AI & EQUALITY

Possibilities and Perils of Artificial Intelligence in Open-Source Human Rights Investigations

Vyoma Raman and Camille Chabot

December 2023

Abstract

This article provides an overview of the role of social media, artificial intelligence (AI), and large language models (LLMs) in open-source human rights investigations. Drawing on our experiences as lead student researchers at the Human Rights Center's Investigations Lab at UC Berkeley School of Law's, we reflect on the transformative potential of social media in uncovering and documenting human rights violations, particularly in contexts where traditional investigative methods face limitations. We highlight the importance of ethical considerations and describe the Berkeley Protocol on Digital Open-Source Investigations, a guiding framework for online research. We delve into the current applications of AI in open-source investigations, focusing on the role of algorithms in social media and how human rights researchers manipulate these algorithms to curate relevant media. We examine AI's contribution to data collection, noting its ability to accelerate the process of gathering large volumes of data but also highlighting challenges in verifying data relevance and accuracy. We discuss the use of AI in content verification for reverse image searching, deep fake detection, and language translation.

Furthermore, we explore the potential use of LLMs in open-source human rights investigations. LLMs offer opportunities for enhanced content discovery, verification, and research resiliency. They can assist in understanding user intent, analyzing diverse sources, and providing comprehensive results. LLMs also aid in content verification by rapidly analyzing and summarizing data, identifying patterns, and facilitating report generation. Moreover, LLMs have the potential to mitigate the emotional impact of research on investigators by identifying and warning about distressing content. However, we emphasize the need for careful consideration of the benefits, limitations, and ethical implications associated with using LLMs. Monitoring for biases, misinformation, and interpretability challenges is crucial. We underscore the importance of supplementing LLM outputs with diverse sources, addressing biases, and ensuring transparency in legal contexts. While LLMs offer promise in enhancing open-source human rights investigations, their use should be judicious, complemented by human expertise, and subjected to rigorous ethical scrutiny.

Introduction

2017 marked a watershed moment [1] in the application of open-source information (OSI) to human rights investigations: the International Criminal Court issued the first arrest warrant reliant on social media evidence, targeting Libyan military officer Mahmoud Mustafa Busayf Al-Werfalli. OSI2 encompasses information that one can readily acquire from the internet, available to any member of the public without the need for special legal status or unauthorized access. The warrant hinged on a series of videos obtained from social media, evidencing seven executions in Benghazi between June 2016 and July 2017, which implicated Al-Werfalli in war crimes. This case underscored the transformative potential of social media in the field of human rights investigations, a potential that is ripe for further enhancement with the emergence of artificial intelligence.

As lead student researchers at the Human Rights Center's (HRC) Investigations Lab at the UC Berkeley School of Law, we use OSI to investigate human rights violations in locales such as Iran, Western Sahara, Latin America, and the U.S. The HRC has played a pioneering role in harnessing OSI for human rights violation investigations. Our training in open-source investigation methods has fueled our involvement in several projects probing human rights violations in partnership with NGOs, international organizations, courts, and journalists to uncover and fact-check public information relevant to human rights crises.

Open Source Investigations

As open-source investigators, our everyday work involves probing social media to unearth and document human rights violations that traditional investigative methods, like interviews and on-the-ground inquiries, often fall short of addressing.

Social Media's Revolution for Human Rights Investigations

The expansive reach of social media has revolutionized investigative practices adopted by legal experts, journalists, and human rights activists. The proliferation of civilian-generated visual content enables the use of real-time information in reports and legal proceedings, narrowing the explosion of first-hand accounts and citizen-driven, unofficial narratives of human rights violations amplifies our capacity to form a comprehensive understanding of on-ground realities, thus laying the groundwork for accountability.

Though open-source investigations frequently complement traditional research methods, they truly shine in contexts where hostility on the ground stifles conventional media and in-person investigations. In such cases, they offer a unique window into otherwise unreachable incidents. To illustrate, our partnership with Amnesty International facilitated an open-source investigation into potential crimes against humanity in Iran amid ongoing protests triggered by the in-custody death of Mahsa Amini. The constraints of activist detention and censorship by the Islamic Republic rendered on-the-ground work unfeasible.

Lastly, the dissemination of incidents on globally accessible social media platforms has broadened public awareness of violations far beyond the reach of conventional reports. The real-time, multi-perspective documentation of human rights incidents through social media has spurred the rise of citizen science in the field of human rights, empowering ordinary citizens to document human rights violations worldwide using their connected devices. Entities like Bellingcat, a Netherlands-based investigative journalism organization, and Amnesty International's university network, the Digital Verification Corps, exemplify this pioneering approach of mobilizing newly-trained citizens to conduct human rights investigations using OSI.

Methods and Ethics in Open-Source Investigations

Open-source investigations stand at the precipice of continuous evolution, demanding investigators to persistently acclimate to new methods and technologies. Yet, the bedrock principles of these investigations have proven resistant to the tides of time.

One such cornerstone is the Berkeley Protocol on Digital Open-Source Investigations3, a comprehensive set of guidelines for professionally and ethically conducting online research into alleged human rights violations and international crimes. The Protocol provides guidance on methodologies for collecting, analyzing, and archiving digital information within the context of human rights investigations. It also focuses on safeguarding the physical, psychological, and digital well-being of online investigators and first responders, recognizing that their work may place them in potentially threatening situations.

Effective and ethical open-source investigations hinge on diligent preparation. Investigators commence their work by assessing potential risks and threats and examining the digital media, and platforms used on the ground, the key actors involved, and the vernacular surrounding specific violations. Particular consideration is also given to technology accessibility to discern any discrepancies between online information and on-the-ground realities. For instance, the well-documented gender digital divide might result in underrepresentation of violence against women on social media platforms as compared to violence against men. This preparatory phase enables researchers to pinpoint potential risks, biases, and gaps, and counter them using various mitigation strategies.

Open-source investigations typically begin with the discovery of content relevant to the investigation's objective. Investigators utilize a combination of keywords in pertinent languages to find written or visual content that addresses the central questions of the investigation across various online and social media platforms. Tools like TweetDeck, CrowdTangle, and Boolean searches are instrumental in sifting through multiple media platforms and accessing user-generated posts and documents.

The adept use of these tools can significantly contribute to confirming established media trends and unearthing content that has been under-reported. For example, during our investigation of Title 42 in partnership with Human Rights First, we aimed to unveil occurrences of lesser-reported violence perpetrated against asylum seekers affected by the emergency health law at the Mexican border. Our focus spanned a range of incidents, from sexual violence to attacks against LGBTQ+ individuals. The pervasive social taboo surrounding these forms of violence often inhibits victims from discussing their experiences with traditional media outlets. Therefore, open-source investigations provide a privileged avenue to bring these concealed incidents to light.

When gathering content, it is paramount to adopt practices that deliberately counteract confirmation bias. This is achieved by involving researchers from a range of backgrounds and expertise and by utilizing VPNs to diversify search results. As discovery primarily hinges on keywords, different investigators can yield disparate outcomes based on their keyword selection, their chosen platforms, and their device's location. Therefore, to amass comprehensive evidence reflecting the actual situation on the ground, it is vital to employ carefully selected research teams and connect to multiple servers using VPNs. Researchers also often employ sock puppet accounts or fictitious online identities to offset data-driven personalization and safeguard their online and physical safety.

Leveraging user-generated content as reliable evidence for advocacy, trials, or journalistic stories, requires verification for authenticity. It is not uncommon for users to re-use photos and videos from past incidents or present misleading contexts, making it crucial for online investigators to fact-check the location (geolocation) and time (chronolocation) of visual content.

Tools like InVid are employed to help researchers determine the first instance a video or image was posted, locate critical frames in each video, and enlarge parts of images for various verification tasks. Reverse image search engines like Yandex and Google, and 3D map software like Google Earth Pro and PeakVisor, assist in finding correlations between visual content and identifiable locations, and in pinpointing precise coordinates. Investigators might also need to discern the precise moment an incident occurred; apps like SunCalc are used to estimate the day and time based on the sun's position. Lastly, in scenarios where content is at risk of being removed, such as graphic content taken down from social media platforms, investigators archive and protect it from destruction using tools like Hunchly.

During the investigation process, online investigators may encounter distressing material and vivid descriptions of human rights violations, which can lead to secondary trauma and PTSD. The Berkeley Protocol provides a set of guidelines to mitigate these risks. The Protocol recommends that researchers should be aware of their own and their colleagues' typical behavior, noticing any changes in eating, sleeping, and recreation habits. They can also employ various techniques to minimize exposure to harmful content, such as turning off audio, minimizing the screen, hiding violent material, using grayscale mode, working in pairs, and avoiding late-night work. Furthermore, researchers strive to foster a sense of community and camaraderie, which is crucial for maintaining good mental health in online investigations4.

Current Uses of AI in Open-Source Investigations

The essence of artificial intelligence is deeply intertwined with the current methodologies adopted for discovery and verification in open-source investigations. The tools that we deploy in our research frequently lean on algorithms to sift through, amass, and analyze pertinent open-source content.

Algorithms in Social Media Feeds

Algorithms serve as the unseen puppet masters of the social media experience, subtly molding the data landscapes navigated by users on a daily basis. These algorithms leverage vast data repositories of past interactions and user similarities to tailor content aimed at captivating and retaining user attention. News feed algorithms, which are AI systems determining the most engaging and pertinent content to exhibit in a user's news feed or timeline, embody this concept of data-driven personalization. They significantly dictate the flow and nature of information range of factors, such as the posts users engage with, the frequency of interactions with certain users, the time spent on various types of posts, and even the speed of scrolling. Consequently, these algorithms curate a personalized feed comprising videos, images, and other posts, fine-tuned to a user's preferences and geared towards prolonging their time spent on the platform.

Human rights researchers have discovered ways to manipulate these algorithms for purposes that transcend simple content consumption. For instance, a frequently employed technique to curate relevant media involves creating artificial social media profiles, or "sock puppets," which strategically interact with specific content. These profiles typically don't reflect the personal interests of their creators; rather, creators systematically engage with a distinct research topic of interest to unearth relevant content and analyze the surfaced information. By interacting with posts, following pages, and clicking on content that aligns with a specific theme, researchers can manipulate the news feed algorithm to present information that corresponds with the sock puppet's designed interest.

We have integrated this methodology into our human rights research, delving into topics as diverse as student protests in Iran and police brutality in Western Sahara. For example, while investigating an incident of excessive force in Smara, we employed a sock puppet with benign interests in sports to discreetly and anonymously probe evidence of Sahrawi activism and authorities' reaction to it. By selectively viewing videos and interacting with posts from Moroccan authorities and Sahrawi activists, we began noticing similar content appearing in our feed. Thus, we were able to manipulate the news feed algorithm to curate content relevant to our investigation.

Despite their utility, news feeds can inadvertently propagate detrimental narratives. Designed to favor engaging content, these algorithms might unwittingly amplify misinformation or propaganda that elicits strong emotional reactions and, in turn, user engagement. This could potentially lead to the creation of echo chambers, where users are consistently exposed to content that reinforces their existing beliefs, thus obstructing the dissemination of accurate information.

Data Collection Using AI

The recent advent of artificial intelligence has tremendously accelerated the process of data scraping from myriad sources, far surpassing the pace achievable by researchers manually conducting discovery. AI-driven data collection operates on the same fundamental principles as traditional discovery, primarily the employment of keywords and specific time frames to amass pertinent content. However, its transformative edge lies in its ability to gather substantially larger volumes of data within a significantly reduced timeframe. For instance, Amnesty International's

Digital Verification Corps leaned on an AI data scraping tool in their investigation into police brutality in Iran following the custodial death of Mahsa Amini. The combination of an internet shutdown and the sporadic nature of data availability posed formidable challenges to procuring evidence of crimes against humanity. The AI data scraping tool, configured with keywords provided by Amnesty International's Iran researchers, managed to collect and systematically arrange data on over three thousand incidents.

However, the incorporation of AI in data collection has brought forth its own set of unique challenges that researchers must grapple with. In this investigation, we found ourselves burdened with verifying the relevance of the massive volume of data collected. This was due to the scraping tool's occasional misattribution of media to inaccurate locations or time frames or the inclusion of content irrelevant to the investigation. Additionally, the imperative to ensure a diversity of perspectives in the discovery process equally applies to the development of the data scraping tool. Therefore, special care must be taken with respect to who designs the tool and the keywords they deploy, to avoid skewing outcomes and risking partiality.

AI Content Verification Tools

In addition to gathering content, AI systems such as Yandex and Google Reverse Image Search have emerged as resourceful allies for verifying visual content. These tools use computer vision algorithms to capture the different features of a given image—its colors, shapes, textures, and more—to create a unique fingerprint. The AI compares this fingerprint to those within its extensive database, seeking similar patterns. The harmony between AI and computer vision technologies thus enables these tools to swiftly pinpoint duplicates, altered images, or potential sources, bolstering the efficacy of content verification.

Verifying content also involves confirming that it was not artificially generated as a fake or deep fake. InVid has integrated multiple AI filters into its forensic capabilities to detect additions and modifications in images. These employ machine learning algorithms that sharpen their understanding of vast datasets of genuine and manipulated images. Instead of merely analyzing an image, the AI scrutinizes its intricate aspects, such as statistical patterns and visual attributes, looking for signs of tampering. By comparing the altered image with a reference or original image, the system can detect discrepancies in pixel-level details, inconsistencies in lighting and shadows, anomalies in noise patterns, or artifacts introduced during the editing process. These filters act as red flags, highlighting areas of the image that are likely to have been modified, thereby assisting investigators in identifying potential manipulations and deep fakes.

AI plays a crucial role in investigations conducted in foreign languages by enabling the translation of words from both text and images. Popular tools like Yandex and Google Translate are widely used for this purpose, proving particularly valuable when dealing with ideographic languages, where researchers cannot simply retype the text unless they have familiarity with the language. However, it’s important to acknowledge that these translation tools are not exempt from gender bias. Gender bias in translation arises when AI algorithms or the datasets used to train translation models exhibit biases in how they handle gender-specific language or cultural nuances linked to gender. For instance, certain languages may have grammatical rules or sentence structures that convey gender information, and if the translation models are not appropriately trained or calibrated, they might inadvertently reinforce or perpetuate gender stereotypes. To ensure fair and accurate translations for all users, it is essential to address and mitigate gender bias through continuous research, data curation, and algorithmic improvements.

Leveraging Large Language Models (LLMs) for Human Rights Investigations

As we explore the potential application of large language models (LLMs) like GPT-4 in open-source human rights investigations, we enter a realm of cautious optimism. These AI systems have the potential to be intricate data processing tools due to their capabilities of context recognition, semantic interpretation, and pattern detection. The prospect of transforming open-source investigation is tempting, but it is vital to navigating this path with measured steps, acknowledging the possible benefits, inevitable limitations, and the need for careful supervision.

LLMs in Content Discovery

The initial stage of any open-source human rights investigation is akin to navigating a labyrinth of online information. Traditionally, investigators rely on their expertise and intuition, employing search engines, social media platforms, and similar resources as described earlier. Investigators use keyword searches, explore hashtags, follow leads from related articles or posts, and track certain individuals, organizations, or locations over time. While these methods can yield useful information, they can be time-consuming in an urgent situation and still leave investigators open to the risk of missing crucial details.

In this context, LLMs propose a different approach. They don’t merely align with keywords or phrases like traditional search algorithms. Instead, they attempt to decode the underlying intent behind a user’s query. This includes understanding the context, recognizing sentiment and connotation, and, consequently, producing results that may be more accurate and comprehensive.

We contributed to OSI discovery and verification efforts of police brutality in Chile in October 2019. While we employed a variety of boolean search techniques to produce results, an LLM could augment this process. By recognizing the deeper context of the query, could assemble a broader spectrum of related information. This might include civilian documentation of police violence, related protests, official responses, legal proceedings, and public reactions. Beyond understanding user intent, LLMs are also adept at comprehending and synthesizing information from diverse sources of online media. They can analyze text, identify key themes and entities, extract relevant facts, and even summarize lengthy documents. This ability to process and make sense of large amounts of information can be a game-changer for investigators, helping them sift through the online information deluge more effectively. However, their efficiency doesn't replace the discerning eye of an investigator but supplements it by helping them navigate the digital deluge more effectively.

By combining their understanding of user intent and online media, LLMs can serve as supplementary tools for investigators to discover more relevant content on the Internet. They can extract first-hand social media accounts of an incident, find a particular type of media (e.g. videos) about it, and aggregate and summarize relevant press releases. In addition, they could suggest related topics or entities to explore, highlight emerging trends or patterns, or even point out inconsistencies or gaps in the available information that warrant further investigation.

For instance, in our investigation of murders of Indigenous environmental defenders in the Amazon basin, an LLM could have hastened our identification of related issues that have caused the violence, such as the illegal lumber trade, agriculture lobbying, and demand for transition minerals used for renewable energy. It could also surface content from lesser-known or non-English sources, thereby providing a more diverse and comprehensive view of the situation.

Despite these possible advantages, the use of LLMs demands continual monitoring and recalibration. The issue-laden release5 of Microsoft's Bing AI serves as a reminder of the potential unpredictability of LLMs. The risk is real: these models can reflect and even amplify biases, hostility, and misinformation prevalent in their training data, which often includes a representative, yet flawed, slice of the internet. In the face of these risks, we must approach

LLMs not as a panacea, but as a potentially valuable tool that must be wielded with care and vigilance.

LLMs in Content Verification

The journey through an open-source human rights investigation, while undeniably important, is often intricate and laborious. A critical juncture of this journey is the verification stage, where investigators sift through the data they discovered, analyze it, and craft comprehensive reports to corroborate incidents. In this process, human researchers carry the bulk of the load, meticulously picking through the information piece by piece – a task whose time-consuming nature often conflicts with its need to respond to rapidly evolving situations.

Here, LLMs could potentially serve as valuable adjuncts. Their ability to parse through colossal amounts of data swiftly suggests a more efficient path to unearthing the facts surrounding a human rights violation. Imagine an LLM rapidly condensing a spreadsheet populated with thousands of entries into a concise and comprehensible summary, outlining vital incident parameters such as primary locations, types of violations, prevalent keywords, incident dates, and sources. This goes beyond merely repackaging raw data into a user-friendly format; it allows investigators to glean hidden patterns and trends, thereby enabling the insightful navigation of an investigation. The recognition of a sudden increase in specific types of violations, or the emergence of new hotspots, can shape the direction of an investigation. The capacity for LLMs to digest content and produce coherent written summaries make them more powerful for this work than traditional data analysis.

LLMs' language versatility is also a significant boon. Human rights incidents span the globe, and LLMs' capability to automatically translate content can potentially lessen the dependency on human translators, lending efficiency and scalability to the process.

An integral part of open-source investigations is the production of detailed verification reports. These meticulous documents, crucial for any ensuing legal proceedings, outline the process by which each piece of evidence was found and fact-checked. With machine-assisted workflows, investigators could potentially find original online sources for discovered content more easily, interpret multilingual content, identify video locations by providing written descriptions of specific landmarks, and integrate their findings into a coherent, standardized report. This could streamline the report generation process while maintaining consistency across different cases.

Guided by clear investigative objectives, LLMs can possibly optimize the verification process by identifying and prioritizing content that aligns with the investigation's goals. Their understanding of content, context, and intent, allows them to discern and rank different scenarios across diverse forms of content. For instance, in video content, an LLM could distinguish a peaceful protest from a violent altercation based on post descriptions and comments. This discerning capability becomes paramount in human rights investigations where the evidence of violence and abuses are typically the central focus.

Let's consider an LLM trained on post descriptions of varying crowd situations and their comments. It learns to recognize textual patterns and signals associated with violent incidents, such as specific actions, language, or emotional tone. When presented with new posts, the LLM could analyze associated textual metadata, comments, or transcriptions, and rank the media by the likelihood of containing relevant content. This could allow investigators to focus their efforts on the most promising leads, potentially improving the efficiency of the investigative process.

Despite these potential advantages, we must remember that LLMs are not infallible. They are tools that could amplify our efforts, but they also require vigilant oversight and rigorous testing. As we walk the fine line between potential benefits and pitfalls, we must strive to leverage these technologies judiciously and ethically in our mission to uphold human rights.

LLMs in Research Resiliency

The pursuit of truth in open-source human rights investigations, while undeniably vital, carries a heavy emotional burden. Investigators are routinely exposed to graphic and distressing content, whether it is a video capturing brutal violence or a chilling first-hand account. This constant exposure can lead to vicarious trauma, a condition akin to post-traumatic stress disorder affecting individuals who regularly witness the traumatic experiences of others. Compounded by the ceaseless nature of this work, fueled by a profound commitment to human rights, the risk of burnout becomes a grim reality. Here, investigators may feel trapped, unable to step away without the gnawing fear of compromising their mission.

In this context, LLMs may present a potential shield by helping to mitigate the impact of traumatic content on investigators. By analyzing and sifting through vast amounts of data, these AI systems can identify content relevant to the investigation, potentially reducing investigators' exposure to irrelevant graphic material. Additionally, they can be trained to identify potentially distressing content and provide advance warnings, allowing researchers to prepare themselves before engaging with such material.

Consider an LLM trained to discern textual and visual cues indicative of violent or traumatic incidents, such as descriptions of violent acts, blood, or explosions. It could be programmed to flag descriptions of violent acts or visual depictions of blood or explosions, subsequently alerting investigators about the nature of the content they are about to encounter. This preemptive caution creates a protective buffer, affording investigators a chance to either mentally brace themselves or delegate the task if they feel ill-equipped to handle it at that moment.

We have previously worked on developing a video-viewing platform with capabilities like facial blurring, object tracking, audio analysis, and grayscaling. An LLM could be integrated into this to display content warnings and offer suggestions for less distressing ways of engaging with content. If an LLM identifies the presence of graphic elements like blood in a video, it might recommend viewing the content in grayscale to lessen the graphic impact. Similarly, it could propose muting the audio if it detects potential auditory triggers such as explosions or gunshots. We have worked on developing a video-viewing platform with these capabilities.

While these adjustments might seem small, they can significantly contribute to preserving the emotional well-being of investigators, thereby supporting the sustainability of their crucial work. By providing such protective mechanisms against the emotional toll of human rights investigations, LLMs could potentially bolster investigator resilience and strengthen the overall capacity of organizations conducting these pivotal inquiries.

However, as we contemplate these potential benefits, we must also maintain a measured perspective. LLMs are not a cure-all solution; they should be used in tandem with robust mental health support structures, including access to counseling and strategies for self-care and stress management. While LLMs might aid in mitigating trauma exposure, the true backbone of emotional well-being and long-term sustainability for investigators remains rooted in the human-centric support systems within an organization.

Ethics and Implications of Using LLMs in Human Rights Investigations

As we explore the potential of LLMs in open-source human rights investigations, we must also navigate the ethical implications inherent in their application. With the potential for transformative progress comes the reality of novel challenges; using LLMs within the field of human rights investigations is not without its potential pitfalls and ethical quandaries.

It is crucial not to overlook the risks that could arise from an overreliance of researchers on LLMs. As powerful as these tools can be, they are not infallible and should not be treated as an absolute source of truth. LLMs, at their core, are models trained on vast but nonetheless finite data sets. While this data provides the models with a broad basis for understanding and generating language, they are not without their inherent limitations. Investigators must bear in mind the fallibility of LLMs and consider their outputs as starting points for investigation, rather than definitive conclusions.

Misinformation is a significant concern when dealing with LLMs. Despite their impressive ability to produce contextually accurate information, LLMs generate responses based on patterns found in their training data. They do not have access to real-time or situation-specific information beyond what they were trained on. Thus, an LLM may create an output that appears authentic and credible but might not be accurate or relevant in the current context. For instance, during an investigation, an LLM might reference sources or media from similar scenarios in its training data, which are not relevant to the unique situation under investigation.

Another critical issue with LLM usage is the risk of amplifying bias. LLMs, despite their expansive scope, can inadvertently perpetuate existing biases in the data they were trained on. If their training data lack representation from certain demographics or omit information about specific issues, these shortcomings will likely be reflected in the LLM's output. This can result in a skewed representation of reality, potentially overlooking marginalized communities or underrepresented issues. An LLM, for example, may disproportionately surface content about dominant groups relevant to a particular conflict while failing to look for underreported human rights abuses in marginalized communities. Hence, it becomes imperative for human rights investigators to consciously incorporate diverse sources and viewpoints alongside LLM outputs to ensure a holistic understanding of the situation.

Interpretability is another significant consideration when using LLMs in a legal context. Human rights investigations often demand a high level of transparency and accountability. Each step of the investigative process must be justified and capable of being explained. However, LLMs, often lack interpretability and are considered a "black box."6 Understanding the decision-making process of an LLM—how it decided to conduct specific queries or rank the relevance of content—can be an incredibly challenging task.

The lack of interpretability becomes particularly significant when the research findings contribute to legal proceedings. If human rights investigators heavily rely on LLM-generated outputs, it may become challenging to justify these findings or validate their reliability in court. After all, a court might find it hard to accept evidence whose derivation cannot be entirely explained or justified. Thus, it becomes crucial for investigators to consider how LLM outputs will be used downstream and to design investigation workflows accordingly.

Conclusion

The integration of social media and AI has revolutionized open-source human rights investigations, expanding our capabilities in discovering, verifying, and collecting content to uphold human rights and promote accountability worldwide. Within this context, large language models (LLMs) hold immense promise and complexity. These AI systems possess sophisticated data processing abilities that with the potential to transform the way we uncover information, validate evidence, and protect researchers from vicarious trauma. LLMs empower investigators to navigate the vast expanse of online information, pinpoint relevant content, and extract invaluable insights. They offer the potential to streamline the verification process, automate data analysis, and shield investigators from emotional strain. However, it is crucial to recognize that LLMs are not flawless and can inadvertently perpetuate biases, amplify misinformation, and lack transparency in their decision-making. To harness the true potential of LLMs, human rights organizations must uphold ethical considerations, complement LLM outputs with diverse sources and human expertise, and ensure that their application adheres to principles of transparency, accountability, and fairness. By proceeding on this path with careful oversight and judiciousness, LLMs can play a meaningful role in advancing human rights and instigating positive change.

While LLMs offer great potential for expanding the investigative capacities of human rights researchers and activists, they also present new challenges that must be addressed by professionals in the field. The recruitment, training, and monitoring of human rights investigators need to consider the implications of AI, including dangers such as overreliance on AI tools,

References

[1] Irving, Emma. "And so It Begins… Social Media Evidence in an ICC Arrest Warrant." Opinio Juris, Sept. 2018, opiniojuris.org/2017/08/17/and-so-it-begins-social-media-evidence-in-an-icc-arrest-warrant

[2]OHCHR. "Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source and Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law." OHCHR, www.ohchr.org/en/publications/policy-and-methodological-publications/berkeley-protocol-digital-open-source.

[3] OHCHR. "Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source and Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law." OHCHR, www.ohchr.org/en/publications/policy-and-methodological-publications/berkeley-protocol-digital-open-source.

[4] 4 OHCHR. "Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source and Information in Investigating Violation of International Criminal, Human Rights and Humanitarian Law." OHCHR. 47-48. www.ohchr.org/en/publications/policy-and-methodological-publications/berkeley-protocol-digital-open-source

[5] Roose, Kevin. "Why a Conversation With Bing's Chatbot Left Me Deeply Unsettled." The New York Times, 17 Feb. 2023, www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

[6] Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A Survey of Methods for Explaining Black Box Models." ACM Computing Surveys 51, no. 5 (2018): 1–42. https://doi.org/10.1145/3236009.

Biographies

Vyoma Raman is an incoming master's student in the Department of Computer Science at Stanford University and a recent graduate of UC Berkeley with bachelor's degrees in Computer Science and Interdisciplinary Studies. Her interests lie at the intersection of artificial intelligence and human rights, with a particular focus on algorithmic justice and disability studies. Currently, Vyoma is an affiliate researcher at Berkeley Artificial Intelligence Research Lab and the UC Berkeley Human Rights Center. Previously, she has interned on multiple responsible AI product teams at Microsoft and conducted research on bias in synthetic media with Berkeley's Natural Language Processing Group and School of Information.

Camille Chabot is a recent graduate of UC Berkeley, holding bachelor's degrees in Global Studies, Human Rights, and Chinese, as well as a BA in Politics, Government & Law from Sciences Po Paris. With expertise in open source investigations from UC Berkeley Human Rights Center's Investigations Lab, she now works as a research consultant, exploring the impact of Large Language Models (LLMs) like ChatGPT on various professional fields. Camille's focus is on utilizing international law to bridge the understanding gaps between Eastern and Western perspectives on human rights and uphold a minimum threshold for human dignity. Additionally, she is an incoming master's student at the Yenching Academy of Peking University, where she will further expand her knowledge of Chinese culture and international law.