Open Source Information's Blind Spot

Human and Machine Bias in International Criminal Investigations

Yvonne McDermott,*Alexa Koenig** and Daragh Murray***

Abstract

Digital open source information has been heralded for its democratizing potential, insofar as it allows access to a much broader range of sources and voices than would normally be consulted through traditional methods of information gathering for international criminal investigations. It also helps to overcome some of the physical access barriers that are commonplace in international criminal investigations. At a time when the use of digital open source information is becoming more widespread, this article warns of the cognitive and technical biases that can impact upon two key stages of an investigation: finding relevant information and analysing that information. At the information-gathering stage, there are particular crimes, regions and groups of people whose experiences are more likely to be overlooked or hidden in digital open source investigations. When it comes to analysing digital open source information, there is a danger that cognitive and technical biases may influence which information is deemed most relevant and useful to an international criminal investigation, and how that information is interpreted. This article proposes some steps that can be taken to mitigate these risks.

* Professor of Law, Swansea University, UK. [Yvonne.McDermottRees@swansea.ac.uk]

** Executive Director and Lecturer-in-Residence, Human Rights Center, University of California, Berkeley, USA. [kalexam@berkeley.edu]

*** Senior Lecturer, Human Rights Centre and School of Law, University of Essex, Colchester, UK. [d.murray@essex.ac.uk]

This work was supported by the United Kingdom Economic and Social Research Council (grant number ES/R00899X/1) as part of the *OSR4Rights* project. The authors thank Sam Dubberley, Emma Irving, Nema Milaninia, and the anonymous reviewers for their feedback on an earlier draft.

Journal of International Criminal Justice (2021), 1 of 21 doi:10.1093/jicj/mqab006

1. Introduction

Digital open source information[1] — defined as information on the internet that any member of the public can obtain by request, purchase or observation[2] — is revolutionizing the investigation and prosecution of international crimes.[3] Nowhere is this more apparent than before the International Criminal Court (ICC), where the Prosecutor has introduced new forms of digital open source evidence in a number of cases.[4] Satellite imagery, videos and photographs have been used to identify particular sites in Mali,[5] while social media evidence has been introduced to show direct evidence of alleged violations,[6] and to demonstrate a claimed relationship between key individuals.[7] Open source information also plays an increasingly important role in the work of human rights fact-finding missions, commissions of inquiry and other forms of investigation.[8]

Researchers have identified three key advantages to using digital open source information for investigations. First, where direct access to crime sites

1 In this article, the terms open source information and open source evidence are distinguished, with 'evidence' solely referring to materials submitted as part of a criminal process. 'Information' covers all other materials.

2 Human Rights Center, University of California, Berkeley/UN Office of the High Commissioner for Human Rights, *Berkeley Protocol on Open Source Investigations* ('Berkeley Protocol'), 1 December 2020, at 6–7; S. Dubberley, A. Koenig and D. Murray, 'Introduction: The Emergence of Digital Witness', in S. Dubberley, A. Koenig and D. Murray (eds), *Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation and Accountability* (Oxford University Press, 2020) 3, at 9.

3 M. Aksenova, M. Bergsmo and C. Stahn, 'Non-Criminal Justice Fact-Work in the Age of Accountability', in M. Bergsmo and C. Stahn (eds), *Quality Control in Fact-Finding* (2nd edn., TOAEP, 2020) 1, at 9–10; E. Irving, 'And So It Begins ... Social Media Evidence in an ICC Arrest Warrant', *Opinio Juris*, 17 August 2017; A. Koenig et al., 'Open Source Fact-Finding in Preliminary Examinations', in M. Bergsmo and C. Stahn (eds), *Quality Control in Preliminary Examination: Volume 2* (TOAEP, 2018) 681.

4 L. Freeman, 'Prosecuting Atrocity Crimes with Open Source Evidence: Lessons from the International Criminal Court', in Dubberley, Koenig and Murray (eds), *supra* note 2, 48, at 52.

5 Transcript, *Al Hassan* (ICC-01/12-01/18-T-027-Red-ENG), Trial Chamber X, 21 September 2020; Judgment and Sentence, *Al Mahdi* (ICC-01/12-01/15-171), Trial Chamber VIII, 27 September 2016.

6 Warrant of Arrest, *Al-Werfalli* (ICC-01/11-01/17-2), Pre-Trial Chamber I, 15 August 2017, §§ 11–22; Second Warrant of Arrest, *Al-Werfalli* (ICC-01/11-01/17-13), Pre-Trial Chamber I, 5 July 2018, §§ 17–18.

7 Decision on 'Prosecution's Fifth Request for the Admission of Evidence from the Bar Table', *Bemba et al.* (ICC-01/05-01/13-1524), Trial Chamber VII, 14 December 2015. See further, Prosecution's Fifth Request for the Admission of Evidence from the Bar Table, *Bemba et al.* (ICC-01/05-01/13-1498), 30 November 2015, §§ 17–18.

8 See e.g. *Detailed Findings of the Independent International Fact-finding Mission on the Bolivarian Republic of Venezuela*, A/HRC/45/CRP.11, 15 September 2020; *Report of the Detailed Findings of the Independent International Commission of Inquiry on the Protests in the Occupied Palestinian Territory*, UN Doc. A/HRC/40/CRP.2, 18 March 2019; *Report of the Detailed Findings of the Group of Eminent International and Regional Experts on Yemen*, UN Doc. A/HRC/45/CRP.7, 29 September 2020; *Report of the Independent International Fact-finding Mission on Myanmar*, UN Doc. A/HRC/39/64, 18 September 2018.

has been denied or is impossible for security or logistical reasons, open source information can be an important source of lead, linkage, contextual and corroborating evidence.[9] The fact that this material can be obtained remotely also helps minimize the risk to witnesses: instead of asking an individual or individuals to testify to the relationship between alleged perpetrators, or the destruction of cultural property, or the alleged commission of crimes, for example, this may be shown through verifiable publicly available information.

The second major acknowledged advantage is open source information's democratizing potential, insofar as it provides an avenue through which ordinary people in affected regions can tell their stories and directly influence international fact-finding processes.[10] As Dyer and Ivens have noted, open source investigations 'can centre the experiences of groups whose voices are too often heavily mediated, marginalised or excluded in conventional reporting'.[11] As well as playing an important role in contemporary fact-finding processes,[12] citizen documentation of human rights violations online can also help to counter disinformation and denialist propaganda by all actors involved in a conflict, including powerful states.[13]

The third identified advantage to the use of digital open source information in investigating international crimes is its relative objectivity. For example, unlike a witness — and despite technical limitations — a satellite image cannot forget salient facts, misremember key details, or be motivated by self-interest or allegiance to a particular group. As such, it could be a means to mitigate against 'the frailties of human perceptions'[14] and the well-documented issue of witness credibility that have affected international criminal trials for decades.[15] In addition, open source information, particularly citizen-generated photo and video content posted online, may reduce the

9 F. Abrahams and D. Murray, 'Open Source Information: Part of the Puzzle', in Dubberley, Koenig and Murray (eds), *supra* note 2, 317; A. Koenig and L. Freeman, 'Open Source Investigations for Legal Accountability: Challenges and Best Practices', in Dubberley, Koenig and Murray (eds), *supra* note 2, 331, at 332–333.

10 UN Human Rights Council, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, A/HRC/29/37, 24 April 2015, § 39; M.K. Land, 'Democratizing Human Rights Fact-Finding', in P. Alston and S. Knuckey (eds), *The Transformation of Human Rights Fact-Finding* (Oxford University Press, 2016), 399, 402.

11 S. Dyer and G. Ivens, 'What would a Feminist Open Source Investigation look like?', 1 *Digital War* (2020), https://doi.org/10.1057/s42984-020-00008-9, 1.

12 S. Gregory, 'Ubiquitous Witnesses: Who Creates the Evidence and the Live(d) Experience of Human Rights Violations?' 18 *Information, Communication and Society* (2015) 1378; Irving, *supra* note 3.

13 A. Geis and G. Schlag, '"The Facts cannot be Denied": Legitimacy, War and the Use of Chemical Weapons in Syria', 7 *Global Discourse* (2017) 285; J. Deutch, 'Challenges in Codifying Events Within Large and Diverse Data Sets of Human Rights Documentation: Memory, Intent, and Bias', 14 *International Journal of Communication* (2020) 5055, at 5056.

14 Judgment, *Kupreškić et al.* (IT-95-16-A), Appeals Chamber, 23 October 2001, § 34.

15 N. Combs, *Fact-Finding without Facts: The Uncertain Evidentiary Foundations of International Criminal Convictions* (Cambridge University Press, 2010); S. De Smet, 'Justified Belief in the Unbelievable', in Bergsmo and Stahn (eds), *supra* note 3, 81, at 128–129; Abrahams and Murray, *supra* note 9, 324.

need for intermediaries, whose role has been problematic in trial practice,[16] insofar as open source materials can provide a direct link to witnesses and survivors.

The authors recognize the potentially significant advantages associated with using open source information, and believe that this form of information and evidence should play a key role in future investigations. However, as investigators increasingly turn to open source techniques, this article sounds a word of caution against overlooking a number of biases and blind spots that can hinder the utility of open source information in the investigation of international crimes. While it is true that open source information may be used to indicate the potential commission of international crimes or to show networks between key actors, to prove the culpability of leaders and/or the existence of a plan or policy, it is likely that in most cases witness testimony (particularly that of 'insider' witnesses) will still be needed. Open source information should neither be viewed as a panacea, nor in a vacuum. Furthermore, and as shall be shown in this article, while open source information has a clearly democratizing potential, there is a risk that a rush towards greater adoption of open source research methods in investigations may inadvertently silence some of the most marginalized populations. In this article, we argue that digital open source information can be as vulnerable to subjectivity and bias as any other form of evidence.

Before turning to our analysis of where and when bias may emerge in digital open source investigations, it is useful to define the term. While the ordinary meaning of 'bias' links to a prejudice or particular bent,[17] the scientific literature describes bias as, '[a]ny process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth'.[18] In the context of international criminal investigations, a bias can similarly be described as any systematic distortion or error, due to a design problem, an interfering factor, or a judgement, that can affect the conception, design, or conduct of an investigation, or the collection, analysis, interpretation, presentation, or discussion of the evidence.[19]

16 Redacted Decision on the Prosecution's Urgent Request for Variation of the Time-Limit to Disclose the Identity of Intermediary 143 or Alternatively to Stay Proceedings Pending Further Consultations with the VWU, *Lubanga* (ICC-01/04-01/06-2517-Red), Trial Chamber I, 8 July 2010; E. Haslam and R. Edwards, 'Managing a New ''Partnership'': ''Professionalization'', Intermediaries and the International Criminal Court', 24 *Criminal Law Forum* (2013) 49.

17 The *Oxford English Dictionary* defines 'bias' as, 'An inclination, leaning, tendency, bent; a preponderating disposition or propensity; predisposition *towards*; predilection; prejudice'.

18 D.L. Sackett, 'Bias in Analytic Research', 32 *Journal of Chronic Diseases* (1979) 51.

19 This definition is heavily influenced by the comprehensive definition of bias produced by Aronson following a systematic review of definitions of 'bias' that have previously been proposed in statistical, epidemiological, and sociological texts and an extraction of key themes from those definitions: J. Aronson, 'A Word about Evidence: 6. Bias – a proposed definition', *Catalogue of Bias*, 15 June 2018, available online at https://catalogofbias.org/2018/06/15/a-word-about-evidence-6-bias-a-proposed-definition/ (visited 2 November 2020). Aronson's definition is: 'A systematic distortion, due to a design problem, an interfering factor, or a judgement, that can affect the conception, design, or conduct of a study, or the collection, analysis, interpretation,

A considerable body of literature is dedicated to the hundreds of forms of bias that can impact upon almost every aspect of human life, from medicine,[20] to education,[21] to technology.[22] The University of Oxford's 'Catalogue of Bias', a project that maps and defines the main biases that impact on research, defines over 60 different biases.[23] For the purposes of this article, we broadly categorize biases into two overarching categories: (i) technical biases, which are biases linked to decisions made by computer systems, and (ii) cognitive biases, which are systematic errors in thinking or reasoning that impact upon human decision-making. Technical biases become relevant when using technology to conduct fact-finding and analysis, while cognitive biases relate to how people perceive and make sense of information.

In reality, these two broad categories are heavily intertwined, insofar as human biases feed into machine biases, and vice versa.[24] However, the most common biases can be loosely defined as being either technical or cognitive in nature. Within 'technical' bias, the following sub-categories are most relevant:

- *Access bias*: The bias around who has access to the technologies and tools needed for documenting events.[25]
- *Algorithmic bias*: The bias embedded in the design of algorithms and their use, often due to already-biased training data.[26] Algorithmic bias can impact what results users see when they conduct a search, and the order in which results are presented.

presentation, or discussion of outcome data, causing erroneous overestimation or underestimation of the probable size of an effect or association.'

20 Sackett, *supra* note 18; D. Chavalarias and J.P. Ioannidis, 'Science Mapping Analysis Characterizes 235 Biases in Biomedical Research', 63 *Journal of Clinical Epidemiology* (2010) 1205.

21 R.L. Linn and C.E. Werts, 'Considerations for Studies of Test Bias', 8 *Journal of Educational Measurement* (1971) 1; C. Jencks, 'Racial bias in testing', in C. Jencks and M. Phillips (eds), *The Black–White test score gap* (Brookings Institution Press, 1998), 55.

22 E.g. Z. Obermeyer et al., 'Dissecting Racial Bias in an Algorithm used to Manage the Health of Populations', 366 *Science* (2019) 447.

23 https://catalogofbias.org (visited 2 November 2020). See further, S. Tanveer, 'Catalogue of Bias: a resource review', 9 November 2019, available online at https://s4be.cochrane.org/blog/2019/05/09/catalogue-of-bias-a-resource-review/ (visited 2 November 2020).

24 Z. Rahman, 'Tech Bias, People Bias', *The Engine Room*, 11 December 2019, available online at https://www.theengineroom.org/tech-bias-people-bias/ (visited 2 November 2020) ('Technology's problems are not, and never have been, just about technology … problems arise not 'just' due to technical mistakes but because of very human decision making, whether that is a human trusting a machine over another human; poorly executed data analysis; or, often, technology being built to reinforce human prejudices').

25 Berkeley Protocol, *supra* note 2, 12.

26 E.g. M. Garcia, 'Racist in the Machine: The Disturbing Implication of Algorithmic Bias', 33 *World Policy Journal* (2016) 111; F. Zuiderveen Borgesius, *Discrimination, Artificial Intelligence, and Algorithmic Decision Making* (Council of Europe, 2018).

- *Machine bias*: The bias that arises from technical constraints or limitations in the design process or computer tools.[27] This might include, for example, where an attempt is made to encode nuanced human experiences or concepts into computer systems.[28]
- *Emergent bias*: Where the knowledge, values, or expertise of the users of a product or system are different to those assumed or prioritized by the designers of that product or system, creating difficulties in how the system is actually used.[29]

Cognitive bias refers to any distorted evaluation of information by humans.[30] The following forms of cognitive bias are most prevalent:

- *Anchoring*: The tendency to rely too heavily on an initial piece of information (or 'anchor') in later decisions in a way that causes the investigator to discount, overvalue or misinterpret later information.[31]
- *Automation bias*: The tendency to defer to suggestions made by automated decision-making systems, particularly in circumstances where a human decision-maker would have reached a different conclusion.[32]
- *Availability heuristic*: The tendency to base decisions or conclusions on information that can be easily accessed or brought to mind.[33]
- *Confirmation bias*: The tendency to search for or favour information that supports one's favoured hypotheses, while disregarding, avoiding or rejecting information that counters them.[34]
- *Groupthink*: Defined as 'a mode of thinking that people engage in when they are deeply involved in a cohesive in-group, when the members' strivings for unanimity override their motivation to realistically appraise alternative courses of action.'[35] In other words, one person's interpretation of the

27 B. Friedman and H. Nissenbaum, 'Bias in Computer Systems', 14 *ACM Transactions on Information Systems* (1996) 330, at 335 (the authors use the term 'technical bias' for this sub-category, but to avoid confusion with the overarching category, we have re-named it 'machine bias').

28 *Ibid.* See further, Institute of Technological Ethics, *Three Kinds of Bias in Computer Systems*, available online at https://www.technologicalethics.org/three-kinds-of-bias (visited 5 November 2020).

29 Friedman and Nissenbaum, *supra* note 27, 335.

30 D. Simon, *In Doubt: The Psychology of the Criminal Justice Process* (Harvard University Press, 2012), at 38.

31 A. Tversky and D. Kahneman, 'Judgment under Uncertainty: Heuristics and Biases', 185 *Science* (1974) 1124, at 1128: Participants were asked to spin a roulette wheel and then guess what percentage of states in the United Nations were African nations. Those participants who received 10 on the roulette wheel guessed 25%, on average, while the average guess for those who landed on 65 was 45%.

32 R. Parasuraman and D.H. Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration', 52 *Human Factors* (2010) 381.

33 Tversky and Kahneman, *supra* note 31, at 1127–1128.

34 R. Nickerson, 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises', 2 *Review of General Psychology* (1998) 175.

35 I. Janis, *Victims of Groupthink* (Houghton Mifflin, 1972), 9.

data can influence that of others. This may also occur consequent to online narratives.

- *Information bias*: Also called observation or measurement bias, information bias is 'any systematic difference from the truth that arises in the collection, recall, recording and handling of information in a study.'[36] Incomplete data is a key cause of information bias.[37]
- *Selection bias*: When particular groups or individuals are over- or under-represented in a study population, leading to a systematic error in extrapolating results to a broader population.[38]

While the above list is by no means a comprehensive overview of all possible relevant biases, this article demonstrates how technical and cognitive biases such as these can potentially permeate international criminal investigations that use digital open source information. Section 2 outlines the risk of bias affecting the evidence-gathering phase of investigations. Section 3 examines how cognitive and technical biases may creep into international criminal investigators' analysis of open source information. Section 4 discusses means to mitigate against these biases.

2. Biases in Gathering Open Source Information

Successful investigations are founded upon collecting and preserving a broad range of reliable, relevant and probative evidence.[39] However, in conducting open source investigations, the ability to find information on all of the potential international crimes committed in an area under investigation may be limited by technical and cognitive biases in ways that ultimately hinder this objective.

A. Technical Biases

A digital open source investigation will typically involve querying Internet search engines (such as Google or DuckDuckGo), as well as social media platforms (such as Facebook, Twitter, Snapchat,[40] Instagram, TikTok or

36 C.R. Bankhead, E.A. Spencer and D. Nunan, 'Information Bias', *Catalogue of Bias*, available online at https://catalogofbias.org/biases/information-bias/ (visited 10 November 2020).

37 C.J. Howe, L.E. Cain and J.W. Hogan, 'Are all Biases Missing Data Problems?' 2 *Current Epidemiology Reports* (2015) 162.

38 D.G. Kleinbaum, K.M. Sullivan and N.D. Barker, 'Is There Something Wrong? Validity and Bias', in D.G. Kleinbaum, K.M. Sullivan and N.D. Barker, *A Pocket Guide to Epidemiology* (Springer, 2007) 109.

39 *Independent Expert Review of the International Criminal Court and the Rome Statute System: Final Report*, 30 September 2020, at 253–255; Office of the Prosecutor, *OTP Strategic Plan 2019-2021*, § 14.

40 A guide to using Snapchat for these purposes is available online at https://citizenevidence.org/2019/12/10/how-to-use-snapchat-to-monitor-breaking-events/ (visited 3 February 2021).

Parler[41]).[42] Each of these platforms and search engines rely on proprietary algorithms that determine which search results are given priority. While the algorithms and how they work are kept secret to protect commercial interests and prevent manipulation,[43] they appear to be informed by factors like: keyword matches; the number of views a page or piece of content has already received (or number of followers for a social media profile); the time of day it was posted; geographic location; and the searcher's search history.[44] The voices that these algorithms amplify — for example, by prioritizing paid accounts, or websites or user profiles that already receive a lot of traffic or have many followers (including media outlets) over smaller, independent, websites or pages — may be informed by the power structures in society and may in turn perpetuate the marginalization of less powerful actors. Thus, each platform's algorithm can lead to relevant material being overlooked. A completely neutral search is not possible, even when the investigator is aware of the risk of algorithmic bias and takes steps to mitigate against it, such as by deleting cookies before commencing their search; deploying a virtual private network; setting up a 'blank' profile to conduct searches that are free of influence from one's personal search history, and using diverse filters and search terms.[45]

New technologies are continually being developed to support open source human rights investigators in discovering the most relevant open source evidence, for example through the automated detection of particular weapons in large caches of videos or photographs.[46] While these are undoubtedly positive developments, investigators should be aware of the risk posed by machine bias when automating tasks, and vigilant as to the limitations of object detection technologies in identifying the full range of international crimes documented through open sources. As Koenig and Egan have noted, a dependence on automation risks increasing 'the tendency towards certain crimes becoming hypervisible, such as chemical weapons attacks or the bombing of hospitals, potentially drawing attention to those crimes like shiny objects, while

41 Since the time of writing Parler has been taken offline.

42 See e.g. P. Myers, 'How to Conduct Discovery Using Open Source Methods', in Dubberley, Koenig and Murray (eds), *supra* note 2, 107.

43 J. Burrell, 'How the Machine ''Thinks'': Understanding Opacity in Machine Learning Algorithms', 3 *Big Data & Society* (2016) 1, at 3–4.

44 D. Davies, 'How Search Engine Algorithms Work: Everything You Need to Know', *Search Engine Journal*, 25 May 2020, available online at https://www.searchenginejournal.com/search-engines/algorithms/#close (visited 9 November 2020).

45 See Berkeley Protocol, *supra* note 2, 11–12.

46 For example, VFRAME have developed object detection algorithms to identify cluster munitions, Forensic Architecture have developed object detection algorithms to detect tear gas canisters, and the European Human Rights Advocacy Centre and Foresnic Architecture used machine learning to gather and present evidence of Russian military presence in Ukraine, for a submission to the European Court of Human Rights, as outlined online at https://ehrac.org.uk/news/machine-learning-russian-military-presence-eastern-ukraine/ (visited 3 February 2021).

detracting from other online information, such as that related to sexual violence, which may not as easily be captured by machines.'[47]

Emergent bias can also be an issue in evidence gathering processes, insofar as social media platforms have become 'accidental archives' of human rights documentation, contrary to platform creators' initial expectations of how their platforms would be used.[48] For these reasons, there are inherent limitations that can be problematic for open source investigators — for example, many platforms and social messaging services erase the metadata from an image when it is uploaded, which means that relevant information for investigators (such as the time, date and location of creation) cannot be extracted from the content itself.[49] Moreover, because content is typically uploaded by members of the public (not trained investigators), and often in situations of extreme stress, they may not label or tag content in a way that makes it easily accessible for investigators. For example, people may take to social media to discuss and share recordings of an air strike on a hospital, but instead of outlining information such as the precise location and specific details of the strike, they are much more likely to use every day and exclamatory language (terms such as, 'what just happened?!?' or 'holy shit!!'), in the local language. The fact that users are documenting their lived experiences as the platforms intended — to communicate with their peers — rather than deliberately documenting potential human rights violations, makes it very difficult for investigators to find the content using typical search terms, such as 'explosion', 'bomb' or 'air strike'.

B. Cognitive Biases

When conducting manual searches, the information available to an open source investigator is necessarily circumscribed by the search terms they use: a search can only return results in response to a specific query inputted by the investigator. However, knowing what to look for, and what search terms to use, is not always straightforward, and may result in information bias. The most widely understood influence on the discovery phase of an open source investigation is the language used. This can play out in a number of ways. Most obviously, if an investigator is searching in a language that is different from the primary language of those involved in an incident and uploading content, only a subset of the total available content will be returned as results. That being said, some uploaders may be aware that content in one

47 A. Koenig and U. Egan, 'Hiding in Plain Site: Using Online Open Source Information to Investigate Sexual Violence and Gender-Based Crimes', in J. Dawes and A.S. Moore (eds), *Technologies of Human Rights Representation* (SUNY Press, 2021).

48 J. Deutch, 'Challenges in Codifying Events Within Large and Diverse Data Sets of Human Rights Documentation: Memory, Intent, and Bias', 14 *International Journal of Communication* (2020) 5055, at 5056.

49 For example, location information is turned off by default on Twitter, and metadata is scraped from images on Facebook, Twitter and WhatsApp. Of course, the camera-original content will still retain metadata, if this can be tracked down: https://citizenevidence.org/2020/04/20/sending-encrypted-photos-while-preserving-metadata/ (visited 3 February 2021).

of the most widely-spoken world languages may receive more attention and may tag their posts accordingly, but to do so requires both a level of tech-savviness and language skills that many people will not have, especially in regions where education and mobility are limited.[50]

Being aware of the need to search for content in local languages, investigative teams typically incorporate people with relevant language skills or, less ideally, use language translation software. This is not always sufficient, however, and context-specific knowledge may also be required in order to adequately locate and interpret content. It may be the case that uploaders use local terminology that is unfamiliar to outsiders, even if they speak the language. In the Gaza Strip, for example, '*zenana*' (a slang Arabic term used to refer to a nagging wife) is also the term typically used to refer to a drone.[51] An Arabic-speaking investigator who does not know this may not find relevant information if their search terms are limited to formal terminology. Similarly, Koenig and Egan have found that people often use coded language to refer to sexual and gender-based violence.[52] Contrary to popular belief, their research indicates that information pertaining to such crimes was being shared online (by survivors, perpetrators and bystanders), but that investigators were not always aware of such practices and did not realize how to look for that evidence.[53]

The fact that certain types of crime are more readily discoverable or more overtly visible may lead to a form of selection bias, where the information that is identified and used does not represent the broader situation.[54] Representativeness and selection bias are common issues in investigations generally; in interviewing witnesses, for example, access to witnesses can be mediated by gatekeepers or intermediaries, meaning that a full range of perspectives may not be gathered through face-to-face interviews alone.[55] In open source investigations, in contrast, anybody with access to a mobile phone (and a data plan) can theoretically share content online.[56] In practice, however, we know that access to technology and the internet is heavily mediated by financial,

50 J. Aronson, 'Mobile Phones, Social Media and Big Data in Human Rights Fact-Finding: Possibilities, Challenges, and Limitations', in Alston and Knuckey (eds), *supra* note 10, 441. See further, N. Milaninia, 'Biases in Machine Learning Models and Big Data Analytics', International Review of the Red Cross (forthcoming, 2021) online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3744164 (visited 4 March 2021), 14 (noting that 'less than 0.1% of the content on the Internet is in Pashto and 3% in Dari, the dominant languages in Afghanistan').

51 S. Wilson, 'In Gaza, lives Shaped by Drones', *Washington Post*, 3 December 2011; J. Cook, 'Gaza: Life and Death under Israel's Drones', *Al Jazeera*, 28 November 2013.

52 Koenig and Egan, *supra* note 47.

53 *Ibid.*

54 S. Edwards, 'Open Source Investigations for Human Rights: Current and Future Challenges', in Dubberley, Koenig and Murray (eds), *supra* note 2, 87, at 91–92.

55 OHCHR, *Integrating a Gender Perspective into Human Rights Investigations* (United Nations, 2018), at 16–23.

56 Land, *supra* note 10; Dyer and Ivens, *supra* note 11; R.J. Hamilton, 'User-Generated Evidence', 57 *Colombia Journal of Transnational Law* (2018) 1.

social, educational, technical, political, geographic and physical constraints,[57] meaning that the same risks of selection bias and over-representation of certain perspectives (e.g. those of digitally literate, young, urban, men) remain. Relatedly, it is a truism that certain types of violation are more amenable to open source documentation than others. Witnesses will often record and post content relating to visible violations, such as murder, destruction of property, and artillery or air strikes, but may not have access to other more 'hidden' atrocities, such as torture, ill-treatment of detainees or starvation.[58]

Where key hypotheses have been identified early in the evidence-gathering process,[59] cognitive shortcuts may lead the investigator to fall prey to confirmation bias or anchoring — where later evidence discovered reinforces the initial hypothesis, while contrary or exculpatory information is overlooked or disregarded. Lidén has written convincingly of the increased susceptibility to confirmation bias that can result from information overload and the limits of human cognitive and processing capabilities.[60] This risk is even greater in the context of open source investigations, where the volume of content through which investigators have to trawl to find the most relevant evidence can be overwhelming.[61] Aside from being a matter of investigative good practice, overcoming confirmation bias is particularly important in the context of tribunals such as the ICC, where there is an obligation 'to investigate incriminating and exonerating circumstances equally', in order to establish the truth.[62]

57 Edwards, *supra* note 54, at 93; Y. Ng, 'File Sharing and Communication During an Internet Shutdown', *Witness blog*, 31 January 2020, available online at https://blog.witness.org/2020/02/file-sharing-communication-internet-shutdown/ (visited 19 November 2020); U. Egan, 'Intersectionality and International Criminal Investigations in a Digital Age', *Opinio Juris*, 19 December 2019, available online at http://opiniojuris.org/2019/12/19/digital-accountability-symposium-intersectionality-and-international-criminal-investigations-in-a-digital-age/ (visited 19 November 2020).

58 Y. McDermott, D. Murray, and A. Koenig, 'Whose Stories Get Told, and by Whom? Representativeness in Open Source Human Rights Investigations', *Opinio Juris*, 19 December 2019, available online at http://opiniojuris.org/2019/12/19/digital-accountability-symposium-whose-stories-get-told-and-by-whom-representativeness-in-open-source-human-rights-investigations/ (visited 19 November 2020). It should be noted that such content is often posted by perpetrators, and so may be available,

59 As Agirre explains, investigations are an iterative process, and pretending to start with a 'blank canvas' or a complete absence of hypotheses would be fallacious; it is better to acknowledge them from the outset so that they can be tested and interrogated: X. Agirre Aranburu, 'The Contribution of Analysis to the Quality Control in Criminal Investigation', in X. Agirre, M. Bergsmo, S. De Smet and C. Stahn (eds), *Quality Control in Criminal Investigation* (TOAEP, 2020) 117, at 125 *et seq*.

60 M. Lidén, 'Confirmation Bias in Investigations of Core International Crimes: Risk Factors and Quality Control Techniques', in Agirre, Bergsmo, De Smet and Stahn (eds), *supra* note 59, 461, at 502.

61 M. Puttick, *Eyes on the Ground: Realizing the Potential of Civilian-led Monitoring in Armed Conflict* (Ceasefire, 2017), at 24; Deutch, *supra* note 48, at 5055 (noting that, 'For the conflict in Syria, which began in 2011, there are more hours of user-generated content documenting rights violations uploaded to digital platforms than there have been hours in the conflict itself.')

62 Art. 54(1)(a) ICCSt.; Internal Rule 55(5) ECCC IR.

This problem is often compounded when particular incidents receive a great deal of public attention, and the same content related to these incidents is shared and uploaded multiple times, potentially drowning out material related to other incidents. In the course of our research interviews, several investigators expressed concern about open source information that depicted particularly shocking incidents, such as attacks on children, being shared widely. A significant amount of content related to a limited number of incidents may drown out other incidents, making them less visible and less accessible to investigators during the discovery phase. A reliance on the availability heuristic can lead to systematic biases in an investigative plan, if these emblematic incidents, which may be the most readily retrievable from one's memory, shape the investigation to the detriment of other incidents. Moreover, there are limits to the utility of information relating one particularly shocking incident in legal accountability processes, where the aim is, for example, to demonstrate intent, or to show a widespread or systematic attack against a civilian population.

Emblematic incidents can also affect the investigative process in a number of other ways. First, there is a real risk that the digital environment will contaminate eyewitnesses' memories, since witnesses' perceptions of events can be shaped by information they later learn about those events,[63] negatively affecting the accuracy or reliability of their testimony. Secondly, emblematic incidents shared widely on social media can shape witnesses' interactions with investigators, by giving rise to an expectation on the part of the witness that the investigator will want to hear about the emblematic incident, rather than other incidents that they witnessed, but which they perceive to be less important.[64]

3. Biases in the Analysis of Open Source Information

Cognitive and technical biases not only emerge during the search or discovery phase of an investigation, but also during analysis — the act of interpreting discovered information, including assessing its meaning, reliability and probative value, and linking it to potential crimes within the jurisdiction of the investigator. Given that technical and cognitive biases can interfere with each stage of the investigation process, there is a very real risk that these biases will compound at the analysis stage.

A. Technical Biases

In light of the enormous volume of digital open source information that can be generated, tools that filter and categorize that information can be incredibly helpful to investigators. For example, object detection software can sift through

63 E. Loftus et al., *Eyewitness Testimony: Civil and Criminal* (6th edn, LexisNexis, 2020) §§ 4–7[a]; Abrahams and Murray, *supra* note 9, at 324.

64 McDermott et al., *supra* note 58.

huge masses of video evidence that have been collected from a region, to identify potential matches for a specific object, such as a cluster munition.[65] This can potentially save hundreds of analyst hours that would otherwise be spent viewing all the videos in the dataset. Other tools may use facial recognition technology to help investigators identify key perpetrators or actors in a collected set of images,[66] while a number of tools are in development which aim to detect incitement or hate speech.[67]

Aside from the risk of automation bias, where the results given by such automated tools are perceived as more accurate than human analysis,[68] investigators should be mindful of the presence of algorithmic bias. Underlying algorithms are informed by training data, and that data may have inherent biases or discrimination that the investigator may be unaware of. For example, facial recognition algorithms tend to be trained on large datasets of celebrities, which means that the lack of diversity in Hollywood may produce a racially biased model.[69] Similarly, the datasets used to train hate speech detection tools may reflect key terms and phrases at the time that the data was collected, but given that language is ever evolving, and hate speech is inherently context-dependent, these tools are likely to overlook newer terms or phrases.[70] This is illustrated by the use of memes in China, which mock officials or official policy, but are designed to evade Internet censorship.[71] Equally, and as noted above, there is a risk that the availability of certain tools will focus investigators' attention on potential violations discoverable by these tools, at the expense of other violations.[72]

B. Cognitive Biases

As established above, cognitive biases 'result from … subconscious mental procedures for processing information'.[73] These biases are 'simplifying

65 K. Hao, 'Human Rights Activists want to use AI to Help Prove War Crimes in Court', *MIT Tech Review*, 25 June 2020, available online at https://www.technologyreview.com/2020/06/25/1004466/ai-could-help-human-rights-activists-prove-war-crimes/ (visited 18 January 2021).

66 For example, the *OSR4Rights* project has developed 'Facesearch', a face matching tool enabling human rights investigators to locate whether the same person appears in a collection of images.

67 See, for example, Hatebase, 'How it works', available online at https://hatebase.org/how_it_works (visited 20 November 2020); HADES, the hate speech detection tool in development by the *OSR4Rights* project; E & T, 'AI Tools could Consider Context when Detecting Hate Speech', 24 February 2020, available online at https://eandt.theiet.org/content/articles/2020/02/ai-tools-could-consider-context-when-detecting-hate-speech/ (visited 20 November 2020).

68 See above, text to note 32.

69 ODSC, 'The Impact of Racial Bias in Facial Recognition Software', *Medium*, 15 December 2018.

70 V. Ikoro, 'Learning to Detect Hate Speech Better: An Investigation of the Contribution of FastText and Hatebase Features' (Working Paper, 2020, on file with authors).

71 K. Keng Kuek Ser, 'Want to Circumvent China's Great Firewall? Learn these 9 Phrases First', *The World*, 20 July 2015; A. Abad-Santos, 'How Memes Became the Best Weapon Against Chinese Internet Censorship', *The Atlantic*, 4 June 2013.

72 See text to notes 46 and 47 above.

73 R.J. Heuer, Jr, *Psychology of Intelligence Analysis* (Central Intelligence Agency, 1999), available online at https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/art12.html (visited 20 November 2020).

strategies and rules of thumb to ease the burden of mentally processing information to make judgments and decisions' that can lead to 'mental errors'.[74] On the positive side, such biases often result in errors that are consistent and predictable thanks to the fact that many of these biases have been the subject of intensive psychological research, and thus can be accounted for. While simple awareness is often not enough to counter the likelihood that cognitive biases influence analysis, their consistency and predictability mean that investigators can create protocols to help safeguard against common errors (as discussed in greater detail in Section 4 below). Cognitive biases are also a feature of traditional 'offline' investigations, but — as discussed below — open source investigations raise a number of new issues.

The anchoring bias can play an especially critical and distorting role during the interpretation of digital open source information. The first cognitive or affective impression created through the analysis of images found online, for example, can be difficult to reverse and can easily impact later interpretation of visual assets. If the investigator has a dominant interpretation of a related event from early online searching or through other sources, there is a very real risk that the investigator will interpret the visuals in ways that conform to their initial perspective.

Importantly, despite their perceived objectivity, subjective considerations are needed to make sense of visual and other sensory information.[75] As explained by researchers Malcolm, Groen and Bakr, 'the diagnostic value of a visual property depends on a combination of the current goal and prior experience of the observer, as well as its availability within the scene and relationship to other properties'.[76] What we perceive is not the product of passive experience, but ultimately a construction of reality.[77] During the analysis process, 'we tend to perceive what we expect to perceive'.[78] This is especially true of visual information, such as videos, which are commonly relied upon in open source analysis. Humans have a tendency to value and weigh sensory information — such as information that incorporates sight and sound — more heavily than abstract information, such as numbers or statistics. As has repeatedly been established, 'A … single, vivid case [often] outweighs a much larger body of statistical evidence or conclusions reached by abstract reasoning.'[79] Given this, a widely circulated video of an atrocity may readily drown out other, less emotive and less visceral information, such as physical evidence or traditional documents, even when the weight of competing evidence points away from what a video apparently depicts.

74 *Ibid.*, at 111. Cognitive limitations cause people to employ various simplifying strategies and rules of thumb to ease the burden of mentally processing information to make judgments and decisions.

75 See e.g. G.L. Malcolm, I.I.A. Groen and C.I. Bakr, 'Making Sense of Real-World Scenes', 20 *Trends in Cognitive Science* (2016) 843.

76 *Ibid.*

77 Heuer, *supra* note 73, chapter 2.

78 *Ibid.*

79 Heuer, *supra* note 73, chapter 10.

While videos are typically perceived as objective representations of an event, authors of one study found that 'the ways in which people watch video, as well as the vividness of the format itself, may [actually] encourage biased decision-making'.[80] For example, numerous studies have established that observers of video evidence can also reach very different conclusions about what the video depicts, who should be punished, and how harshly, depending on the viewer's identity, including group affiliation — such as race or gender — and based on who they identify with in the video, if anyone.

Associational biases, especially implicit associational biases that rely on stereotypes — such as associations between race and perceived rates of criminality — can also be dangerous influences on the interpretation of digital open source information. In a relatively recent study that looked into associational biases and juror interpretations of police body camera footage, the researchers concluded that 'while body cameras can help provide evidence in credibility battles between opposing witnesses, they are dangerous because they give a narrow perspective of an encounter that may simply reinforce the implicit biases of those who watch the video'.[81] As pointed out by the authors, 'what any viewer sees is ''influenced by the viewer's cultural, demographic, social, political, and ideological characteristics'''.[82]

Investigators are similarly known to be biased in ways that lead to them perceiving cause and effect between data points even when such connections do not exist. This is due to the human propensity to seek order and meaning in the world. Investigators and other fact-finders are vulnerable to creating mental stories that provide coherence between information found online — even when the coherence is a fallacy.[83] Thus, investigators may fall victim to the very human tendency to build a coherent narrative that 'makes sense' of disparate data points that are in fact random, developing false narratives that are 'internally consistent as well as consistent with the available evidence',[84] but nevertheless inaccurate.

Even the apparently objective act of storing information can reflect the politics, perceptions and biases of the individual investigator, through the filenames, data categories and/or tags they choose in preserving and archiving evidence.[85] For example, a video showing violence against protestors may be stored in an archive of evidence using any of the following terms: 'police brutality'; 'disproportionate force', 'violence against protestors', 'attack', 'police

80 Y. Granot et al., 'In the Eyes of the Law: Perception versus reality in appraisals of video evidence', 24 *Psychology, Public Policy, and the Law* (2018) 93.

81 M.A. Birck, 'Do You See What I See? Problems with Juror Bias in Viewing Body-Camera Video Evidence', 24 *Michigan Journal of Race and Law* (2018) 153, at 157.

82 *Ibid.* (quoting H.M. Wasserman, 'Moral Panics and Body Cameras', 92 *Washington University Law Review* (2015) 831).

83 See e.g. N. Pennington and R. Hastie, 'Explaining the Evidence: Tests of the Story Model for Juror Decision Making', 62 *Journal of Personality and Social Psychology* (1992) 189.

84 Heuer, *supra* note 73, chapter 11.

85 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, at 46.

suppress riot' or 'police re-establish control'. Each of those terms has its own weight and meaning and reflect the subjective views of the person storing and indexing the information. On the other hand, a dispassionate and detached description (such as, 'person falls to the ground') can be meaningless and may lead to the evidence being overlooked in later reviews.

4. Mitigating Biases in Open Source Investigations of International Crimes

As mentioned above, simple awareness of the technical and cognitive biases that can impact upon the discovery and analysis of open source information cannot, of itself, alleviate such biases. However, awareness that biases can distort the information identification and analysis process[86] is a crucial first step in designing effective counter-strategies.

The psychological and practitioner literatures provide many potentially useful tools and techniques that should form a crucial part of the open source investigator's toolkit. In this section, we discuss some of the strategies identified in those literatures that can be integrated into investigators' workflows to mitigate the impact of bias. We organize this section around three key categories of bias touched on above: access bias, algorithmic bias and cognitive bias.

With regards to access bias, having a clear investigative plan that takes into account the 'blind spots' of open source information outlined in Section 2 above, will help investigators overcome, or mitigate against, informational gaps.[87] A good example of this can be found in the report of the independent international Commission of Inquiry on the 2018 protests in the Occupied Palestinian Territory.[88] Given that women and girls were, generally speaking,[89] less likely to have participated in the demonstrations than men and boys, the harms they suffered were less visible in the numerous open source videos and photographs which tended to depict the killing or injury of civilians. However, the Commission noted that the 'low proportion of women and girls injured and killed compared to men and boys should be understood within the prevalent social context in Gaza':[90] in this context, the absence of open source

86 Often referred to as the 'discovery' process.

87 For a sample investigation plan and technical landscape assessment, see Berkeley Protocol, *supra* note 2, annexes I and III.

88 *Report of the Detailed Findings of the Independent International Commission of Inquiry on the Protests in the Occupied Palestinian Territory*, UN Doc. A/HRC/40/CRP.2, 18 March 2019.

89 One notable exception was the killing of Dr Rouzan Al-Najjar, as expertly documented by *Forensic Architecture* and the *New York Times*: see D.M. Halbfinger, 'A Day, a Life: When a Medic was Killed in Gaza, Was It an Accident?' *New York Times*, 30 December 2018, available online at https://www.nytimes.com/2018/12/30/world/middleeast/gaza-medic-israel-shooting.html (visited 20 November 2020). The Commission of Inquiry outlined over 300 injuries to women, through live ammunition, tear gas, rubber coated metal bullets and shrapnel: *ibid.*, § 591.

90 *Ibid.*, § 592.

information should not necessarily indicate an absence of harm. In response, the COI took additional steps to examine the gendered impact of violations, for example women's disproportionate accumulation of debt, the added burden of breadwinning and/or caregiving for the wounded, and increased incidence of domestic violence.[91] An awareness of the intersectional factors that can lead to relative invisibility in digital spaces and identification of other ways in which relevant information may be documented — for example verbally, or in non-digital documentary form, or through physical evidence — can go a long way in encouraging investigators to dig deeper and ensure a full representation of a broad cross-section of impacted communities.[92]

As for algorithmic bias, including the possibility that online searches will favour results from particular parts of the world (as with the bias of social media platforms in producing information from the global north and west) or relevant to particular populations, investigators should conduct a digital landscape assessment prior to commencing any collection or monitoring.[93] Such assessments require the investigator to methodically document the online sources (including social media sites) used in the relevant jurisdiction(s). This information should be farther parsed with an eye to how use of the internet, including social media sites, and other digital technologies varies on the basis of age, gender and rural and urban divides, among other demographic factors. Templates and protocols that force investigators to plan and structure their searches can help with consciously expanding the sources they use for information collection in ways that may counteract *both* algorithmic and cognitive biases.

Algorithmic bias can also distort information discovery, collection and analysis when digital tools are used to conduct analytic tasks at scale — for example, when facial recognition technologies are used to identify potential persons of interest from enormous datasets that humans are incapable of combing, but which prove more accurate for identifying white versus black faces (making the risk of misidentification especially acute for people of colour). First, it's important for investigators to understand the biases that may be prevalent in the data used to train the tool. Secondly, the investigator may want to create 'weights' to help counter those biases (i.e. by reducing the likelihood threshold when searching for bias-affected categories of persons), or complement automated processes with human review — using the tool to bring the dataset down to human scale but then conducting an independent analysis of the resulting dataset.

Ideally, such automated tools are used in ways that complement human weaknesses (such as limitations on humans' abilities to review hundreds of thousands of hours of video, as experienced, for example, with the Syrian conflict). Of course, investigators must be careful to avoid known biases that

91 *Ibid.*, §§ 603–614.

92 Egan, *supra* note 57.

93 For a digital landscape assessment template that spotlights relevant categories of information, see Berkeley Protocol, *supra* note 2, Annex III, at 85.

come with using digital methods; research has found that digital outputs, despite best efforts at independence, may be given undue weight.[94] This has the impact of distorting or biasing researchers' allegedly 'independent' conclusions in favour of the results proposed by automated tools.

To counter both the machine and human biases noted above, international criminal investigators may also wish to adopt some of the following processes that have been found to help offset human error. These include establishing evidence review panels to 'stress test' the evidence and the strength of any working hypotheses,[95] appointing a 'devil's advocate' or 'red team' to challenge the arguments, or build competing hypotheses that can be tested against the collected evidence.[96] It is especially important that investigators document and report any assumptions that have been used in reaching analytical conclusions, which helps with external review. 'Evidence Review Boards' can also prove helpful: they are groups of senior colleagues whose role is to assess the case as a whole based on the evidence gathered, and on that basis determine whether the case can proceed to the filing of charges, or whether further investigation or amendment of the proposed charges is warranted.[97] These measures can help to overcome cognitive biases by incorporating the perspectives of analysts with some degree of distance from the investigative team. As Heuer notes: '[analysts conducting evidence review] often see things or ask questions that the author has not seen or asked. Because they are not so absorbed in the substance, they are better able to identify the assumptions and assess the argumentation, internal consistency, logic, and relationship of the evidence to the conclusion'.[98]

Co-authors with experience in running open source investigation and verification labs in universities incorporate these methods through systems of peer review, similarly analysing multiple working hypotheses (even charting or writing out all facts that support each hypothesis), and testing the null hypothesis — essentially, seeing if they can disprove the favoured working hypothesis or hypotheses. This allows them to question not only conclusions with regards to the 'five Ws' (who, what, when, where, and why), but also clearly assert *how* we know each fact to be true, to see if there may be problematic analytical gaps or assumptions that have created blind spots in our analysis.

Such peer review can also assist with a related mechanism for identifying and compensating for blind spots: ensuring that diverse investigators with different cultural schema and individual perspectives have a chance to review and analyse the evidence. Given research that establishes how peoples'

94 J.M. Logg, J.A. Minson and D.A. Moore, 'Algorithm Appreciation: People Prefer Algorithmic To Human Judgment', 151 *Organizational Behavior and Human Decision Processes* (2019) 90.

95 M. Bergsmo, *Towards a Culture of Quality Control in Criminal Investigations*, FICHL Policy Brief Series No. 94 (2019), at 4.

96 Agirre, *supra* note 59, at 257–259. See also Heuer, *supra* note 73, at 97, for a 'Step-by-Step Outline of Analysis of Competing Hypotheses'.

97 Agirre, *supra* note 59, at 259–272; Heuer, *supra* note 73, chapter 14.

98 Heuer, *supra* note 73, chapter 14.

identities influence and thus bias their interpretation of collected data — and especially visual data, such as the photographs and videos common to digital open source investigations — ensuring that investigators bring diverse perspectives to the collection and analysis of relevant information can serve as a helpful check on the quality of the investigation.

Also critical is engaging in known processes for verifying open source materials — specifically, testing what the investigator has been told the item represents and/or what the investigator believes it represents — by consistently following a three-step process. On every piece of original data (such as a photograph or video) the investigator should conduct (i) a technical analysis (seeing if the item has any metadata attached, such as geocoordinates or timestamps, that can serve as a lead for further verification), (ii) content analysis (comparing landmarks and other visual information in the item against other sources such as satellite imagery or reliable photographs or videos to see if that information is consistent with the asserted place, time, etc.) and (iii) source analysis (evaluating the reliability of the source for the information that the item represents).[99]

A now-infamous video called 'Syrian Hero Boy' illustrates the importance of consistently engaging in these three steps, including — and especially — source analysis. In 2014, an online video went viral that purported to show a young boy in Syria rescuing a young girl from where she was trapped during an active shooting. Several media outlets shared the video as a heartwarming story of bravery and kindness in the context of a horrific war. However, several media outlets declined to amplify the video, expressing scepticism of its authenticity. Ultimately, the latter were right to be sceptical: it was later revealed that the incident had been staged on the Gladiator film set in Malta.[100] The director had shot the video and posted it online, positioning it as factual, as part of an attempt to bring broader global attention to the horrors of the Syrian war and inspire international aid. One relatively common difference between those who declined to show the video and those who did was that the former engaged in the third step in the verification process: source analysis. They were unable to find more than one video from the source, which was unusual for documenters in Syria, or to find any clear information about the source's offline identity, which called the source's reliability into question. This case is ultimately a particularly excellent example of poor forensic analysis, with one BBC reporter stating (based on content analysis) that the one thing that could be conclusively determined was that the film was shot in Syria — which of course, was not accurate.

As for content analysis, investigators can and should use a variety of methods to check whether what's depicted or otherwise included in the data is consistent with what they've been told and/or with their working hypothesis.

99 Berkeley Protocol, *supra* note 2, 63–68.

100 See e.g. '#BBCTrending: Syrian 'Hero Boy' Video Faked by Norwegian Director', *BBC*, 14 November 2014, available online at https://www.bbc.com/news/blogs-trending-30057401 (visited 2 February 2021).

For example, if a video is allegedly of a particular incident in a particular village in Myanmar in 2017: is the video's content consistent with what one would have seen in 2017, at the alleged location in Myanmar? Equally, if a video claims to have been taken during the coronavirus pandemic, are people wearing face masks? Common verification practices include conducting a reverse image search of any photo (or stills, if a video) to see if the visuals have been graphed by one or more search engines and if so, if the information surrounding that image is consistent with the hypothesis; checking to see whether built or natural landmarks are consistent with what can be found in satellite imagery from that location and time, etc.[101]

Triangulation with other data types is also helpful for challenging one's working hypothesis. At a minimum, open source research methods should always be used in conjunction with other investigative methods to help counter the technical and cognitive biases endemic to digital research, with investigators aiming for diverse physical, documentary and testimonial evidence for each fact at issue in their case. This may be, for example, through complementing open source information with information gathered through witnesses, local networks or other means to capture the perspectives of marginalized populations. Some investigators use open source tools, such as satellite imagery, to supplement their more traditional investigative activities. One notable example of this is the use of satellite imagery to trace the actual journey that witnesses described in their interviews with investigators. In this way, open source tools can pave the way for an investigation underpinned by 'radical empathy', or the ability to understand and appreciate another person's standpoint or experiences, and allow that perspective to prevail over one's personal reaction or feelings.[102] This method can also be a means to check the credibility of a witness's account, for example if the journey they described varies inexplicably and significantly from the route visible in satellite imagery taken at around the same time.[103]

Finally, investigators should be mindful of shifting perceptions of what constitutes 'good' evidence in the digital age.[104] Several investigators we consulted in our research were concerned that the growing salience of open source information in mass atrocity investigations may undermine the perceived value of witness testimony.[105] One investigator we interviewed noted that those parts of an investigation that are overwhelmingly testimony-based are perceived as having less weight. In a Keynesian sense, that is correct, given

101 For more on verification, see A. Toler, 'How to Verify and Authenticate User-Generated Content', in Dubberley, Koenig and Murray (eds), *supra* note 2, 185.

102 M.L. Caswell and M. Cifor, 'From Human Rights to Feminist Ethics: Radical Empathy in the Archives', 81 *Archivaria* (2016) 23, at 25. See also, Amnesty International, 'The Hidden US War in Somalia', 19 March 2019, at 40–43.

103 It is possible to view historical satellite imagery using DigitalGlobe (now Maxar) and other satellite imagery sites. This means that, even if a route has changed over time (e.g. a road has collapsed due to flooding or an earthquake), it should still be possible to view what that route looked like at around the same time as events described.

104 McDermott et al., *supra* note 56.

105 *Ibid.*

that the quantum of total evidence is lower.[106] However, this perception is problematic in the hierarchies it creates — both between the violations that tend to be more grounded in testimony and those that are more typically visible in open source information, and between the organizations that have the skills and capacity to collect and analyse open source information and those that do not. International criminal investigators should push back against this facet of the new advocacy environment, underscoring the continuing importance of traditional investigative methods, especially when carried out in conjunction with digital fact-finding. Academics and the third sector also have a role in ensuring a more level playing field amongst human rights organizations, through collaboration and meaningful partnership, knowledge transfer, training and the democratization of the open source investigative space by sharing open source technology and sharing knowledge, experience and methods.

5. Conclusion

Digital open source information is emerging as a critical tool in the toolkit of international criminal investigators. While the international community is increasingly embracing the potential of digital information and open source methods to strengthen fact-finding and verification, it is important to simultaneously acknowledge such processes' weak spots and the potential for bias to impact upon the collection and analysis of such information. In this article, we have highlighted some key technical and cognitive biases that can impact upon digital open source investigations. These include the risk of access bias determining where and from whom digital open source information comes from; algorithmic bias shaping search results and the analysis and filtering of information gathered; and cognitive biases leading to systematic errors in the gathering or interpretation of information. These findings are relevant not just to the specific context of international criminal investigations, but to anyone (for example, journalists, academic researchers, or investigators of domestic crimes) seeking to use digital open source information to establish and prove relevant facts.

Thankfully, with careful attention to the abundant research that has been conducted into the technical and cognitive biases, and through carefully designed investigations (and investigation teams) to minimize those weaknesses, international criminal investigators should be able to overcome those hurdles. In so doing, they can ensure that digital open source information realizes its full potential to strengthen justice and accountability for the world's most grave crimes.

106 Keynes defined 'weight' as 'the amount of evidence upon which each probability is founded': J.M. Keynes, *A Treatise on Probability* (Macmillan, 1921), at 356.