Bad Speech, Good Evidence: Content Moderation in the Context of Open-Source Investigations

Hillary Hubley | ORCID: 0000-0003-3204-3972
Amsterdam Law School, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
hh2407@columbia.edu

Abstract

This article explores how content moderation on social media platforms impacts the work of open-source investigators through its routine removal of content having evidentiary value. These practices have rendered social media platforms susceptible to public criticism and scrutiny. However, these criticisms have largely been levied by a community who cares about content removal’s impact on free expression online. This swath of concerns does not comport with those of international criminal investigators who have increasingly turned to platforms for evidence gathering. Rather than confronting the issue, investigators have absorbed the costs by downplaying the impact of content removal on their work and by seeking to preserve the content on their own. I examine the disconnect between these two groups in their respective approaches to the problem of content removal and argue that both communities can stand to benefit from joining forces and taking notice of the convergence of their respective concerns.

Keywords

open-source investigation – content moderation – free expression – social media – international criminal investigations – digital evidence

1 Introduction

Social media has quickly become the world’s most favoured source for communication and news. It is estimated that 3.6 billion people—roughly half of the global population—use social media, and this number is projected to increase
to almost 4.41 billion by 2025.\(^1\) This development has brought people closer together and allowed them to create a record of their experiences, good and bad.

Indeed, social media has served as a platform for otherwise disenfranchised and powerless individuals to document the injustices they experience. This content not only serves the purpose of rendering its viewers secondary witnesses to atrocities, but can also prove to be valuable as evidence should a criminal proceeding arise. Both perpetrators and victims of human rights abuses use social media, which facilitates the unparalleled opportunity to view a conflict from both sides.

Investigators and journalists have harnessed this moment by relying on social media as a tool for furthering their investigations. It is not only tempting, but well-advised to review all relevant online content mobilised in service of fact-finding. This practice, termed open-source investigation, has gained immense traction and support among practitioners in the international criminal space.

However, investigators relying upon social media platforms enter a fraught landscape. The technology they are using is far from perfect, and there are many factors at play which can impede an open-source investigation. One of them is the ominous possibility of the evidence disappearing.

This article focuses on such a problem—the removal of digital content by social media platforms endeavouring to moderate their spaces. Unsurprisingly, it is the content of potential probative value that runs the greatest risk of a takedown.\(^2\) Once this content is removed, unless it is independently preserved or successfully restored, it cannot be used by either party in a legal proceeding.

Content moderation affects many facets of the investigation, including whether evidence of potential probative value is available, and for how long. As courts and tribunals begin to welcome the by-product of these open-source investigations as evidence, \textit{all} stakeholders should be aware of the possibility of it disappearing without recourse and their own limitations in engaging with this problem without addressing it. These concerns are palpable, and investigators’ efforts to downplay them or shoulder the consequences on their own

\begin{enumerate}
\item Number of Social Media Users Worldwide, \textit{Statista}, available online at www.statista.com/statistics/278414/number-of-worldwide-social-network-users (accessed 16 June 2021).
\item Human Rights Center, UC Berkeley School of Law and United Nations Human Rights Office of the High Commissioner, ‘Berkeley Protocol on Digital Open-Source Investigations’ (hereinafter ‘Berkeley Protocol’), available online at www.ohchr.org/Documents/Publications/OHCHR_BerkeleyProtocol.pdf, para.150. See also ‘Video Unavailable: Social Media Platforms Remove Evidence of War Crimes,’ \textit{Human Rights Watch} (September 2020) (hereinafter ‘HRW’).
\end{enumerate}
does a disservice to the practice of open-source investigations and the broader content moderation debate.

The disconnect between how the respective communities view and approach the issue is rooted in a divergence in discourse. To some, admissibility and viability of content in the field of criminal justice becomes a secondary concern when juxtaposed with the broader threat to freedom of expression. However, by viewing content moderation primarily as a speech issue, one may capture many but not all of its possible consequences. This article focuses on one of these consequences—the problem of ‘bad speech, good evidence,’ or content that is indeed harmful and destructive, but could serve as valuable evidence in a criminal proceeding.

2 Content Moderation in the Human Rights Arena

2.1 Content Moderation: An Introduction

UN Special Rapporteur on Freedom of Expression David Kaye defines content moderation as: ‘the process by which Internet companies determine whether user-generated content meets the standards articulated in their terms of service and other rules.’ Internet companies, or for the purpose of this article—platforms—can be defined as ‘services that ‘host, organise, and circulate users’ shared content or social actions for them,’ without having produced a bulk of that content themselves.

Moderation can also be viewed in the more abstract sense. Tarleton Gillespie argues that moderation is the commodity platforms offer, making order out of chaos and simplifying access to information its users are looking for.

Whether they want to or not, platforms find that they must serve as the setters of norms, interpreters of laws, arbiters of taste, adjudicators of disputes, and enforcers of whatever rules they choose to establish. Having in many ways taken custody of the web, they now find themselves its custodians.

---

3 Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (A/HRC/38/25), 6 April 2018 (hereinafter ‘Kaye Report’), footnote 2.
4 B. Sander, ‘Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation,’ 43 Fordham International Law Journal (2020) 939–1006, footnote 21.
5 T. Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, (Yale University Press, New Haven, CT, 2018), p. 11.
6 Ibid., p. 6.
Indeed, Gillespie purports: ‘Moderation is not an ancillary aspect of what platforms do. It is essential, constitutional, definitional. Not only can platforms not survive without moderation, there are no platforms without it.’7 Given the critical importance of this act, content moderation can be a conduit for understanding platforms more broadly. It also underscores how relevant this practice is to someone who relies on these platforms for investigative purposes.

Understanding content moderation is not easy. All platforms handle it differently, and their approaches evolve over time.8 Furthermore, content moderation functions in secret, and that is by design. As Gillespie states, ‘social media platforms are vocal about how much content they make available, but quiet about how much they remove…with so much available, it can start to seem as if nothing is unavailable.’9

However, lots of content is rendered unavailable every day. Content removal is one form of content moderation that platforms often rely upon, particularly regarding extremist or violent content.10 This process is driven by platforms themselves. There is very little they are uniformly required to take down by law,11 and some jurisdictions do not even have the capacity to regulate online content.12 As a result, their policies for takedown often go beyond pure legal compliance. In turn, platforms rarely engage in line-drawing between what must and what may be removed. This is further complicated in the international context, given the patchwork of practices that span the jurisdictions at issue in an international criminal investigation.

Sander considers four driving forces that motivate platforms to remove content: corporate philosophy, regulatory compliance, profit maximisation, and public outcry.

---

7 Ibid., p. 21.
8 Supra note 4, p. 949.
9 Supra note 5, p. 7.
10 See I. van der Vegt, P. Gill, S. Macdonald and B. Kleinberg, ‘Shedding Light on Terrorist and Extremist Content Removal’, Global Research Network on Terrorism and Technology (2019), available online at https://rusi.org/explore-our-research/publications/special-resources/shedding-light-on-terrorist-and-extremist-content-removal.
11 See Directive No. 2000/31/EC of the European Parliament and of the Council of 8 June 2000; 47 U.S.C. para. 230. See also Kaye Report, supra note 3, para.14. Rather than focusing on what must be taken down, the focus shifts to indemnifying platforms for hosting content provided they introduce comprehensive takedown and monitoring schemes.
12 R. Hamilton, ‘Governing the Global Public Square’, 62 Harvard International Law Journal (2021) 117–174, pp. 125–126. Hamilton attempts to illustrate how focus on the US and EU has rendered scholarship on content moderation myopic to the relationships between platforms and states in many parts of the Global South, explaining that not everywhere is the ‘rule of law in operation’ such that the state is capable of regulating content online.
Corporate philosophy can be defined as the high-level goals platforms wish to achieve in designing an experience for their users. It is typically what motivates the company’s terms of service—their own rules of the road for posting content and participating in their network.\textsuperscript{13}

Regulatory compliance captures the body of laws and norms from national and supranational bodies. Sander bifurcates this into mandatory and informal regulation. Mandatory regulation consists of the existing law governing platform regulation, such as NetzDG in Germany which levies heavy fines on platforms if they do not remove ‘clearly illicit’ content within 24 hours.\textsuperscript{14} A recent example is the European Parliament regulation which mandates platforms remove terrorist content ‘within one hour’\textsuperscript{15} of receiving an order to do so from the authorities. This gives member states the opportunity to compel removal of content anywhere within the bloc,\textsuperscript{16} and requires platforms take up ‘specific measures’ to protect against the dissemination of terrorist content.\textsuperscript{17}

More broadly, many states have content restriction laws that compel certain content to be removed,\textsuperscript{18} or intermediary laws which require the implementation of rigorous procedural safeguards and monitoring schemes to render platforms immune from prosecution.\textsuperscript{19} These laws may incentivise removal by granting safe harbour to platforms that employ a notice-and-takedown regime.\textsuperscript{20} Informal regulation may entail a state agency reaching out to a platform to compel takedowns on an \textit{ad hoc} basis, or engaging directly with platforms in an effort to exert influence over their terms of service.\textsuperscript{21}

Profit maximisation is a motivation Gillespie alludes to in his work. Ultimately, platforms are for-profit entities seeking to design a popular

\begin{thebibliography}{99}
\item [13] \textit{Supra} note 4, p. 949.
\item [14] Beschlussempfehlung und Bericht [Resolution and Report], Deutscher Bundestag: Drucksache [BT] 18/13013, official English translation.
\item [15] Regulation No. 14308/1/20 REV 1 of the European Parliament and of the Council on addressing the dissemination of terrorist content online, Annex 1 to Article 3.
\item [16] \textit{Ibid.}, Articles 3 and 4.
\item [17] \textit{Ibid.}, Article 5.
\item [18] \textit{Supra} note 4, p. 950.
\item [19] \textit{See, e.g.}, D. Keller, ‘Internet Platforms: Observations on Speech, Danger, and Money’, \textit{Hoover Institution’s Aegis Paper Series} 1807 (2018), pp. 18–20.
\item [20] E. Newland, C. Nolan, C. Wong and J. York, ‘Account Deactivation and Content Removal: Guiding Principles and Practices for Companies and Users,’ \textit{The Berkman Center for Internet & Society and The Center for Democracy & Technology} (September 2011), available online at https://cyber.harvard.edu/sites/cyber.harvard.edu/files/Final_Report_on_Account_Deactivation_and_Content_Removal.pdf.
\item [21] \textit{Supra} note 4, p. 951. Indeed, the mere shadow or threat of regulation can be construed as informal regulation at work.
\end{thebibliography}
product, and their currency of choice is human attention. Curating content, and scrubbing undesirable content, helps serve the goals of ‘surveillance capitalism’ by luring users into engaging with the platform. This ‘incentivizes online platforms to moderate content in ways that aim to maximize both user engagement and advertising revenue on their platforms.’

Finally, there is the motivation of public outcry. Platforms are not immune to criticism and can tend to buckle when confronted with overwhelming public response. Although these responses are rendered most effective through high profile media campaigns or impact litigation, platforms may be inclined to listen to a groundswell of public support—or criticism—and respond accordingly.

2.2 Content Moderation in Practice

Understanding what motivates platforms to moderate content can help shed light on how they do so. Although each platform employs its own means and methods, they all follow a similar protocol.

Ex post moderation is the removal of content that occurs after it is posted and available for consumption. Indeed, for many of these platforms ‘the norm is to allow all uploads without pre-screening.’ As a result, ex-post moderation commonly occurs when a user of a platform flags content they believe to be hateful, extremist, or otherwise inappropriate, based on their own subjective standards. This ‘community flagging’ approach is not the final straw, however, as anything that is flagged enters a queue to be reviewed by a content moderator.

Professional content moderators, or ‘commercial content moderators,’ are hired by platforms to view and remove extremist content. Their role is to view all content ‘flagged’ for review, either by the ‘community’ or by an algorithm employed by the platform. Commercial content moderators span the

---

22 Ibid., p. 952.
23 Ibid., p. 954.
24 Ibid.
25 S. Roberts, Behind the Screen: Content Moderation in the Shadows of Social Media (Yale University Press, New Haven, CT, 2019), p. 35.
26 A.R. Al Jaloud, H. Al Khatib, J. Deutch, D. Kayyali and J. York, Caught in the Net: the impact of ‘extremist’ speech regulations on human rights content, Whitepaper, Electronic Frontier Foundation, the Syrian Archive, and witness (2019), available online at https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content, p. 4.
27 The term ‘commercial content moderator’ is coined by Sarah Roberts in her book Behind the Screen (Yale University Press, New haven, CT, 2019).
globe, and often make snap, human decisions about whether the content at issue indeed violates the platform's policies.28

Although content moderation was initially performed entirely by humans, their role is growing increasingly obsolete.29 Algorithms can assist with ex-post moderation by reviewing content flagged by users, or by flagging content on their own.

This algorithmic flagging, also known as ‘automated flagging,’ is becoming the preferred mode of moderation for many platforms. Instead of paying human beings who are outmatched by endlessly proliferating content, platforms can ‘use their own proprietary tools to automatically detect potentially violating content.’30 While this model can work in tandem with a human moderator, eventually the choices algorithms make may become so nuanced that there will be no need for human input.31 Automated flagging also nicely complements the business model of many of these platforms—curating content that captures user attention and optimises engagement.32

Flagging triggered by such an algorithm and sanctioned either by man or machine is ex-ante moderation, or moderation prior to publication.33 These decisions are made in secret, by machines, ‘shielded from any external review.’34 Furthermore, ex-ante moderation is becoming ubiquitous. In 2018, 73% of the videos removed by YouTube were done so by machines and prior to publication.35

This type of review is particularly common on terrorist and extremist content. Here, platforms will use a tool known as hashing.

28 See, e.g., supra note 3, para. 35.
29 See e.g., supra note 2, p. 7: 93.4% of content taken down by platforms was flagged by an automated system. In March 2020, YouTube insinuated it would commence removing content without human review. See also The Impact of COVID-19 on Content Moderation, Syrian Archive (10 June 2020), available online at syrianarchive.org/en/lost-found/covid19-impact-on-content-moderation (accessed 17 June 2021). ‘Overnight these measures put in place to stop the transmission of COVID-19 have brought forward a future in which A.I., in the absence of its human co-workers, is empowered to aggressively moderate our social media feeds.’
30 Supra note 26, p. 4.
31 Ibid., p. 5.
32 T. Dias Oliva, ‘Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression’, 20 Human Rights Law Review (December 2020) 607–640, doi.org/10.1093/hrlr/ngaa032.
33 Supra note 4, p. 946.
34 Supra note 26, p. 5.
35 Ibid.
Hashes are digital fingerprints that companies use to identify and remove content from their platforms. They are essentially unique, and allow for easy identification of specific content. When an image is identified as ‘terrorist content,’ it is tagged with a hash and entered into a database, allowing any future uploads of the same image to be easily identified.36

In other words, platforms rely on the properties found in one piece of terrorist content to identify and prophylactically remove others.

Hashing, unlike other moderation tools, is not platform specific. In fact, all major social networks are part of a ‘hash-sharing consortium’ designed to combat the posting of terrorist-content. Founded in 2016, the Global Internet Forum to Counter Terrorism (‘gifct’) built up a database of ‘terrorist imagery’ in order to create hashes.37 Presently, virtually all platforms rely on gifct’s over 300,000 hashes of terrorist content to identify and block publication on their networks.38 These hashes inform when content has violated a platform’s terms of service—not the law—and facilitate a pre-publication takedown.39

GIFCT is not open to the public and has rejected requests for information and pleas to respect human rights. Its opacity renders it impossible to review or challenge, and it remains unclear what standards and practises the consortium employs.40 Furthermore, the fact that there is only one consortium renders it highly susceptible to error.41 From an evidentiary standpoint, this means most content of potential probative value is likely being removed through GIFCT’s hashing consortium.

There is no uniform notice-and-takedown procedure for expunged content. Sometimes the original poster will receive an automated message from the platform;42 other times no notice will be conferred.43 There is also no uniformity regarding preservation of the removed content. It is not known

36 Ibid.
37 Ibid.
38 Supra note 2, p. 9.
39 D. Kayyali, ‘European ‘terrorist content proposal is dangerous for human rights globally,’ Mnemonic (24 November 2020), available online at mnemonic.org/en/content-moderation/european-terrorist-content-proposal-dangerous-human-rights-globally (accessed 17 June 2021).
40 Supra note 2, p. 11.
41 Supra note 26, p. 5.
42 See, e.g., A. Ascher-Shapiro, ‘YouTube and Facebook are Removing Evidence of Atrocities, Jeopardizing Cases Against War Criminals,’ The Intercept (2 November 2017), available online at theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya (accessed 17 June 2021).
43 Supra note 3, para. 37.
how long platforms preserve expunged content before deleting it from their servers or even whether the content is, in fact, ever deleted from its servers.\textsuperscript{44} For instance, Facebook purports to preserve content on its servers for 90 days upon receiving a ‘valid request’ from law enforcement or some other entity, but there have been instances in which the content has remained for longer periods. YouTube once restored content two years after takedown.\textsuperscript{45}

Should an individual wish to challenge the platform’s decision to remove content, they will typically have to lodge an appeal with the platform. The same is true for independent human rights organisations and members of the media, as no special route to recovery exists for these groups.\textsuperscript{46} However, such processes are notoriously ill-fated and time-consuming,\textsuperscript{47} and have been lambasted by human rights groups as a denial of due process due to their opacity and lack of accessibility.\textsuperscript{48}

Takedowns occur regularly. In the period of January to March 2020, YouTube removed over 6 million videos; 93.4\% of which was flagged by a machine and 49.9\% before any user saw it.\textsuperscript{49} Facebook removed nearly 50 million pieces of purportedly hateful, terrorist, or graphic content in the same period.\textsuperscript{50} Such an operation becomes much easier with the help of automated systems and an unchecked review process.

\section*{2.3 Criticisms of Content Moderation}

It is evident that content moderation is a far from perfect system. The criticisms levied against it can loosely be divided into three buckets—moderation as a speech problem, moderation as a transparency and due process problem, and moderation as a memory problem. While all these concerns are legitimate and compelling, none neatly address the issues faced by open-source investigators.

\subsection*{2.3.1 Free Speech Problem}
Free speech academics and activists have deemed content moderation a growing, omniscient threat to free expression, emblematic of the shaky ground that speech rests upon in the digital age.\textsuperscript{51} Although the government is not at the

\begin{footnotesize}
\begin{enumerate}
\item Supra note 2, p. 2.
\item Ibid.
\item Ibid., p. 31.
\item Ibid., para. 38.
\item Supra note 4, p. 1002.
\item Supra note 2, p. 7.
\item Ibid.
\item See, e.g., ‘UN Expert: Content moderation should not trample free speech,’ \textit{UNHCR} (19 July 2018), available online at www.ohchr.org/EN/NewsEvents/Pages/FreedomExpressionReport.aspx (accessed 17 June 2021).
\end{enumerate}
\end{footnotesize}
helm of such censorship, free expression experts note government action often explicitly or implicitly motivates such behaviour.\textsuperscript{52} Kaye argues that when the government has the power to define the scope of what content should be moderated, it has the capacity to abuse this privilege.\textsuperscript{53} Both states and law enforcement have been known to pressure platforms to remove content in an expeditious, and potentially unreasonable timeframe, sometimes resorting to coercion.\textsuperscript{54} As Sander posits, ‘such schemes not only incentivize platforms to sacrifice thoughtful deliberation in favour of speed but also circumvent the rule of law by enabling States to avoid seeking the removal of illegal content through formal legal avenues.’\textsuperscript{55}

Although many jurisdictions view platforms as ‘intermediaries, and confer them some level of protection and immunity for their decisions’,\textsuperscript{56} it is evident that the censorial practices they employ are an effort to remain in good standing with the government. Viewed in this light, content moderation can be characterised as a form of collateral censorship.\textsuperscript{57}

These practices have also served to silence marginalised populations who rely upon platforms as a tool of both communicating and documenting their experiences.\textsuperscript{58} While governments may have a vested interest in censoring dissident voices, this practice is only as successful as the platforms that facilitate it, and engagement with these platforms confers them even greater power. As Gillespie alludes to in the title of his book, they become the ‘custodians’ of the internet; the arbiters of speech. Platforms’ ability to control what information its users receive and believe poses a grave and insidious threat to expression.\textsuperscript{59}

Some free expression experts believe content moderation can fit neatly into the human rights framework of balancing rights against restrictions.\textsuperscript{60} In order for speech restrictions to be compliant with the global right to free

\textsuperscript{52} Supra note 3, para. 12.
\textsuperscript{53} Ibid., para. 13.
\textsuperscript{54} See, e.g., ibid., para. 16, referring to 24-hour takedown requirements imposed by certain states and para. 20, referring to state threats to ban the use of platforms provided they do not remove ‘offensive’ content.
\textsuperscript{55} Supra note 4, p. 960.
\textsuperscript{56} See, e.g., Section 230 of the Communications Decency Act which immunizes platforms from being classified as “publishers” for speech purposes, thereby indemnifying them. This is particularly relevant considering all major platforms are headquartered in the United States.
\textsuperscript{57} See, e.g., J. Balkin, ‘Old-School/New-School Speech Regulation’, 127 Harvard Law Review (2014) 2296–2342, pp. 2298, 2324.
\textsuperscript{58} Supra note 26, pp. 6–7.
\textsuperscript{59} Supra note 4, pp. 941–942.
\textsuperscript{60} See, e.g., M. Lwin, ‘Applying International Human Rights Law for Use by Facebook,’ 4 Yale Journal on Regulation Online Bulletin (2020) 53–85.
expression—as conferred by Article 19 of the International Covenant on Civil and Political Rights—they must be prescribed by law, have a legitimate aim, and be necessary (or proportionate). Kaye argues that platforms’ moderation policies are not compliant with Article 19 for a number of reasons, and that they should alter the manner in which they moderate content to ‘ensure content-related actions will be guided by the same standards of legality, necessity, and legitimacy that bind State regulation of expression.’

While important, compliance with Article 19 will bear little on the content that is valuable to international criminal investigators. Indeed, because that content can be hateful, inciteful, and is often at a minimum unpleasant to consume, it is likely that platforms would be justified in removing it. However, removing content without recourse may be construed as ‘unnecessary,’ that is, a disproportionate response given the lack of safeguards and procedural rights conferred to users. Pushing for a more fulsome and accessible appeals process would ameliorate some of the challenges faced by investigators.

### 2.3.2 Due Process Problem

There is a nexus between speech safeguards and procedural safeguards, and viewing content moderation as a speech problem may help address some of the procedural concerns that plague platforms. According to Roberts, platforms operate with a ‘logic of opacity’—the less the public knows about how moderation works, the more efficiently platforms run, and the better they can maintain the semblance of objectivity. This lack of transparency is pervasive—it extends beyond labour practices and into algorithmic decision making. It also affects avenues of recourse, as it is virtually impossible to mount a challenge against a decision when no one knows why content is removed. Furthermore, the limited options for appeal exacerbates a pre-existing access problem. While law enforcement has the power of subpoena, individuals lack such standing, and third parties even more so. Indeed, only resourced individuals may have the power to appeal, and such an imbalance translates to favourable treatment for those with influence. Viewing the due-process concern as

---

61 Article 19, ICCPR. See also supra note 3, paras 6–7.
62 Ibid., para. 45.
63 A.V. Banchik, ‘Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights-related content,’ 23 New Media & Society (2020) 1527n1544, doi.org/10.1177/1461444820912724.
64 Supra note 4, pp. 959–960.
65 Supra note 2, p. 2.
66 Supra note 4, p. 961.
two-fold—both a transparency and an access to justice problem—helps bolster the claim that content moderation runs afoul of human rights.

2.3.3 Memory Problem

An oft overlooked virtue of social media platforms is their ability to serve as an archive. Users possess the capability to document, share, and revel in their past and present activities. While much of this content remains unproblematic and preserved, human-rights related content runs a greater risk of being taken down.\textsuperscript{67} Documentation of human rights abuses on social media is not only often the most potent, visceral account of what happened but sometimes the only one that exists.

Deleting such content has the effect of redacting one’s chronicle of events and ‘underscore[s] the shortcomings of these platforms as persistent archives of historical events, particularly for eyewitness media documenting war and conflict.’\textsuperscript{68} Viewing media as a substitute for memory, content moderation poses a grave threat to our ability to remember atrocities.\textsuperscript{69} This not only represents platforms not living up to their potential to serve as archives of such history, but also impedes upon the ability to process and discuss global events. This is especially troubling when one considers the individuals who have potentially risked their lives to expose the world to what they have experienced.\textsuperscript{70} This, coupled with the proliferation of faked, or manipulated, content obscures our ability to create a record.

All three issues are real and pressing. However, they are broad-based and as a result obscure the day-to-day benefits and drawbacks content moderation confers to its users. Content moderation itself is not all bad. When levied in a cumbersome and sweeping manner, it runs afoul of human rights, but it also serves to limit the spread of decidedly unprotected speech, speech that incites violence. Indeed, there is no alternative to a hasty and binding process when content is uploaded at such a breakneck pace.

There is a chasm between the practical concerns of investigators regarding preservation and use of content and the theoretical concerns of these critics. There is content that exists that does not deserve unencumbered speech protection; content that due to its nature may be better off obscured from public view. Herein lies the problem of content that is bad speech but good evidence. Bolstering individual posters’ speech rights will not have an immediate impact

\textsuperscript{67} Supra note 26, p. 7.
\textsuperscript{68} Supra note 63, p. 3.
\textsuperscript{69} See, e.g., ibid., p. 13.
\textsuperscript{70} Supra note 26, p. 7.
in this space; nor will narrowing the scope of what gets taken down—or relying on human moderators,71 given the potential probative value of the most problematic content. Increasing transparency and due process rights only will help to the extent investigators exercise and engage with these rights. And memory, although important, is ancillary to probative weight in the context of a criminal proceeding.

3 Digital Evidence: The Open-Source Approach

3.1 Open-Source Investigations: An Introduction

Open-source investigations have become increasingly prevalent in the international criminal space.72 They are defined as ‘the process of identifying, collecting, and/or analysing open-source information as part of an investigative process.’73 Open-source information, is ‘publicly available information that anyone can obtain by request, purchase, or observation.’74 Online open-source investigation is a subset that pertains to employing these processes online.75 These terms encompass information derived from social media, a practice that is becoming ‘more and more important in international criminal and human rights investigations.’76

Open-source investigations (‘osint’) arose from the work of journalists, who for the past decade have used these techniques to monitor evolving situations around the globe. Indeed, ‘from journalism, these investigative practices and workflows migrated to the field of human rights: to NGO reporting and advocacy, fact-finding commissions, and criminal prosecutions…Today, many human rights institutions have either developed their internal capacity to leverage eyewitness media or outsourced this task.’77

Open-source information comprises both ‘user-generated’ and ‘machine-generated’ data, including metadata.78 Investigators are trained to observe, capture, and preserve all relevant information—a nuanced process

---

71 Supra note 63, p. 7.
72 Berkeley Protocol, supra note 2, para. 2.
73 A. Koenig, S. Dubberley and D. Murray, Digital Witness: Open Source Information for Human Rights Investigation, Documentation, and Accountability (Oxford University Press, Oxford, 2020), p. 9.
74 Ibid.
75 Ibid.
76 Ibid., p. 51.
77 Supra note 63, p. 3.
78 Berkeley Protocol, supra note 2, para. 1.
that extends beyond simply taking a ‘screengrab.’ Indeed, information about the content may prove to be more valuable than the content itself.

Information scraped from platforms can be invaluable to both fact-finding and criminal proceedings. It can be used to ‘corroborate witness testimony and to confirm specific details about an incident including the exact time and location, identities of the perpetrators, and how the crimes were carried out or their aftermaths.’79 In fact, OSINT can circumvent the need for state cooperation or law enforcement muscle to conduct on-the-ground investigation. It also limits reliance on ‘cumbersome bureaucratic procedures’ endemic in cross-border evidence collection.80

Social media provides fertile ground for this emerging investigative method. Notably, its ubiquity and volume of content renders it incredibly useful. Platforms provide a repository free of charge for documentation of events, and smartphones ‘created the possibility of user-generated evidence...being produced on a mass scale.’81 This development was nicely encapsulated by one Google employee referring to the Syrian civil war of 2011 as ‘the first YouTube conflict in the same way that Vietnam was the first television conflict.’82 In other words, despite their drawbacks, platforms such as Facebook, YouTube and Twitter are where the content is. As a result, ‘practitioners have little choice to resort to the dominant platforms and tolerate their limitations.’83

3.2 From Memory to Evidence
In order for information obtained through the course of an open-source investigation to be utilised to the fullest extent in court, it must be admitted as evidence. Indeed, ultimately what distinguishes ‘information’ from ‘evidence’ is the latter’s ‘evidentiary value that may be admitted in order to establish facts in legal proceedings.’84 A good deal is required to render digital evidence admissible,85 in spite of its increasing ubiquity, a phenomena some have termed ‘the coming storm.’86 The obstacles to admitting content found on social media are

79 HRW, supra note 2, p. 13.
80 Berkeley Protocol, supra note 2, para. 8.
81 R. Hamilton, ‘Social Media Platforms in International Criminal Investigations,’ 52 Case Western Research Journal of International Law (2020) 213–223, p. 216.
82 Ibid.
83 Supra note 63, p. 4.
84 Berkeley Protocol, supra note 2, para. 21. While somewhat circular, ‘it is important not to misuse or overuse the term ‘evidence’ when referring to ‘information’ generally.’
85 See e.g.: Rome Statute of the International Criminal Court, Article 69, which governs the admissibility of evidence generally.
86 S. Helmeci, ‘The Coming Storm: Possibilities for Preserving Evidence of War Crimes Published on Major Online Service Providers,’ ICC Forum (31 May 2020), available online at iccforum.com/forum/permalink/122/33563 (accessed 17 June 2021).
warranted, to be sure, but also reflect a structural impediment for courts and tribunals to swiftly adapt to evolving technologies.

For open-source investigators, the path to admissibility is a thorny one, rendered even more complicated by content moderation. This is perhaps best captured in the evolving practices of the ICC, where evidentiary determinations are made with ample judicial discretion. Bound by Rules of Procedure and Evidence, Judges are instructed to admit or exclude evidence based on its relevance, probative weight, and potential prejudice, all weighed in a holistic manner at the close of trial.87

Much is required on the part of investigators as they must preserve and maintain thorough records of their findings to have them admitted in court. These obstacles, among others culminated in the recent release of the Berkeley Protocol on Open-Source Investigations. The Protocol endeavours to standardise practices and further legitimise OSINT by introducing ‘common standards for capturing, preserving and analysing open-source information that can be used as evidence in criminal trials.’88 Sound methodology when capturing and preserving open-source information will more likely result in a finding of admissibility, and a conferral of probative weight.89

Irrespective of this development, there are obstacles to admissibility that loom large. For instance, lots of valuable content gets uploaded by anonymous users, a practice which has been referred to as ‘highly problematic’90 by the ICC. There are also questions about authenticity91 and reliability,92 as well as acceptance of open-source investigations as a legitimate means of evidence gathering. This distrust of OSINT is understandable but can present a roadblock—one that shifts the burden onto investigators and researchers to document and preserve, even as content disappears.

3.3 Content Moderation in the Open-Source Arena

Although content moderation has catastrophic impacts on open-source intelligence gathering, its covert manner of operating has tended to elude investigators.

In fact, most investigators only seem to understand the urgency of the issue—and the ‘content’s ephemerality through first-hand experience of losing

87 Rome Statute of the International Criminal Court, Article 69.
88 Berkeley Protocol, supra note 2, v.
89 Ibid., para. 55.
90 ICC, Prosecutor v. Gbagbo, Case No. ICC-02/11-01/11, Decision Adjourning the Hearing on the Confirmation of Charges, 3 June 2013, para. 29.
91 Supra note 73, p. 65.
92 Ibid.
access to removed content.’ In spite of this, it seems the concern is now widespread among investigators. For example, the first compendium dedicated to open-source investigations referred to content removal in these terms:

This impermanence of access to bits and bytes—the fact that the digital open-source universe is not necessarily an ever-accumulating thing, that today’s open source can be tomorrow’s closed source—is perhaps the most significant challenge facing open-source human rights investigators.

This sobering view of content moderation epitomises the challenge posed by relying on social media platforms to conduct investigations. Indeed, ‘beyond the question of how to handle the evidence is the challenge of obtaining it in the first place.’

Content moderation’s impact on open-source information is undeniable. None of these steps to render such information admissible as evidence can be achieved if the content disappears. ‘The implications for human rights investigations are obvious: if the original materials or documentation used in an investigation are no longer accessible, the credibility of any-fact finding is contingent on the word of the investigator or some later or lesser version of, or testament to, the content in question.’ Furthermore, there’s an understanding among investigators and practitioners that this content can be odious enough to warrant removal. The issue with removing content, however, is compromising its potential evidentiary value. Unfortunately, ‘social media companies can, and do, remove content with little regard for its evidentiary value.’ In fact, platforms are well aware of what they are doing, but ‘one person’s extremist propaganda is another person’s war-crime evidence.’ Banchik has observed investigators and practitioners letting platforms off the hook, viewing the ‘topography of takedowns as far more complex.’ However, given the ‘logic of opacity’ articulated in Part 1, they know little about how removals work and what may be precipitating them.

93 Supra note 63, p. 5.
94 Supra note 73, p. 95.
95 Supra note 42.
96 Supra note 73, p. 96.
97 Supra note 42.
98 Ibid.
99 Supra note 63, p. 7.
100 Ibid., p. 9.
They also seem to misunderstand or downplay the consequences of an active moderation regime. By viewing moderation practices as a black box beyond their control, investigators position themselves between a rock and a hard place regarding content. Content removal makes it increasingly more difficult to access, and preserve information, in a manner palatable for criminal proceedings. Investigators appear to cope with this by establishing their own standards for authenticating and preserving content. Academics, in turn, focus their energies on the challenges of working with the evidence—authenticating, documenting, and sifting through the sheer volumes of it.101

It is argued that focusing on these sets of challenges presupposes the existence and accessibility of open-source information and perpetuates the myth that content moderation bears no discernible weight on osint. By contrast, these measures render the work of investigators much more difficult and acutely impact progress. This impact is felt in three major ways—‘actual removal’ of content (occurring after an investigator has discovered it); ‘anticipated removal’ of content (occurring before an investigator—or any user—has seen it); and through perpetuating a power imbalance between resourced and ill-resourced investigative operations, and between civil society and law enforcement.

3.3.1 Actual Removal
Actual removal is probably the most palatable and visceral effect of content moderation. According to empirical studies, this has investigators feeling lucky when they do discover content in time, and ‘irresponsible’ or ‘stupid’ for not preserving it themselves.102 This has resulted in practitioners over-preserving content, ‘snagging a copy’ whenever possible as ‘you can’t trust others to ensure its existence.’103

There are a handful of ways to preserve content once it has been viewed. The simplest is to store the link or create a permalink. Another means is taking a screengrab, although this will not capture all properties of the content including its metadata. Finally, there is the option of downloading the content, one which many investigators employ. However, a caveat to this approach is that it expressly violates all platforms’ terms of service, amounting to a breach of contract.104 This seems to deter very few investigators, including one who proclaimed ‘we don’t have any other choice.’105

---

101 Supra note 81, p. 218.
102 Supra note 63, p. 6.
103 Ibid., p. 9.
104 Berkeley Protocol, supra note 2, para. 65.
105 Supra note 63, p. 9.
While violating the Terms of Service does not serve as a strong deterrent, or dilute the content’s admissibility, it makes downloading content a more foreboding technical accomplishment. Indeed, the process of preserving content is a costly and time-consuming one, particularly because it requires resources to store. Some practitioners deal with this by gauging how likely the content in question will be removed before moving forward, others temporarily save copies and ultimately reupload them to the same platforms, viewing it as ‘a public service’.

None of these solutions are enviable. They are an exercise in prioritisation, and hinge upon the judgment of the investigator and the savviness of his or her operation. ‘Investigators must prioritize their efforts in discovery and preservation of open-source materials based on a mix of relevance to key questions of fact, the relative abundance (or paucity) of materials that address those questions, and some measure of the risk to the accessibility of the evidence.’

3.3.2 Anticipated Removal

Perhaps a bigger issue is the prophylactic removal of content. Most content is removed via automation—either through automated flagging, hashing, or some other means. ‘Where the loss of information is precipitated by algorithms, however, it is unlikely that the material can be discovered and secured before it is removed from the open. Indeed, as suggested by the YouTube transparency report, most materials that are ultimately removed will never be seen by anyone on the platform.’ In other words, most content is removed before investigators even have the opportunity to act on it.

The staggering impact ex-ante removal has on digital evidence gathering is perhaps best illustrated with an analogy. Imagine a third party had the power to be the first to an active crime scene, arriving before law enforcement, and investigators for both parties. If, in this window of time, it removed certain items at random—potential pieces of physical or documentary evidence—such removal could dramatically alter the scope of the investigation. Some of the items might be crucial to the investigation, others may be relatively useless. Regardless, both prosecution and defence would be flying blind, operating on the assumption nothing is missing, not knowing what they do not know.

---

106 Ibid., pp. 9–10.
107 Berkeley Protocol, supra note 2, para. 159.
108 Supra note 63, p. 11.
109 Supra note 73, p. 97.
110 Ibid.
Such a hypothetical would likely result in a severe miscarriage of justice. Making such a mental leap into the open-source context at a minimum casts doubt on open-source investigations as a form of evidence gathering. After all, how efficacious can such a method be if what remains for human consumption is just a sliver of what exists? Unlike physical crime scenes, which we endeavour to protect from interference, no safeguards exist for ex-ante content removal. One ICC employee lamented: ‘It’s something that keeps me awake at night… the idea that there’s a video or photo out there that I could use, but before we identify it or preserve it, it disappears.’

3.3.3 Power Imbalance
Finally, content moderation exacerbates a power imbalance endemic in criminal investigations. This imbalance is captured in two ways—both in the avenues of appealing a takedown and regaining access to the expunged content, and in the ability to preserve content and exert influence over platforms.

When it comes to reversing a takedown, the surest bet is issuing a subpoena. This power is conferred exclusively to domestic law enforcement agencies, rendering international criminal tribunals reliant on state cooperation. International investigators, including those working for the prosecution, ‘lack law enforcement powers and standing.’ This is in part because all major platforms are operated in the US and are therefore required to abide by the Stored Communications Act (‘SCA’) which prohibits turning over content to law enforcement in the absence of a subpoena. The SCA is silent regarding foreign law enforcement subpoena power. European courts, as a result, have to rely on mutual legal assistance treaties which are slow and cumbersome. This is even more problematic in the international criminal space due to the American Servicemembers Protection Act which expressly forbids platforms from turning over evidence to the ICC, or otherwise cooperating in an investigation. Furthermore, such practices are ‘premised on authorities knowing about content that was taken offline…[and] do not address’ content that disappears without a trace.

For those that lack subpoena power—however insufficient it may be—the only option is to file an appeal. As aforementioned, such processes are so
byzantine that many decline to pursue them.118 This remedy restores the content for everyone, unlike subpoena power which in effect confers access exclusively to the requesting party. Some NGO’s have been quite successful in lodging appeals, but according to Banchik ‘practitioners were unequally able to leverage connections with employees at platform companies.’119 Furthermore, ‘cynicism and confusion about the appeals process’ prevents many from appealing.120

Another imbalance perpetuated by content moderation is the race to preserve expunged content. Given the sheer cost and manpower required to self-preserve content, more resourced operations have a much easier time doing so. Preservation-related expenses are incurred by the investigators themselves, and many operate either independently or for non-profits. Ultimately, ‘[w]e have huge equity concerns: What stories are we losing? Whose voices are we not hearing? Who’s in a dire situation who we don’t know about?’121

In spite of all these undeniable adverse impacts, there is much hope and optimism surrounding open-source investigations. Perhaps in an effort to legitimise such a mode of evidence-gathering, investigators and criminal lawyers are disinclined to acknowledge or confront the threat of evidence disappearing. However, sidestepping or downplaying the issue does not make it disappear. In fact, it will likely exacerbate the problem and deepen pre-existing inequities.

4 A Fused Approach

4.1 Core Features

Both aforementioned attitudes toward content moderation are incomplete. Where one may be lacking in practicability, the other is lacking in administra-bility. However, both have merit. Fusing these views would allow practitioners to graft from each, harnessing their strengths and mitigating their weaknesses. It is argued there are three core features of a fused approach: a nuanced understanding of the issue, successful deployment of advocacy, and resources specifically allocated to offsetting the adverse impacts of content moderation. While these features are not easily to implement, I believe adopting them will allow

118 HRW, supra note 2, p. 31.
119 Supra note 63, p. 2.
120 Ibid., p. 12.
121 Supra note 42.
both investigators and human rights practitioners to erode the moderation scheme and effect change from the ground up.

4.1.1 Nuance
The first step anyone invested in the content moderation issue must take is a step back. One must recognise that not all content is created equal, and that it is neither desirable nor feasible to leave platforms unmoderated. Indeed, sometimes the most valuable evidence is not only unprotected speech but is unsuitable for human consumption. As one practitioner put it, ‘As a prosecutor in this field of law, I’m worried...but as a citizen, I’m a bit relieved.’

After recognising that content must be moderated in some capacity, it’s critical to examine the platforms themselves. They are not static tools, but rather dynamic actors in the ‘landscape of international criminal investigations.’ As Rebecca Hamilton argues, ‘digital evidence does not just bring a new form of evidence into the international criminal justice system, it brings in a host of new actors as well.’ These actors include platforms who operate ‘in a regulatory landscape that is in flux’ and ‘cannot be expected to prioritize the goals of justice and accountability in the face of business demands.’ Furthermore, the risk to evidence is not static and may dramatically increase over time as the result of the actions of third parties, or even as a result of the investigation itself....Even materials that are reasonably deemed to be at low risk of disappearing could unexpectedly become at high risk of loss or tampering, based on unpredictable developments or activity by external parties.

Although ‘[t]he temptation to look to simple solutions to the complex problem of extremism online is strong,’ acknowledging that content moderation is a dynamic, evolving issue with multiple stakeholders is a preliminary and necessary step to effecting change. As is often the case, an employing an interdisciplinary approach will fortify this understanding.

122 Supra note 42.
123 Supra note 81, p. 214.
124 Ibid., p. 217.
125 Ibid., p. 223.
126 Supra note 73, p. 97.
127 Supra note 26, p. 10.
4.1.2 Advocacy

Public pressure is often the best method of encouraging large entities such as platforms to implement changes. This is perhaps even more true in the content moderation space. An approach which directly challenges this status quo, through more traditional, confrontational means of activism, or through more cooperative means such as engaging with the platform to restore expunged content, seems to have reaped some success for the human rights community. One salient example was YouTube’s restoration of Syrian-related content in 2017.128

Platforms’ removal of content with potential evidentiary value is a pressing issue in a way the more general threats they pose to free speech is not. A targeted campaign against removal of content with evidentiary value could encapsulate some means employed by activists more generally, putting open-source investigations at the focus.

4.1.3 Resources

As previously mentioned, anyone wishing to challenge the content moderation regime needs resources to do so. These resources need not be purely financial, as labour and experience are also instrumental. A well-funded operation will be able to cover the costs of self-preservation as a means of mitigating the effect of takedowns.129 ‘Storage for video is expensive. Journalists, smaller organizations and activist groups often lack the technical resources to preserve eyewitness videos on their own.’130 At a baseline, anyone wishing to preserve their own content—particularly in a manner that renders it admissible in court—need invest in funding their own archive and hiring personnel to properly manage it. ‘To be blunt: running servers and operating a large-scale archive costs a considerable amount of money and effort and requires technical skills to maintain.’131

Human capital is also a resource. A larger staff of trained personnel will result in a more comprehensive monitoring of platforms and removed content.

---

128 A. Rosen, ‘Erasing History: YouTube’s Deletion of Syria War Videos Concerns Human Rights Groups,’ Fast Company (7 March 2018), available online at www.fastcompany.com/4054041/erasing-history-youtubes-deletion-of-syria-war-videos-concerns-human-rights-groups (accessed 17 June 2021): ‘The subsequent outcry led to media coverage in the New York Times and the Intercept. Many videos…were eventually restored, and the pace of removals slowed.’

129 See supra note 73, p. 98.

130 S. Ristovska, ‘Human Rights Collectives as Visual Experts: The Case of Syrian Archive’, 18 Visual Communication (2019) 333–351, at 349.

131 Supra note 73, p. 173.
If there is no one employed to take this issue seriously, it will be much easier for expunged content to fall to the wayside, rendering the work of investigators less effective. These labour-related resources extend beyond investigators to archivists and legal personnel who can assist in rendering any found content as useful as possible. Having experienced staff on hand—or at a minimum, personnel willing to be trained—will be an invaluable resource in navigating the complex regime of takedowns.

4.2 Case Study: The Syrian Archive
When examining potential implementations of these core features, one need look no further than the Syrian Archive. The Syrian Archive was created in 2014 by Hadi Al Khatib. He conceived of the initiative as a means of ‘creating an evidence-based tool for reporting, advocacy, and accountability purposes.’ It was started by an interdisciplinary team of technologists, journalists, and lawyers. The group’s focus was the Syrian conflict which began in 2011 and has ‘more hours of user-generated digital content about the conflict than there have been hours of the conflict itself.’ Indeed, it was the pressing ‘need to securely preserve the online content coming from the Syrian conflict’ that motivated Al Khatib to found Syrian Archive. The organisation opted to service these goals by archiving content ‘from thousands of social media channels and accounts—images, videos and posts that are both invaluable historical artifacts and potential evidence of human rights abuses.’

The Syrian Archive employs a sophisticated yet intuitive methodology. This ‘organized, secure and open-source repository’ focuses on several pressure points: identification, collection and secure preservation, processing, review, and publication. It does so in a completely transparent manner, rendering it more palatable to a court or tribunal. Although the details of this workflow are beyond the purview of this article, the Archive generally starts with

132 J. Deutch and H. Habal, ‘The Syrian Archive: A Methodological Case Study of Open-Source Investigation of State Crime Using Video Evidence from Social Media Platforms,’ 7 State Crime Journal (2018) 46–76, p. 48.
133 Ibid., p. 52.
134 ‘Lost and Found,’ Syrian Archive, available online at syrianarchive.org/en/lost-found (accessed 2 April 2021).
135 The term ‘open-source’ employed in this context refers to technology whose code is made publicly available and can be modified. See also supra note 73, p. 172. ‘For Syrian Archive, because all stages of its archival work are open source, it is possible to show how the data was discovered, how it was acquired, and the process by which data has been transformed, processed, and analysed.’
136 Supra note 132, p. 53.
137 Supra note 73, p. 173.
prophylactically identifying content that may be pertinent by ‘scraping’ it from platforms based on whether it comes from a credible source.\textsuperscript{138} This content (and its metadata) is then stored on servers throughout the world. The metadata is then extracted and processed, the content is verified, and ultimately published on Syrian Archive’s website, accessible to the general public.\textsuperscript{139} Currently, there are ‘1.5 million data points, over twenty terabytes of video and image data’\textsuperscript{140} available for view on the Syrian Archive.

Much of this methodology evolved over time. In particular, the work of the Syrian Archive was put to the test in 2017. YouTube unveiled a new algorithm that engaged in a priority system—allowing ‘extremist’ content to skip the queue and be removed by automated processes rather than a human moderator.\textsuperscript{141} This resulted in hundreds of thousands of videos documenting the Syrian conflict being removed in one fell swoop.\textsuperscript{142} For open-source investigators, this was a massive setback.\textsuperscript{143} However, it was then that the Archive’s ‘parallel evidence locker’\textsuperscript{144} became even more critical. In fact, Syrian Archive was able to work with YouTube to ultimately restore expunged content. The Syrian Archive’s model also extended to other conflict zones, as its parent company Mnemonic (also founded in 2017) began spearheading archiving initiatives for Sudan and Yemen as well.\textsuperscript{145}

Although these initiatives are primarily archives, preservation of expunged content is only a part of their mission. For instance, Syrian Archive has ‘led the field in discussions on content moderation’ as ‘one of the only groups worldwide who has quantitative data on the real impact of content moderation policies.’\textsuperscript{146} Indeed, the Archive publishes a monthly takedown report that tracks the quantitative and qualitative patterns of content removal.\textsuperscript{147} It also is dedicated to reforming the system through long-term advocacy by publishing

\begin{itemize}
\item \textsuperscript{138} Supra note 132, p. 58.
\item \textsuperscript{139} ‘The Archive,’ Syrian Archive, available online at syrianarchive.org/en/data-archive (accessed 2 April 2021).
\item \textsuperscript{140} Supra note 73, p. 165.
\item \textsuperscript{141} M. Browne, ‘YouTube Removes Videos Showing Atrocities in Syria,’ New York Times (22 August 2017), available online at www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html (accessed 17 June 2021).
\item \textsuperscript{142} Supra note 12, at 91.
\item \textsuperscript{143} See supra note 141, ‘Things just got a lot harder now that the videos that were on YouTube are no longer around.’
\item \textsuperscript{144} Supra note 42.
\item \textsuperscript{145} ‘About Us,’ Mnemonic, available online at mnemonic.org/en/about (accessed 2 April 2021).
\item \textsuperscript{146} Supra note 134.
\item \textsuperscript{147} See, e.g., ‘Removals of Syrian Human Rights Content: May 2020,’ Syrian Archive, available online at syrianarchive.org/en/lost-found/june20-takedowns (accessed 17 June 2021).
\end{itemize}
scholarship, working with platforms, and training professionals in how to use open-source investigative methods. The Archive has even participated in Gifct and the Christchurch Call Advisory Network.

In its purest form, the Archive serves as an alternative to platforms for those searching for potential evidence of war crimes in Syria. This curation of content valuable to human rights monitors and investigators not only preserves it, but brings the content moderation issue to light and pressures platforms to do better. Indeed, ‘by collecting and verifying visual evidence, archiving and preserving it, conducting visual investigations and training, Syrian Archive is helping promote best practices and standards for rendering eyewitness video meaningful in institutional and legal contexts for human rights purposes.’

This aim is buttressed by a nuanced understanding of the content moderation problem. Syrian Archive recognises the virtues of open-source investigation, and that such methods are valuable in courtrooms and beyond. However, rather than seeking to side-step the content moderation issue, it addresses it head on—producing reports of takedowns and confronting hard-learned truths about the even greater risk automation poses to the moderation regime. It seeks to offset the adverse impacts of content moderation while simultaneously studying them and noting their existence. Most importantly, however, by positioning itself as an intermediary, a developed understanding of content moderation is a precondition to the Archive’s success. It endeavours to preserve—and make available—videos it sees as valuable and probative, engaging a demographic of journalists and human rights practitioners rather than impressionable citizenry.

The Syrian Archive also takes its role as an advocate in this space seriously. It engages with platforms, with the media, and with NGOs in the publication of pertinent scholarship. In addition to engaging with such venerable

---

148 ‘Home,’ Syrian Archive, available online at https://syrianarchive.org/ (accessed 17 June 2021): ‘We...train others and work with platforms to reduce the impact of their harmful policies.’

149 ‘Advisory Network,’ Christchurch Call Advisory Network, available online at www.christchurchcall.com/advisory-network.html (accessed 17 June 2021).

150 Supra note 130, p. 19–20 (emphasis added).

151 Supra note 29.

152 B. Perrigo, ‘These Tech Companies Managed to Eradicate ISIS Content. But They’re Also Erasing Crucial Evidence of War Crimes,’ Time (11 April 2020), available online at time.com/5798001/facebook-youtube-algorithms-extremism/ (accessed 17 June 2021).

153 H. Al Khatib and D. Kayyali, ‘YouTube is Erasing History,’ New York Times (23 October 2019), available online at www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html (accessed 17 June 2021).

154 See, e.g., supra note 26.
institutions, the Archive also cultivates a strong grassroots presence. It seeks to train individuals in open-source investigative methods, and advocate for individuals whose content has been expunged.\textsuperscript{155}

Finally, although the Syrian Archive is a non-profit institution, it has levied its resources in an admirable way, putting the content moderation issue front and centre. With a small number of full-time employees,\textsuperscript{156} and a meagre budget it has managed to create a secure and robust archive of digital evidence. Its means of preserving content are methodologically sound and compliant with the Berkeley Protocol. Indeed, the Syrian Archive predates the release of the Protocol and was influential in its development. It also expends time and energy understanding and analysing takedowns, tapping into a small network of experienced personnel.

The Archive’s credibility cannot be understated. Despite its small size, the impressive nature of its operation has catapulted it onto the global stage, rendering it a ‘pioneer.’\textsuperscript{157} By expending its resources to help bring the content moderation issue to light, the Archive helps in ‘ensuring that eyewitness videos have a higher likelihood of playing evidentiary and forensic roles whether in UN investigations or possible ICC prosecutions.’\textsuperscript{158}

5 Conclusion

Content removal poses a risk to human rights practitioners, lawyers, journalists, and anyone with an interest in maintaining our collective memory of atrocities. In fact, the risk it poses is growing with time, as we rely more heavily on social media platforms, and those platforms in turn rely on automation to make decisions. More content means more artificial intelligence, and increasing regulatory pressures will cause platforms to prophylactically and sweepingly remove content. Open-source investigators need to be on notice of this phenomenon, acknowledging that the platforms they rely on for evidence gathering are dynamic actors with a vested interest of their own. As Rebecca Hamilton astutely states: ‘relying on digital evidence also means relying on the platforms who host it.’\textsuperscript{159}

\begin{thebibliography}{9}
\bibitem{155} \textit{Supra} note 134: ‘Has one of your posts been removed? Get in touch.’
\bibitem{156} \textit{Supra} note 42.
\bibitem{157} \textit{Supra} note 130, p. 22.
\bibitem{158} \textit{Ibid.}, p. 23.
\bibitem{159} \textit{Supra} note 81, p. 223.
\end{thebibliography}
Preparing for the ‘coming storm’ means understanding that platforms have vastly different objectives, and will continue to operate in furtherance of these objectives until they are pressured to do otherwise.

In fact, it is argued that noticing the threat content moderation poses to the work of open-source investigations may in fact motivate investigators, and ultimately courts and tribunals, to act on this issue in a meaningful way. Turning a blind eye to the problem allows it to precipitate, casting an ominous shadow over conversations about admissibility and the legitimacy of this form of fact-finding. Working with each other to understand the scope and severity of the problem, and fostering collaborative relations and practices between institutional agents, could serve to cut against this uncertainty.

In this respect, proponents of open-source investigations may have something to learn from human rights and free expression practitioners who have directly confronted the content moderation issue. By taking up the issue, they are influencing the broader debate and conducting it on their own terms.

Indeed, this group would stand to benefit from giving voice to the concerns of investigators. The content moderation issue has largely been framed as a conceptual, abstract one. Hand-wringing and allegations of censorship also reflect some level of wishful thinking; that platforms will respond to such broad-based concerns. By contrast, practical issues are much harder to ignore. The problem of evidence of genocide or war crimes disappearing is a palpable one. It directly implicates platforms in the calculus and may even render them liable for obstructing justice.

Content moderation is more than just a speech problem and invoking exclusively speech-based concerns does not neatly comport with the scale of operations. As platforms have repeatedly stated, there is too much content to moderate it carefully. In fact, platforms can easily turn the speech argument on its head—*not* removing content is their way of sanctioning free speech, and removing (or curating) content is a way of making speech better and facilitating an open discourse of ideas.

A comprehensive approach to content removal will optimise the work of investigators by making them better and more careful at their job, give human rights and free speech experts a palatable and cogent example of moderation’s adverse impact that they can harness in support of their cause. Most importantly, however, this approach may inspire users to continue documenting and uploading their lived experiences knowing there is a community which supports and relies upon such content in its mission of bringing perpetrators to justice.