyzed daytime satellite images, generating more data points than traditional sources and providing a granular representation of poverty rates across the country. The accuracy of these predictions was validated using the latest census results, ensuring consistency and interoperability across different data sources. This method allowed the government to implement targeted policies and supported progress toward SDG targets 1.1, 1.2, and 10.2, which aim to eradicate extreme poverty, reduce poverty in all its dimensions for at least half of the affected population, and promote social, economic, and political inclusion.[40]

These examples demonstrate how AI technologies can improve data relevance and support targeted interventions aimed at achieving the SDGs. By collecting and processing data from different sources, AI can generate relevant information to populate SDG indicators and track progress. It also supports policymaking by providing timely insights and identifying trends, enabling prompt interventions to address SDG-related challenges and contributing to the achievement of these goals.

However, the use of AI for SDG monitoring is not without concerns. Its accuracy depends heavily on the data used to train its algorithms. For AI to perform effectively in a specific context, the algorithms should be trained on local data to account for local nuances and avoid biased outcomes.[41] Algorithms designed elsewhere may fail to capture local realities, leading to biased results. For instance, in the case of Ghana, the AI relied on an algorithm developed in a European country, which may have not fully reflected local and contextual conditions.[42] A broader challenge is unequal access to AI systems, especially in the Global South, where some countries face significant barriers due to limited access to basic infrastructure, such as internet connectivity and electricity.[43] Section iii. provides a more detailed discussion of the most relevant risks associated with using AI for monitoring purposes, focusing particularly on the human rights field.

AI FOR MONITORING HUMAN RIGHTS

In the context of using AI for social good, its advantages in data collection and analysis also extend to monitoring human rights protections and violations. AI can strengthen the international human rights reporting system by automating key tasks such as collecting and organizing data, analyzing existing datasets to identify new patterns in protections or violations, forecasting trends, and supporting decision-making.[44] It has predominantly been used by NGOs to monitor human rights trends, aiding their advocacy efforts before UN human rights bodies and the general public.

The use of AI for these purposes offers significant advantages, as it can contribute to various stages of human rights monitoring and reporting. AI can gather data on human rights implementation or violations from diverse open sources and generate insights concerning different issues and social groups.[45] This enables organizations and

38 "The 17 Goals | Sustainable Development," United Nations, https://sdgs.un.org/goals.

39 Dilek Fraisl et al., "AI through the Lens of Official Statistics and the Sustainable Development Goals: The Benefits and Risks".

40 "The 17 Goals | Sustainable Development," United Nations, https://sdgs.un.org/goals.

41 Dilek Fraisl et al., "AI through the Lens of Official Statistics and the Sustainable Development Goals: The Benefits and Risks".

42 Ibid.

43 Centre for Intellectual Property and Information Technology Law, The state of AI in Africa report 2023, 2023, https://cipit.strathmore.edu/wp-content/uploads/2023/05/The-State-of-AI-in-Africa-Report-2023-min.pdf.

44 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law".

45 Ibid.

authorities to have a more comprehensive and detailed understanding of human rights concerns and thus take informed and timely actions to address them.

The international human rights monitoring system relies on data from four primary sources: international organizations, national human rights institutions (NHRIs), NGOs, and states. Among these, states have a formal obligation to report on human rights within their jurisdictions.[46] AI can support state reporting by providing help in several ways. For example, AI can detect gaps in existing datasets, identifying missing or outdated information that may need updating.[47] It can also synthesize large volumes of information, presenting it in more accessible and visually comprehensible formats. Additionally, AI can support the drafting of reports and cross-reference their data with other sources to ensure their accuracy.[48]

Here are some of the key advantages of AI use in human rights monitoring:

Automated Data Processing: AI systems can process vast amounts of human rights data, such as reports, images, and social media content, at speeds that goes far beyond human capabilities. For instance, AI has been used to analyze satellite images to detect human rights violations like ethnic violence or forced displacement. In Darfur, satellite imagery has been used to track and report on village destruction, while in Myanmar, AI tools were employed to monitor and highlight attacks on the Rohingya people by analyzing thermal imaging data;[49]

NLP: These tools can identify patterns in text data, track trends in violations, and even predict potential abuses by analyzing government statements, NGO reports, and media articles. By using advanced sentiment analysis, these AI systems can determine not just what violations occurred but also the severity and intensity of these violations, offering deeper insights into the human rights landscape (see the example in the box);[50]

In the context of human rights reports, many NLP techniques rely on quantitative analysis, counting words that relate to key issues but lacking the ability to analyze sentence structure or uncover the meaning conveyed through word connections. PULSAR is a tool that addresses this gap by employing aspect-based sentiment analysis (ABSA), which goes beyond simple word counting. It uses grammatical and syntactic relationships between words to identify the opinions or judgments expressed in the text. Specifically designed for analyzing human rights reports, PULSAR automatically extracts the judgments conveyed within the text. (Park, Greene, Colaresi 2020)

Through automated processing of large volumes of natural language, PULSAR has been trained to identify the central issue or right being discussed, along with the sentiment attached to it, and to link these two elements. By extracting both the issue and the expressed judgment, it creates a new token that can then be quantified for further analysis. (Park, Greene, Colaresi 2020)

PULSAR is an open-source tool intended not to replace human analysis, but to enhance it by automating certain aspects and thus making it more efficient.

46 Ibid.

47 Michael L. Littman et al., "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report" (arXiv, 2022).

48 PwC, Artificial Intelligence for Reporting, 2020, https://www.pwc.com/sg/en/consulting/assets/artificial-intelligence-for-reporting.pdf.

49 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law"

50 Baekkwan Park, Kevin Greene, and Michael Colaresi, "How to Teach Machines to Read Human Rights Reports and Identify Judgments at Scale," Journal of Human Rights 19, no. 1 (2020): 99–116.

Predictive Analytics: AI has the potential to forecast future human rights violations.[51] By training models on historical data, AI can identify patterns that signal escalating risks, such as impending ethnic conflicts or refugee crises. This allows organizations to take preemptive action, potentially preventing further violations before they escalate. For example, the Danish Refugee Council (DRC) used AI and ML to predict displacement trends in 2021 and 2022, allowing for more effective humanitarian responses;[52]

Disaggregation of Data: AI enhances the ability to analyze human rights data at a granular level.[53] By disaggregating data, AI tools can reveal how specific communities or demographics, such as women, children, or ethnic minorities, are disproportionately affected by human rights abuses. This allows human rights organizations to advocate for targeted interventions, ensuring that the most vulnerable populations receive the needed attention;

Validation and Cross-Referencing: AI can improve the accuracy of human rights reporting by cross-referencing data from multiple sources, validating reports, and detecting inconsistencies.[54] This feature is particularly important in relation to state reporting on human rights implementation, since governments are prone to represent the state of things in a better light and thus should be cross-referenced with other reports such as those produces by NGOs.

RISKS OF AI IN HUMAN RIGHTS MONITORING

Using AI for monitoring purposes is not without risks, particularly in sensitive areas such as human rights. While AI technologies offer significant benefits to human rights monitoring, it is crucial to ensure their use remains fair, transparent, and responsible. This fairness is grounded in international human rights principles and protection standards, which serve as essential safeguards to regulate the development and application of AI tools.[55] In the absence of an AI governance framework, these technologies are at risk of perpetuating, or even worsening, human rights violations, especially affecting vulnerable social groups. Since AI algorithms learn from the data they are trained on, the challenges associated with using AI in human rights monitoring are closely tied to the nature, availability, accessibility, and quality of the data itself.[56]

The use of AI in this field must be carefully managed to avoid exacerbating existing inequalities and patterns of discrimination. Here are some of the crucial risks that must be taken into account:

Bias: One of the most significant risks associated with AI is the presence of bias. AI models are trained on data created by humans or human-made systems, meaning any existing human bias is likely to be transferred to the AI.[57] As a result, if algorithms are trained on biased data, their outputs will also reflect and perpetuate those biases and discriminate against minorities or societal, political, and demographic markers underrepresented in that data.[58] This poses a serious issue in human rights monitoring, where biased data can distort the representation of human rights conditions, leading to overlooked violations and reinforcing inequality.[59] AI systems trained on

51 Domenico Zipoli, "The Human Rights Data Revolution".

52 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law".

53 Ibid.

54 Ibid.

55 Nenad Tomašev et al., "AI for Social Good: Unlocking the Opportunity for Positive Impact,"

56 Medha Bankhwal et al., "AI for Social Good: Improving Lives and Protecting the Planet".

57 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law".

58 Domenico Zipoli, "The Human Rights Data Revolution".

59 Ronald Musizvingoza, "AI as a Catalyst for Sustainable Progress: Ensuring Information Integrity," United Nations University, June 2024, https://unu.edu/macau/blog-post/ai-catalyst-sustainable-progress-ensuring-information-integrity.

THE IMPORTANCE OF DATA GOVERNANCE FOR AI RELIABILITY

Data governance is key for ensuring the reliable use of AI in data processing and human rights monitoring. Indeed, AI systems depend on high-quality, reliable, and responsibly managed data throughout its lifecycle, i.e. from collection to deletion. Poor data governance can result in biased or inaccurate AI outcomes, particularly when the data is incomplete or not representative of different segments of society. (Verhulst and Schüür 2023)

Establishing data governance standards is crucial also for ensuring compliance with privacy and data protection laws, helping to reduce risks related to data privacy, security, and transparency. Indeed, only proper governance frameworks can entrench rules that require respect for individuals' consent when collecting, using, and processing their data. Without such standards, safeguarding privacy and ensuring ethical data practices would be difficult to achieve. (Verhulst and Schüür 2023)

The Global Digital Compact (GDC) is a UN initiative aimed at creating a unified framework to regulate digital technology and AI. It offers the opportunity to establish global standards for data governance, which is quite inconsistent across regions and sectors. A key goal of the GDC is to promote the development of strong, representative, and accessible datasets, to help prevent the harmful use of data by AI, thus protecting human rights and supporting sustainable development. (Slotin and Mclaren 2024)

data that predominantly reflect the experiences of certain ethnic or socioeconomic groups may fail to identify or address violations affecting other groups. This can reinforce and worsen the harm faced by those who are already more susceptible to human rights abuses: for instance, AI models may overlook violations against minorities if the data used to train them reflects predominantly the reality of dominant segments of society.[60] AI tools are often limited in their ability to recognize social and cultural nuances, which can lead to these important differences being overlooked during data processing;

Lack of transparency and accountability: Many AI systems, particularly those using deep learning, function as "black boxes," where the decision-making process is not obvious and is often difficult for the general public to understand.[61] In human rights monitoring, transparency is crucial, as the outcomes directly influence policy decisions that shape how human rights are regulated and enforced. When the results of AI data processing are not explainable, the reliability of human rights findings grounded on AI work is undermined and can be dismissed by the end-users.[62] Additionally, this lack of transparency impedes accountability from being established for errors or misjudgments in human rights monitoring, which could lead to further violations. Transparency involves making the processes behind AI outcomes accessible and understandable to end-users, including the public.[63] Trust in AI will only be established if people can comprehend how these systems get to their insights and conclusions, particularly when such outcomes influence decisions that affect their lives;

Surveillance and Privacy Concerns: The use of AI in human rights monitoring often involves collecting and analyzing sensitive personal data, which raises serious privacy concerns.[64] Given that AI systems are data-driven and handle vast datasets, there is a risk of these tools becoming intrusive and crossing into surveillance territory.[65] To prevent this, AI technologies must be designed to ensure data confidentiality and privacy, with strict limits on access to sensitive information. For example, researchers have developed Privacy Enhancing Technologies (PETs), which allow data

60 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law".

61 A. Beduschi, "Human Rights and the Governance of Artificial Intelligence".

62 Domenico Zipoli, "The Human Rights Data Revolution".

63 Mark O Riedl, "Human-Centered Artificial Intelligence and Machine Learning".

64 A. Beduschi, "Human Rights and the Governance of Artificial Intelligence".

65 Domenico Zipoli, "The Human Rights Data Revolution".

to be collected, processed, and shared without compromising privacy. These tools, along with encryption-based data processing methods, could help protect personal information while still enabling effective human rights monitoring;[66]

Data Quality and Availability: AI depends heavily on the availability and quality of data. In many regions, especially in developing countries or conflict zones, data may be incomplete, unreliable, or inconsistently available across different areas. This challenge is closely tied to resource constraints, as developing countries often lack the technologies and tools necessary to collect data effectively and reliably.[67] As a result, the absence of high-quality data for training AI systems can lead to inaccurate or misleading outcomes, creating gaps in information.[68] This can have serious repercussions for human rights monitoring and reporting, potentially distorting key issues or misrepresenting the situation on the ground;

Risk of Misuse by States and Malicious Actors: While AI can be a powerful tool for monitoring human rights violations, it can also be misused by states, particularly those that are the primary violators of human rights. When both the regulation and use of AI technologies are controlled by the state, there is a risk that governments may exploit them to evade accountability, manipulate data, or even perpetuate human rights abuses.[69] Such misuse of AI tools can produce unreliable and manipulated insights into the human rights situation and thus severely undermine public trust in the institutions that rely on these results for policymaking.[70]

Beyond these risks associated with AI use for human rights monitoring, there are also significant environmental and economic impacts that must be acknowledged. Most large AI systems are run in data centers, which have significant environmental impacts.[71] They generate hazardous electronic waste and consume vast amounts of water, while a quarter of the global population lacks access to clean water. Additionally, they require substantial energy, mostly from fossil fuels, contributing to greenhouse gas emissions; a single request to AI platforms like ChatGPT can use ten times the electricity of a Google search.[72] With the number of data centers increasing from 500,000 in 2012 to 8 million today, AI's growing demands pose serious environmental challenges. The environmental impact of AI must be critically assessed, specifically in the context of SDGs and human rights, as environmental degradation from climate change caused by the excessive use of the technology directly harms individuals' rights.

AI use produces high economic expenses tied mostly to its infrastructure. These costs can limit access to advanced AI capabilities, concentrating development and deployment within wealthier states and organizations.[73] Additionally, AI use over time can be expensive due to ongoing costs for energy, processing power, and skilled experts. As demand for AI grows, these costs can become a sustainability challenge, straining budgets and resources.

66 OECD, "Emerging Privacy-Enhancing Technologies," OECD Digital Economy Papers (March 8, 2023).

67 Dilek Fraisl et al., "AI through the Lens of Official Statistics and the Sustainable Development Goals: The Benefits and Risks".

68 Domenico Zipoli, "The Human Rights Data Revolution".

69 Anne Dulka, "The Use of Artificial Intelligence in International Human Rights Law".

70 Ronald Musizvingoza, "AI as a Catalyst for Sustainable Progress: Ensuring Information Integrity".

71 UN Environment Programme, "Environment Under Review," September 21, 2024.

72 Ibid.

73 Tania Babina et al., "Artificial Intelligence, Firm Growth, and Product Innovation," Journal of Financial Economics 151 (January 2024).

Another point of contention which needs to be mentioned specifically in the context of applying AI tools to human rights monitoring is the proper inclusion of values, norms, and ethics when pursuing AI alignment. The development of powerful AI models predominantly in countries of the Global North leads to a significant challenge in AI alignment and value integration. AI models often reflect WEIRD (Western, Educated, Industrialized, Rich, and Democratic) perspectives, disregarding global, regional and local values, norms, and traditions.[74] This creates a tension between the dominant AI paradigm and the need for more inclusive, diverse and pluralistic value sets integrated in AI systems. There is a growing emphasis on incorporating human rights values and local, contextual and regional perspectives into AI systems.[75] Recent changes in tech companies scaling back DEI (Diversity, Equity, Inclusion) programs reflect a shift in priorities that may influence efforts to embed values into AI systems other than those promoted by WEIRD perspectives.

Moreover, interdisciplinary and cross-sector collaboration is crucial in developing and deploying responsible AI systems that address critical societal concerns. A multi-stakeholder approach that brings together computer scientists, social scientists, ethicists, legal experts, human rights practitioners, domain experts, government representatives, and grassroots CSO activists guarantees the creation of AI solutions which align with diverse cultural, ethical, and social viewpoints. Incorporating a global framework of values, or contextual, regional, and local values where appropriate, into AI systems designed for monitoring the delicate and sensitive domain of human rights is essential. This approach ensures that these systems are perceived by the global community as trustworthy, responsible, credible, and fair, and thus fosters widespread acceptance and confidence in their use.

CONCLUSION

AI and ML have transformative potential for addressing global challenges, from achieving the SDGs to improving human rights monitoring and thus contributing to their advancement. By automating many processes that demand high levels of time and effort from human intelligence, AI enhances the ability of organizations and governments to make informed decisions and respond more quickly and efficiently to situations that challenge the achievement of these goals and human rights protection. In particular, in the field of human rights monitoring, the different features of AI can improve the timeliness and accuracy of many processes concerning the analysis of large-scale data for the purpose of establishing the level of human rights protection and identifying violations. However, careful governance and ethical considerations are necessary to mitigate the risks associated with AI, particularly concerning bias, transparency, and privacy. By addressing these challenges, AI can become a powerful force for social good, contributing to a more just, equitable, and sustainable world.

74 Aubra Anthony, Lakshmee Sharma, and Elina Noor, "Advancing a More Global Agenda for Trustworthy Artificial Intelligence - Carnegie Endowment for International Peace | Carnegie Endowment for International Peace," Carnegie Endowment for International Peace, April 30, 2024, https://carnegieendowment.org/research/2024/04/advancing-a-more-global-agenda-for-trustworthy-artificial-intelligence?lang=en.

75 Maria Paz Canales, Ian Barber, and Jacqueline Rowe, "What Would a Human Rights-Based Approach to AI Governance Look Like?," What would a human rights-based approach to AI governance look like? – Global Partners Digital, September 2023, https://www.gp-digital.org/what-would-a-human-rights-based-approach-to-ai-governance-look-like/.

THE GENEVA ACADEMY

The Geneva Academy provides post-graduate education, conducts academic legal research and policy studies, and organizes training courses and expert meetings. We concentrate on branches of international law that relate to situations of armed conflict, protracted violence, and human rights protection.

THE GENEVA ACADEMY HUMAN RIGHTS PLATFORM

The Geneva Human Rights Platform (GHRP) provides a dynamic forum in Geneva for all stakeholders in the field of human rights - experts, practitioners, diplomats and civil society - to discuss and debate topical issues and challenges. Relying on academic research and findings, it enables various actors to become better connected, break down silos and, ultimately, advance human rights.

DISCLAIMER

The Geneva Academy of International Humanitarian Law and Human Rights is an independent academic centre, the publications of which seek to provide insights, analysis and recommendations, based on open and primary sources, to policymakers, researchers, media, the private sector and the interested public. The designations and presentation of materials used, including their respective citations, do not imply the expression of any opinion on the part of the Geneva Academy concerning the legal status of any country, territory, or area or of its authorities, or concerning the delimitation of its boundaries. The views expressed in this publication represent those of the authors and not necessarily those of the Geneva Academy, its donors, parent institutions, governing board or those who have provided input or participated in peer review. The Geneva Academy welcomes the consideration of a wide range of perspectives in the pursuit of a well-informed debate on critical policies, issues and developments in international human rights and humanitarian law.

The Geneva Academy
of International Humanitarian Law
and Human Rights

Villa Moynier
Rue de Lausanne 120B
CP 1063 - 1211 Geneva 1 - Switzerland
Phone: +41 (22) 908 44 83
Email: info@geneva-academy.ch
www.geneva-academy.ch