doi:10.1017/aju.2024.5

SYMPOSIUM ON DIGITAL EVIDENCE

CHALLENGES OF USING DIGITAL EVIDENCE FOR WAR CRIMES PROSECUTIONS: AVAILABILITY, RELIABILITY, ADMISSIBILITY

*Jessica Peake**

Digital evidence has the potential to transform the accountability landscape. However, several obstacles must be overcome to use it effectively for war crimes prosecutions. The availability of digital evidence can be impacted by a range of factors, from Internet connectivity to removal of content by platforms to the difficulty of identifying relevant content through the fast pace of social media. The reliability of digital information is increasingly being called into question because of deliberate disinformation campaigns by parties to conflicts, as well as the fear of media manipulated by Artificial Intelligence (AI). And questions around chain of custody and admissibility of digital evidence have not fully been resolved by international courts. This essay unpacks some of these challenges. It suggests some ways in which governments and big tech should seek to ensure access to digital spaces and put in place measures to increase the integrity of online content, and how investigators and lawyers can gather, authenticate, archive, and establish chain of custody to ensure it can be used in accountability processes.

If we can harness its transformative power, digital evidence can be highly probative in establishing the necessary *mens rea* and *actus reus* of international crimes. Digital information and digital evidence provide vital ways to monitor, track, and seek accountability for international law violations committed during armed conflict. Digital technologies give us access to otherwise difficult to reach places such as conflict zones: satellite imagery and geospatial analysis allow us to monitor troop movements in almost real time; those impacted by the conflict—both civilians and combatants—can upload photo and video content to their social media platforms giving immediate insights into what is happening on the ground; and we can use digital methods to analyze weapons trajectories and destruction of civilian infrastructure.

The Availability of Digital Evidence

The power of social media to enable people to organize and share information during conflict can incentivize repressive governments to restrict—or shut down—access to the Internet to control what people are saying and sharing online. This can be implemented both by a country's own government or by an occupying power. For example, during the violent conflict in the Tigray region of Ethiopia between November 2020 and 2022,[1] the region was subject to an Internet shutdown impacting more than six million people and restricting information emerging from the conflict.[2] This two-year information black hole prevented impacted communities from

* *Director, International & Comparative Law Program; Assistant Director, the Promise Institute for Human Rights, UCLA School of Law, United States.*

[1] African Union, Agreement for Lasting Peace Through a Permanent Cessation of Hostilities Between the Government of the Federal Democracy of the Republic of Ethiopia and the Tigray People's Liberation Front.

[2] Access Now, *Two Years of Internet Shutdowns: People in Tigray, Ethiopia, Deserve Better* (Nov. 4, 2022).

organizing and sharing stories of the harms committed against them, and will impact how this conflict is memorialized and the kinds of accountability projects that are pursued in the future.

Following Russia's unlawful invasion of Ukraine in early 2022, the city of Mariupol was subjected to a two-month siege and information blackout,[3] deliberately caused by Russian forces shelling the last cell tower.[4] Approximately 100,000 people remained in the city, with hundreds of thousands of others having already fled. As the only international journalists left in the city, reporters for the Associated Press continued to document and report by smuggling out memory cards of data from the region, at great personal risk,[5] without which we would not know the extent of the atrocities by Russian forces. Since the October 7, 2023 unlawful attack by Hamas on Israel, Israel has instituted a "complete siege" of Gaza.[6] This has included Internet shutdowns and information blackouts for 2.2 million people,[7] driven by the destruction of communications infrastructure[8] and shortages of fuel and power,[9] which has impeded the ability of the people of Gaza to communicate with one another and to document and share what is happening to them.[10]

Whether digital information can be used to monitor and seek accountability for laws of war violations depends on the availability of information. 2022 saw 187 Internet shutdowns in thirty-five countries, marked by "the weaponization of shutdowns during armed conflict."[11] Information blackouts during conflict can "provide cover for atrocities and breed impunity,"[12] as the world is unable to see the extent of the atrocities being committed on the ground in real time. Power blackouts mean that devices go uncharged and valuable digital evidence for future accountability processes is not captured.

Another way in which access to information uploaded by users in conflict zones is impacted is through social media platforms' content moderation policies, which can result in the removal, flagging, or deprioritization of content.[13] For example, two weeks into the Russia-Ukraine war, YouTube promoted that it had removed more than 15,000 videos for violating their terms of service.[14] As Human Rights Watch has reported, this "prevents the use of that content to investigate people suspected of involvement in serious crimes, including war crimes."[15] While there may be a legitimate public interest in removing content that could be harmful to users, it is imperative that platforms make the content available to researchers and investigators that need it to identify instances of international law violations during conflict. Additionally, social media platforms can use their algorithms to deprioritize content and to silence the voices of some parties to conflict. At the time of writing, the voices of Palestinians and supporters are being silenced through "content removals, suspension or deletion of accounts, inability to engage

[3] *Mariupol: Key Moments in the Siege of the City*, BBC (May 17, 2022).

[4] Access Now, *#KeepIntOn: How to Stop Internet Shutdowns in Ukraine* (Mar. 17, 2022).

[5] Mstyslav Chernov, *20 Days in Mariupol: The Team that Documented the City's Agony*, AP (Mar. 21, 2022).

[6] Isabel Kershner, Aaron Boxerman & Hiba Yazbek, *Israel Orders "Complete Siege" of Gaza and Hamas Threatens to Kill Hostages*, N.Y. Times (Oct. 9, 2023).

[7] Access Now, *Palestine Unplugged: How Israel Disrupts Gaza's Internet* (Nov. 10, 2023).

[8] *Gaza War Inflicts Catastrophic Damage on Infrastructure and Economy*, Reuters (Nov. 17, 2023).

[9] Hiba Yazbek, *Gaza Is Plunged into a Communications Blackout Amid a Severe Fuel Shortage*, N.Y. Times (Nov. 16, 2023).

[10] Human Rights Watch, *Gaza: Communications Blackout Imminent Due to Fuel Shortage, Israel Should End Blockaded, Restore Services* (Nov. 15, 2023).

[11] Access Now, *Weapons of Control, Shields of Impunity* (2020).

[12] Human Rights Watch, *supra* note 10.

[13] Mukund Rathi, *Amidst Invasion of Ukraine, Platforms Continue to Erase Critical War Crimes Documentation*, Electronic Frontier Foundation (Apr. 27, 2022).

[14] YouTubeInsider, @YouTubeInsider, X (Mar. 11, 2022, 12:26 p.m.).

[15] Human Rights Watch, *"Video Unavailable," Social Media Platforms Remove Evidence of War Crimes* (Sept. 10, 2023).

https://doi.org/10.1017/aju.2024.5 Published online by Cambridge University Press

with content, inability to follow or tag accounts, restrictions on the use of features such as Instagram/Facebook Live, and 'shadow banning' [making a user's content no longer visible to others]."[16] Through these measures, platforms are making it harder for people to tell their stories and creating challenges for war crimes investigators to identify potential violations.

When digital evidence is available, it can be difficult to find useful data through the noise of social media. For example, in 2022, 6,000 tweets were posted to Twitter every second[17] and more than 500 hours of video are currently uploaded to YouTube every minute. With that level of usage, discovering relevant content that shows potential violations of international law, or helps to identify alleged victims or perpetrators, is incredibly difficult. Machine learning has huge potential to help with the process of identifying and classifying relevant content and, as the technology improves, will greatly assist war crimes investigators in their monumental challenge of evidence collection and analysis. But machine learning must be implemented cautiously, so as to avoid the perpetuation of bias and discrimination inherent in many technology products.[18]

The Reliability of Digital Evidence

Even where digital information is available, rampant disinformation and deepfakes can call the reliability of content into question. It is well documented that the Internet is replete with disinformation, particularly in the context of elections[19] and COVID-19.[20] Disinformation during warfare is nothing new, but the speed with which disinformation can spread using social media is unprecedented and may affect the reliability of digital evidence, or at least the perception of its reliability. Two days after Hamas's attack on Israel, David Gilbert at Wired reported that the Israel-Hamas war "is drowning X in disinformation."[21] Some of that was driven by recent changes made to the X (formerly Twitter) platform. Some was driven by users posting content from other contexts: in some instances video was recycled from other conflicts, including Syria in 2020,[22] in others scenes from a video game were represented as Hamas attacks on Israel.[23] Additional unverified claims have circulated online, including much speculation about who is responsible for the hospital blast at Al-Ahli Baptist Hospital.[24] In the Russia-Ukraine conflict, Russia has deliberately engaged "systemic information manipulation and disinformation by the Kremlin . . . as an operational tool in its assault on Ukraine"[25] to advance its revisionist history propaganda campaign and to tout the success of its military efforts to garner support at home.[26]

In addition, we have witnessed the explosion of generative AI over the past year, which "can produce various types of content, including text, imagery, audio and synthetic data."[27] The most mainstream is currently text-based

[16] Human Rights Watch, *Meta's Broken Promises, Systemic Censorship of Palestine Content on Instagram and Facebook* (Dec. 21, 2023).

[17] Rohit Shewale, *Twitter Statistics in 2023 – (Facts After "X" Rebranding)*, DemandSage (Sept. 16, 2023).

[18] *See, e.g.*, Safiya U. Noble, Algorithms of Oppression: How Search Engines Reinforce Racism (2018).

[19] *See, e.g.*, Emily Wilder & Julian Crown, *Disinformation & Decentralization: How Viral Videos Helped Spread Election Denialism in Arizona's 2022 Midterm* (Apr. 9, 2023).

[20] *See, e.g.*, Ramez Koury et al., *Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter*, 12 Cureus e7255 (2020).

[21] David Gilbert, *The Israel-Hamas War Is Drowning X in Disinformation*, Wired (Oct. 9, 2023).

[22] Bellingcat, *Hamas Attacks, Israel Bombs Gaza and Misinformation Surges* (Oct. 11, 2023).

[23] Shayan 86, @shayan86, X (Oct. 8, 2023, 8:44 p.m.).

[24] *See, e.g.*, Bellingcat, *Identifying Possible Crater from Gaza Hospital Blast* (Oct. 18, 2023).

[25] Council of Europe, *EU Imposes Sanctions on State-Owned Outlets RT/Russia Today and Sputnik's Broadcasting in the EU* (Mar. 2, 2022).

[26] OECD, *Disinformation and Russia's War of Aggression Against Ukraine: Threats and Governance Responses* (Nov. 3, 2022).

[27] George Lawton, *What Is Generative AI? Everything You Need to Know*, Tech Accelerator.

https://doi.org/10.1017/aju.2024.5 Published online by Cambridge University Press

Chat GPT, but there are steadily more user-friendly AI that can create increasingly realistic photos and videos, such as Dall-E, Craiyon, Deep AI, and Runway. Deepfake technology poses significant challenges to the reliability of user generated content, as it contributes to the "truth decay" of our information ecosystem.[28]

Disinformation and deepfakes help to fuel false narratives and whip up support for parties who manipulate the truth for their own gains. Most of us are not yet equipped with the visual verification techniques necessary to identify fake posts, which can be particularly dangerous in a conflict situation, where inflammatory content may lead to increased violence on the ground. Several platforms, such as Facebook, now have independent fact-checkers to help review content uploaded by users,[29] and many media outlets, such as the New York Times, will provide fact checking services, but they do not catch everything and enthusiasm for these efforts may be waning.[30] Against this backdrop, digital investigators must carefully authenticate content. This is the process by which content is geolocated and chronolocated to ensure that a post is what it claims to be, and not something that is being taken out of context or repurposed. Ensuring content is methodically analyzed in this way will help to dispel some of the concerns with using it in war crimes prosecutions.

The Admissibility of Digital Evidence

Once investigators have discovered relevant digital content of potential war crimes and determined its reliability, the question of its admissibility will need to be addressed in each jurisdiction in which its submission is sought. A primary concern is establishing the chain of custody of digital evidence, which is vital to many criminal justice systems and requires demonstration of where and when evidence was received and stored. Luckily, there has been significant evolution in this space over the past decade: there are now advanced web capture tools like Hunch.ly and digital archives, such as those created and maintained by Mnemonic, to forensically store and secure content. In addition, the Berkeley Protocol on Digital Open-Source Investigations outlines methods to effectively preserve chain of custody and to create minimum standards of practice for digital investigators seeking to gather evidence of international crimes.[31] Recently, the International Criminal Court (ICC) has launched its own evidence submission platform, OTPLink, to provide users with a "seamless and secure method for submitting potential evidence."[32] All of these innovations allow us to place increasing trust in the chain of custody of digital evidence documenting international law violations.

Using the ICC as an example, generally, the court takes a liberal approach to the admission of evidence.[33] It has relied heavily on publicly available information to build its cases since the court's inception,[34] and since the early 2010s has increasingly utilized digital evidence, particularly user-generated content, such as social media posts.

[28] Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Cal. L. Rev. 1753 (2019).

[29] Facebook Help Center, *How Is Facebook Addressing False Information Through Independent Fact-Checkers?*

[30] Tiffany Hsu & Stuart A. Thompson, *Fact Checkers Take Stock of Their Efforts: "It's Not Getting Better,"* N.Y. Times (Sept. 29, 2023).

[31] Berkeley Protocol on Digital Open-Source Investigations: A Practical Guide on the Effective Use of Digital Open-Source Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law (United Nations & University of California, Berkeley eds., 2022).

[32] ICC Press Release, ICC Prosecutor Karim A.A. Kahn KC Announces Launch of Advanced Evidence Submission Platform: OTPLink (May 24, 2023).

[33] Rome Statute of the International Criminal Court, Art. 69(4), July 17, 1998.

[34] Lindsay Freeman, *Prosecuting Atrocity Crimes Using Open Source Evidence, Lessons from the International Criminal Court*, *in* Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability 51 (Sam Duberly, Alexa Koenig & Daragh Murray eds., 2020).

https://doi.org/10.1017/aju.2024.5 Published online by Cambridge University Press

Beginning in 2013, the ICC relied on Facebook to establish a relationship between parties involved in alleged witness tampering in the case against Jean-Pierre Bemba Gombo.[35] In the 2016 case of Ahmad Al-Faqi Al-Mahdi, a range of digital information was used to help to secure Al-Mahdi's guilty plea and his conviction as a co-perpetrator for the war crime of destruction of cultural property.[36] In 2017, the Court relied upon videos uploaded to Facebook that showed the Libyan General Mahmoud Mustafa Busayf al-Werfalli directly involved in the killing of thirty-three persons in order to issue an arrest warrant for him.[37]

The biggest test for the use of digital evidence at the ICC, and particularly user-generated content, is likely to come from cases stemming from Russia's war with Ukraine. Since the armed conflict broke out in February 2022, millions of pieces of content have been collected documenting alleged violations. Groups such as Bellingcat are tracking a wide variety of harms,[38] and Mnemonic has built a digital archive to forensically store evidence.[39] How exactly the Office of the Prosecutor will rely on this information in its investigation and any resulting cases remains to be seen, but it is likely to play a significant role as ICC Chief Prosecutor Karim Khan has emphasized the importance of collecting digital materials in this and other situations.[40]

Digital evidence can be highly probative as evidence of the necessary intention and commission of international crimes. However, as this essay has shown, numerous challenges can arise in its collection. Even when digital evidence is available, it is not a silver bullet; we are unlikely to have a video showing the entire chain of an event, including the perpetrator's role in its commission. But we can use user generated content to corroborate other evidence, to provide contextual information about the conflict, to lead us to potential witnesses who may be able to provide testimony, or as linkage evidence to establish connections and chains of command between alleged perpetrators.[41]

Concluding Observations

To achieve the potential of digital evidence to transform the accountability landscape, we must continue to grapple with how to enhance its availability, reliability, and admissibility before courts. Social media platforms must be required to make potential evidence of war crimes available to investigators and must commit to not silencing voices online. Platforms must also do a better job of identifying disinformation and AI generated deepfakes and placing warnings on posts that are spreading falsehoods. We must harness the power of AI for good and use machine learning to help investigators to identify and analyze relevant content. The ICC is already deploying AI to help with evidence pattern analysis through its digitization project, Project Harmony.[42] This and other initiatives should be supported by governments, big tech, and civil society to create a digital infrastructure capable of harnessing the power of user-generated content to pursue accountability for international crimes.

[35] Prosecutor v. Jean-Pierre Bemba Gombo, ICC-01/05-01/08-2721, Decision on the Admission into Evidence of Items Deferred in the Chamber's "Decision on the Prosecution's Application for Admission of Materials into Evidence Pursuant to Article 64(9) of the Rome Statute" (June 27, 2013).

[36] Prosecutor v. Ahmad Al Faqi Al Mahdi, ICC-01/12-01/15, Judgment and Sentence (Sept. 27, 2017).

[37] Prosecutor v. Mahmoud Mustafa Busayf Al-Werfalli, ICC-01/11-01/17, Warrant of Arrest (Aug. 15, 2007).

[38] Bellingcat, *Civilian Harm in Ukraine Timemap*.

[39] Mnemonic, *Ukrainian Archive*.

[40] ICC Press Release, ICC Prosecutor Karim A.A. Khan QC Announced Deployment of Forensic and Investigative Team to Ukraine, Welcomes Strong Cooperation with the Government of the Netherlands (May 17, 2022).

[41] *See, e.g.*, Bellingcat, *Skripal Poisoner Attended GRU Commander Family Wedding* (Oct. 14, 2019).

[42] ICC Press Release, ICC Office of the Prosecutor to Launch Modern Evidence Management Platform (Feb. 8, 2023).

https://doi.org/10.1017/aju.2024.5 Published online by Cambridge University Press