doi:10.1017/aju.2019.47

SYMPOSIUM ON NON-STATE ACTORS AND NEW TECHNOLOGIES IN ATROCITY PREVENTION

"HALF THE TRUTH IS OFTEN A GREAT LIE": DEEP FAKES, OPEN SOURCE INFORMATION, AND INTERNATIONAL CRIMINAL LAW

*Alexa Koenig**

The video opens with a man—now known to be Commander Al-Werfalli of the elite Libyan Special Forces team al Saiqa—throwing a quick glance over his shoulder as a black SUV rolls off camera. Cradling a gun, he saunters towards three men kneeling on a sidewalk, hands bound behind their backs, faces turned toward the wall. He raises the gun in his left hand. Without pausing, he walks methodically down the row, a bullet punctuating every other step. The men slump forward as they fall. Werfalli's first hint of emotion is visible only when he unloads multiple bullets into the final body.[1]

In February 2011, the International Criminal Court's (ICC's) Chief Prosecutor announced that the Court would open an investigation into alleged crimes in Libya.[2] Within seven years, the ICC issued two warrants for Werfalli. The first alleged that he was responsible for the war crime of murder, based on the deaths of thirty-three people between June 2016 and July 2017.[3] The international legal community considered the warrant a milestone, as it represented the first time the ICC cited videos pulled from social media—often referred to as open source evidence—as a basis for its charges.[4] The second ICC warrant asserted Werfalli's responsibility for the deaths of ten more people; the Prosecutor again relied on public information. This included another video pulled from social media, a UN report,[5] and a statement in which al Saiqa claimed responsibility for the killings.[6]

While the emerging practice of online open source investigations—the process of investigating and analyzing publicly-accessible online information[7]—has tremendous promise for helping to build international criminal cases,[8] the recent success of integrating such content into case files is threatened by *other* emerging

* *Executive Director, Human Rights Center; Lecturer, University of California, Berkeley School of Law.* *Thanks to Lindsay Freeman, Alexandra Huneeus, Andrea Lampros, Daragh Murray, and Larissa van den Herik for their feedback.*

[1] Bellingcat, *Werfalli Libya Video Two*, VIMEO (Sep. 20, 2017) [warning: graphic content].

[2] Int'l Crim. Ct. Press Release, Situation in Libya: ICC Pre-Trial Chamber I Issues a Second Warrant of Arrest for Mahmoud Mustafa Busayf Al-Werfalli for War Crimes (July 5, 2018).

[3] Prosecutor v. Al-Werfalli, Case No. ICC-01/11-01/17, Warrant of Arrest (Aug. 15, 2017).

[4] *Id.*

[5] Cited in the warrant as LBY-OTP-0060.

[6] Prosecutor v. Al-Werfalli, Case No. ICC-01/11-01/17, Second Warrant of Arrest (July 4, 2018).

[7] International Protocol on Open Source Investigations (U.C. Berkeley Human Rights Center, forthcoming 2019).

[8] Lindsay Freeman, *The Impact of Digital Technologies on International Criminal Investigations and Trials*, 41 FORDHAM INT'L L.J. 283 (2017). U.C. Berkeley's Human Rights Center has been leading an international effort to develop standards for the use of open source information as evidence in prosecutions of international crimes.

practices.[9] One particularly acute threat is presented by "deep fakes"—videos generated via algorithms that make it look like a person said or did something she did not.[10] For example, what if the Werfalli videos had been modified to appear as though Werfalli committed the crimes, when really it was his opposition? While in the *Werfalli* case the likelihood of that having happened is low, the dangers of relying on such disinformation are profound: such videos could theoretically result in everything from false convictions to an undermining of trust in entire fields of practice, ranging from journalism to human rights advocacy to law.

This essay discusses how digital content derived from open sources is impacting the practice of international criminal law, expanding the role that non-state actors such as human rights NGOs play in locating, preserving, verifying, and analyzing online visual imagery. These actors use information and communication technologies to send videos and photographs directly to human rights organizations and courts, and compile information that has been posted to social media sites like YouTube, Twitter, and Facebook with the goal of seeing that data used in court.[11] The essay also asks how international lawyers should prepare for the coming storm of deep fakes, especially when they are only just beginning to mainstream the use of open source videos to document international crimes. The essay proposes several strategies, ranging from working cross-disciplinarily to systematically analyzing contextual information.

Assembling Open Source Content

Since 2011, the ICC and other tribunals have increasingly incorporated online open source content into their investigations to corroborate witness testimony and fill evidentiary gaps.[12] The sources of such content are often non-state actors, including survivors and perpetrators. Such content is abundant: the latest statistics establish that as of late 2018, more than six thousand tweets are produced every second,[13] and more than five hundred hours of video are uploaded to YouTube every minute.[14] Whereas only a few years ago, the international legal community had a volume problem based on a scarcity of visual content,[15] today the challenge is to find the signal in all the noise.

Given this volume, many organizations are experimenting with machine learning processes to help locate, analyze, and preserve this extraordinary amount of data. However, the challenges are similarly voluminous, from needing sufficient quantities of training data (for example, to help machines learn to identify certain kinds of weapons across large data sets of videos or photos) to developing algorithms capable of surfacing helpful content. The most critical information may at first not seem to relate to an atrocity: for example, the key piece of evidence may not be a video of an extrajudicial killing, but a satellite image of a bare patch of earth that suggests a mass grave[16] or a still image of two people smoking that ties a commander to a frontline perpetrator.

International investigators are beginning to successfully navigate this new terrain, acknowledging that "the future of [accountability] will be intertwined with the advancement of technology."[17] They have strengthened

[9] *See generally* Rebecca Hamilton, *User-Generated Evidence*, 57 Colum. J. Transnat'l L. 1 (2018).

[10] *See, e.g.*, Jon Christian, *Experts Fear Face Swapping Tech Could Start an International Showdown*, The Outline (Feb. 1, 2018).

[11] *See, e.g.*, Syrian Archive.

[12] Int'l Crim. Ct. Office of the Prosecutor, Strategic Plan 2016-2018, at 23 (Nov. 16, 2015).

[13] Twitter Usage Statistics, Internet Live Stats.

[14] Mark R. Robertson, *500 Hours of Video Uploaded to YouTube Every Minute [Forecast]*, Tubular Insights (Nov. 13, 2015).

[15] Iva Vukusic, *Nineteen Minutes of Horror*, 12 Gen. Studies & Prevention: An Int'l J. 35 (2018).

[16] Jonathan Drake & Theresa Harris, *Geospatial Evidence in International Human Rights Litigation*, AAAS (2018).

[17] Enrique Piracés, *The Future of Human Rights Technology*, *in* New Technologies for Human Rights Law and Practice 289–308 (Molly K. Land & Jay D. Aronson eds., 2018).

https://doi.org/10.1017/aju.2019.47 Published online by Cambridge University Press

networks for sharing digital content, integrated new actors into practice (such as computer programmers and data scientists), and developed new methods for discovering and verifying online information. In addition, several groups have developed guidelines—including an international protocol,[18] model standard operating procedures, and a "video as evidence" field guide[19]—to help the international legal community better understand how to use open source information to build cases around grave international crimes.

All of these efforts harness new technologies to strengthen accountability. However, the recent emergence of "deep fakes" threatens lawyers' ability to trust the videos they find online just as they are starting to embrace open source content.

What Are Deep Fakes and How Do They Work?

The term "deep fake" refers to manufactured imagery that is developed via generative adversarial networks, a process that pits two neural networks against each other. The first network, known as the "generator," produces a sample output (such as an image) based on an underlying dataset of images, which is then evaluated by the "discriminator," which provides critical feedback about the generator's success in replicating the characteristics of the underlying data. The two iterate to generate increasingly realistic "fakes" that come closer and closer to the images in the original dataset and thus to seeming as if a false event actually occurred,[20] perhaps most famously using President Obama as the subject.[21]

The term "deep fake" is a nod to both "deep learning"—a category of artificial intelligence through which algorithms train each other to increasingly perfect their ability to do something[22]—and false (or "fake") information. Most notoriously advanced in the service of pornography, especially by a Reddit user known as "deepfakes," to meld the faces of famous actresses onto the bodies of porn actors—the term also hints at the 1972 American film Deep Throat, an icon from the golden age of pornography. Since its inception, the potential uses for this technology have exploded.

Challenges and Opportunities

The challenges of detecting deep fakes are significant. These include 1) absence of a legal workforce specifically trained in visual verification techniques, which can be a first line of defense against being duped; 2) the increasing sophistication and decreasing costs of deep learning technologies, which means they are becoming increasingly realistic and available to a wide variety of actors; and 3) an information ecosystem in which trust in facts—and in their sources—is being degraded, with malicious actors disseminating headline-grabbing lies.

To date, Robert Chesney and Danielle Citron have drafted the most thorough analysis of the consequences of deep fakes,[23] underscoring that "[o]ne of the prerequisites for democratic discourse is a shared universe of

[18] International Protocol, *supra note* 7.

[19] Kelly Matheson, Video as Evidence Field Guide (2016).

[20] *See, e.g.*, Tianxiang Shen et al., *Deep Fakes Using Generative Adversarial Networks*, NOISELAB (2018).

[21] *Fake Obama Created Using AI Tool to Make Phoney Speeches*, BBC NEWS (July 17, 2017). Audio fakes are also becoming increasingly sophisticated.

[22] *See* Marc Jonathan Blitz, *Lies, Line Drawing, and (Deep) Fake News*, 71 OKLA. L. REV. 59, 106 (2018); *see also* James Vincent, *Why We Need a Better Definition of Deep Fake*, VERGE (May 22, 2018).

[23] Robert Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Security and National Security*, 107 CAL. L. REV. (forthcoming 2019).

https://doi.org/10.1017/aju.2019.47 Published online by Cambridge University Press

facts … supported by empirical evidence."[24] So how do we help safeguard that universe? The next part of this essay suggests three categories of responses that could be integrated into international strategies for combating deep fakes and other disinformation that can contaminate evidence pools. These include legal, pedagogical, and technological responses.[25]

Legal Responses

One option is to disincentivize the malevolent creation and dissemination of deep fakes by imposing legal liability. Legal scholars have begun contemplating the theories that could be used to regulate the creation or use of deep fakes and create remedies for harms.[26]

At the international level, UN Special Rapporteur David Kaye has considered how a regulatory framework might evolve to address the dangers that emerge from artificial intelligence processes. He emphasizes the need to build responses that respect human rights, safeguarding key principles like freedom of expression.[27] A coalition of organizations, including the Office of the High Commissioner of Human Rights, has issued a joint declaration establishing similar high-level guidance.[28]

In the United States, Marc Jonathan Blitz argues that a deep fake, as a variant of fake news, should "lose [free speech] protection … when it is not only a falsity, but a forgery,"[29] such as a manufactured video. However, like Kaye, he notes that "some lies have value,"[30] arguing that we should distinguish between fakes that harm "persons and property" and those that do not.[31] While this distinction may sometimes be difficult to draw, such an approach would ideally permit deep fakes created for entertainment purposes (such as Princess Leia in Star Wars' Rogue One after actress Carrie Fisher's death), while punishing the use of deep fakes to negatively impact security or property rights or the effective functioning of democracies.[32]

Individual states have begun to respond, but so far that response has been problematically piecemeal and narrow. For example, a UK bill focused on "image-based sexual abuse" included penalties for pornographic deep fakes.[33] But harms from deep fakes can extend far beyond reputational damage. These can include identity theft, trademark and copyright violations, election manipulation, eroded trust in institutions, and damaged international relations.[34] Any legislation will need to reflect this reality. Ultimately, gatekeeping control "remains with the companies responsible for our digital infrastructure," and therefore legal responses must be designed around that fact.[35] Because digital technologies disrespect territorial borders, the world will need coordinated mechanisms for detecting and responding to deep fakes.

[24] *Id.* at manuscript p. 21.

[25] While there are many other potential categories of response, they exceed the scope of this essay.

[26] *See, e.g.*, Blitz, *supra* note 22.

[27] David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), Report, UN Doc. A/73/348 (Aug. 29, 2018).

[28] Joint Declaration on Freedom of Expression and "Fake News," Disinformation and Propaganda (Mar. 3, 2017).

[29] Blitz, *supra* note 22, at 63–64.

[30] *Id.* at 82.

[31] *Id.* at 72.

[32] *Id.* at 73–74.

[33] *See, e.g.*, Dan Sabbagh & Sophia Ankel, *Call for Upskirting Bill to Include "Deepfake" Pornography Ban*, GUARDIAN (June 21, 2018).

[34] Chesney & Citron, *supra* note 23, at manuscript p. 4, 20–21.

[35] *Id.* at 10; Kaye, *supra* note 27.

https://doi.org/10.1017/aju.2019.47 Published online by Cambridge University Press

Pedagogical Solutions

Education—in the form of awareness and training—can also mitigate the harms posed by deep fakes. One option is to encourage investigators to embrace "multiple working hypotheses" about the credibility of visual and auditory content, while developing an international coalition of lawyers trained in digital discovery and verification techniques. Ultimately, lawyers must master these techniques—and quickly. While open source investigations have long been conducted by journalists, lawyers are relatively new to their use. Law schools are woefully unprepared to provide this training and will need to catch up fast.[36] However, these skills are not hard to develop. Open source verification relies on a relatively simple two-step process: an evaluation of 1) source, and 2) content. Each step encompasses safeguards that may help to detect deep fakes.

First, open source investigators determine the reliability of their sources. They ask whether the source is someone they know and/or someone they have successfully relied on before. If unknown, is there information that suggests that the source was at the place necessary to have captured the content? For example, if the source alleges that a video was taken in a particular village in Liberia, is that person regularly in Liberia? A scan across social media sites using known profile pictures, phone numbers, or handles provides a helpful check.

Second, open source investigators use multiple methods to corroborate content. For example, time of day can be checked with metadata. While not dispositive, since metadata can be faked, it can provide critical lead information for further corroboration. When missing, metadata can often be reconstructed through the geolocation of photographs and other content.[37] Alleged time of day can be checked against shadow angles; weather records can help confirm whether visual content is consistent with a particular day, time, and place. (For example, a picture with abundant sunlight allegedly captured at 6 a.m. in Sweden in December would be suspect.) Simple methods like reverse image searching stills grabbed from videos can lead to the original videos from which a fake may have been generated.

Finally, given the ubiquity of smartphones, international crimes are often documented by multiple individuals. Investigators can scan social media for tweets or posts that come from similar locations along a particular timeline. The long-term risk is that coordinated but apparently unconnected groups will plant doctored, corroborating information—but even then contradictory content will often be available.

Ultimately, the gold standard for any legal investigation is to triangulate documentary, physical, and testimonial evidence. For both professionals and lay people, training in methods for cross-checking open source content should begin in earnest; reestablishing confidence in facts will help to counter the "truth decay and trust decay"[38] upon which malevolent actors depend.

Technology-Based Solutions

Developers are increasingly experimenting with ways to automate forensic analysis of photographs and videos. For example, InVID offers an online tool that allows verifiers to check the integrity of digital content to assess reliability. Project Maru by Gyfcat uses artificial intelligence to spot deep-fakes based on imperfections in frames. Project Angora, also by Gyfcat, conducts reverse image searches by blocking the face (the segment of the image most likely superimposed on the original footage, as with "celebrity" porn) and searching for the remainder of the image elsewhere online.[39]

[36] For the first textbook on this, see Digital Witness: Using Open Source Information for Human Rights Documentation, Advocacy and Accountability (Sam Dubberley et al. eds., 2019).

[37] *Id.*

[38] Chesney & Citron, *supra* note 23, at 29.

[39] Louise Metsakis, *Artificial Intelligence is Now Fighting Fake Porn*, Wired (Feb. 14, 2018).

https://doi.org/10.1017/aju.2019.47 Published online by Cambridge University Press

While these technologies are promising first steps for detecting fakes, Hany Farid warns that no forensic technology is perfect, and notes that those committed to digital manipulation will start building ways to fool digital detectors into their technology.[40] However, deep-fake technology is also imperfect. Some flaws may not even require forensic tools but may simply be detected with careful observation. For example, deep fakes may include eyes that stay open longer than those of actual humans.[41] Breath rates may also vary from human norms.

Ultimately, at least for now, the danger of being duped is far greater for journalists (who may have to determine a video's veracity within minutes) than for international lawyers. In the latter case, the systematic use of verification and corroboration is key.

Conclusion

In 1758, Benjamin Franklin wrote, "[H]alf the truth is often a great lie." Deep fakes are today's digital manifestation of that statement. Generated from two sets of "real" data—images of one person merged with images of another—such falsehoods are increasingly difficult to parse from facts. As international lawyers look to videos posted online to corroborate the stories of survivors, they must pay attention to the ways in which malevolent actors plant and spread disinformation, making case-building ever more challenging. While advances in technology threaten to obfuscate the who, what, when, and where of international crimes, a careful analysis of "how we know" that a piece of content is what it claims to be will go a long way towards safeguarding our legal institutions.

[40] Chesney & Citron, *supra* note 23, at 30.

[41] Sarah Scoles, *These New Tricks Can Outsmart Deepfake Videos—For Now*, WIRED (Oct. 17, 2018).

https://doi.org/10.1017/aju.2019.47 Published online by Cambridge University Press