International Review of the Red Cross (2020), **102** (913), 199–234.
Digital technologies and war
doi:10.1017/S1816383121000096

IRRC

Biases in machine learning models and big data analytics: The international criminal and humanitarian law implications

Nema Milaninia*

Abstract

Advances in mobile phone technology and social media have created a world where the volume of information generated and shared is outpacing the ability of humans to review and use that data. Machine learning (ML) models and "big data" analytical tools have the power to ease that burden by making sense of this information and providing insights that might not otherwise exist. In the context of international criminal and human rights law, ML is being used for a variety of purposes, including to uncover mass graves in Mexico, find evidence of homes and schools destroyed in Darfur, detect fake videos and doctored evidence, predict the outcomes of judicial hearings at the European Court of Human Rights, and gather evidence of war crimes in Syria. ML models are also increasingly being incorporated by States into weapon systems in order to better enable targeting

* The views expressed herein are those of the author alone and do not necessarily reflect the views of the ICC Office of the Prosecutor or Google. The author would like to thank Nasrina Bargzie, Alexa Koenig, Matthew Cross, Beth Van Schaack, María Sol Beker, Jérôme de Hemptinne, Nikki Ahmadi, Gayane Khechoomian, Yulia Nuzban, and the editors of the *International Review of the Red Cross*, Bruno Demeyere, Sai Venkatesh and Ash Stanley-Ryan, for their considerably important and insightful input and feedback. The article is written without any first-hand knowledge of any of the investigations described herein.

systems to distinguish between civilians, allied soldiers and enemy combatants or even inform decision-making for military attacks.

The same technology, however, also comes with significant risks. ML models and big data analytics are highly susceptible to common human biases. As a result of these biases, ML models have the potential to reinforce and even accelerate existing racial, political or gender inequalities, and can also paint a misleading and distorted picture of the facts on the ground. This article discusses how common human biases can impact ML models and big data analytics, and examines what legal implications these biases can have under international criminal law and international humanitarian law.

Keywords: machine learning, big data, international criminal law, international humanitarian law, biases, International Criminal Court.

Introduction

Due to the proliferation of mobile phone technology, together with the growth of social media through which information can be created and shared, there is exponentially more information being generated today than at any other time in history. For those documenting, investigating and prosecuting international crimes or violations of international humanitarian law (IHL), this means that there is a potential treasure trove of evidence available to uncover mass atrocities and identify those responsible for their commission. While United Nations (UN) commissions of inquiry, human rights monitors and international criminal law (ICL) institutions (such as the International Criminal Court (ICC), Kosovo Specialist Chambers and International Residential Mechanism for Criminal Tribunals) are accustomed to handling large and diverse data sets and evidential pools,[1] these institutions have only just begun to truly grasp "big data" sources – extremely large data sets that tend to require computational analysis – like social media content and other digital media.[2] Simply put, the volume of information now available has outpaced our ability to review and analyze that information using traditional investigative methods. Adding to this, new data sets are often varied and "unstructured" (i.e., do not follow a specified format[3]), such as text,

1 In the *Ratko Mladić* case at the International Criminal Tribunal for the former Yugoslavia (ICTY), for example, 377 witnesses were called and over 10,000 exhibits, including videos, forensic reports, photographs, audio recordings and handwritten documents, were admitted at trial. ICTY, "Case Information Sheet: Ratko Mladić", 2020, available at: https://bit.ly/39CgOaa (all internet references were accessed in January 2021).

2 For a definition of "big data", see Council of Europe, *Guidelines on the Protection of Individuals with Regard to the Processing of Personal Data in a World of Big Data*, 23 January 2017, n. 3, available at: https://bit.ly/34zMcVn ("The term 'Big Data' usually identifies extremely large data sets that may be analysed computationally to extract inferences about data patterns, trends, and correlations").

3 Unstructured data can be human-generated or machine-generated. Some examples of unstructured human-generated data include text files, emails, social media data and mobile data. Examples of unstructured machine-generated data include satellite imagery, scientific data, digital surveillance and

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

audio and video, and require additional pre-processing to derive meaning and support metadata.[4]

Machine learning (ML) models – systems that help software perform a task without explicit programming or rules[5] – and big data analytical tools have the power to ease these burdens by making sense of big data and providing insights that we might not otherwise have, including generating leads, showing patterns or even establishing networks and hierarchies. Non-governmental organizations (NGOs), for example, already use ML models to identify and report child pornography.[6] In Mexico, an ML model designed by local and international NGOs has been used to predict and find the location of mass graves.[7] Researchers at Carnegie Mellon University have designed an ML and computer vision-based video analysis system called Event Labelling through Analytic Media Processing (E-LAMP) to detect objects, sounds, speech, text and event types (murders, rapes or other crimes) in a video collection.[8] ML models have been used by the UK Serious Fraud Office to identify legally privileged material among millions of disclosed documents in an investigation, and by the Norwegian Labour Inspection Authority to predict high-risk workplaces to be inspected by the agency.[9] Finally, researchers at the Syrian Archive have launched VFRAME to detect cluster munition strikes in Syria and Yemen.[10] Much of the work described in these examples would take years for humans to complete; with ML models, it can take just days.

Equally, ML models are increasingly being considered for and deployed in armed conflicts. The US Department of Defense (DoD) is actively seeking to incorporate ML into intelligence collection cells that would comb through footage from unmanned aerial vehicles and automatically identify hostile activity for targeting.[11] It is also using ML models in command and control, to sift through data from multiple domains and combine them into a single source of information to provide a comprehensive picture of friendly and enemy forces and

sensor data. See UN Secretary-General, *Data Strategy for Action by Everyone, Everywhere (2020–2022)*, 2020, p. 81, available at: https://bit.ly/3iqCdY2.

4 This is in contrast to traditional structured data, like bank transactions, which are typically highly organized and formatted in a way that makes them easily searchable in relational databases. *Ibid.*, p. 81.

5 *Ibid.*, p. 80; International Committee of the Red Cross (ICRC), *Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach*, Geneva, 6 June 2019, pp. 1, 10, available at: https://bit.ly/3qtAODc.

6 Nikola Todorovic and Abhi Chaudhuri, "Using AI to Help Organizations Detect and Report Child Sexual Abuse Material Online", *The Keyword*, 3 September 2018, available at: https://bit.ly/2HJx9Qi.

7 Mimi Onuoha, "Machine Learning Is Being Used to Uncover the Mass Graves of Mexico's Missing", *Quartz*, 19 April 2017, available at: https://bit.ly/31PxFDo.

8 Jay D. Aronson, Shicheng Xu and Alex Hauptmann, *Video Analytics for Conflict Monitoring and Human Rights Documentation: Technical Report*, Carnegie Mellon University, July 2015, available at: https://bit.ly/2LXJhiH.

9 Annette Vestby and Jonas Vestby, "Machine Learning and the Police: Asking the Right Questions", *Policing: A Journal of Policy and Practice*, 14 June 2019, p. 5, available at: https://bit.ly/3nVyLp8.

10 Karen Hao, "Human Rights Activists Want to Use AI to Help Prove War Crimes in Court", *MIT Technology Review*, 25 June 2020, available at: https://bit.ly/3e9M1mX.

11 Congressional Research Service, *Artificial Intelligence and National Security*, 10 November 2020, p. 10, available at: https://bit.ly/2XNcEH5.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

assist in decision-making surrounding attacks.[12] Finally, ML is being integrated into autonomous weapons systems, including to select and engage targets.[13]

ML models, like all big data analytics tools, are not inherently objective, however. Engineers train models by feeding them data, and human involvement in the provision and curation of this data can make a model's predictions susceptible to bias.[14] This is because data collection often suffers from biases that lead to the over- or under-representation of certain groups or events, especially in big data, where many data sets have not been created with the rigour of a statistical study but are instead the by-product of other activities with different, often operational, goals.[15] For instance, an image recognition ML model produced by a computer scientist at the University of Virginia disproportionately associated pictures of kitchens with women.[16] The reason for this was that the photos used to train the software often depicted certain activities, like cooking and cleaning, being performed by women rather than men – a predictable gender bias. As a consequence of such biases, outputs from ML models or other big data analytics can be highly skewed.

To date, there is no robust international jurisprudence concerning the legality of ML models, big data analytics or even social media data under ICL or IHL. While the Special Tribunal of Lebanon had to grapple with complex telecoms analysis, for example, the Trial Chamber failed to address any of the particularly salient concerns – indeed, none appear to have even been raised – regarding bias in the collection or interpretation of that data. The closest case in point at the time of writing this article is the *Al-Werfalli* matter, which concerns an ICC arrest warrant largely based on information posted on Facebook and YouTube, but where no ML model was applied.[17] While advocates have called the case an important milestone as the first international arrest warrant based on content from social media,[18] the decision related only to an arrest warrant, meeting the lowest evidentiary threshold – reasonable grounds to believe – provided under ICL.[19] None of the social media evidence used in *Al-Werfalli*'s arrest warrant has at this time been tested on cross-examination or under the higher evidentiary threshold required for conviction at trial (beyond reasonable

12 Theresa Hitchens, "Air Force Expands 5G as It Transforms to Multi-Domain Ops: Donovan", *Breaking Defense*, 4 September 2019, available at: https://breakingdefense.com/2019/09/air-force-expands-5g-as-it-transforms-to-multi-domain-ops-donovan/.

13 Michael N. Schmitt, "Autonomous Weapons Systems and International Humanitarian Law: A Reply to the Critics", *Harvard National Security Journal: Features Online*, 5 February 2013, p. 28, available at: https://bit.ly/3ip5pyh.

14 Facebook, *Facebook's Civil Rights Audit – Final Report*, 8 July 2020, p. 76, available at: https://bit.ly/3nVlCwk.

15 ICRC, above note 5, p. 10.

16 Tom Simonite, "Machines Taught by Photos Learn a Sexist View of Women", *Wired*, 21 August 2017, available at: https://bit.ly/3qvxaIm.

17 ICC, *Prosecutor v. Mahmoud Mustafa Busayf Al-Werfalli*, Case No. ICC-01/11-01/17, Warrant of Arrest (Pre-Trial Chamber I), 15 August 2017.

18 See, for example, Emma Irving, "And So It Begins… Social Media Evidence in an ICC Arrest Warrant", *Opinio Juris*, 17 August 2017, available at: https://bit.ly/3kvEtNI.

19 Rome Statute of the International Criminal Court, UN Doc. A/CONF.183/9, 17 July 1998 (entered into force 1 July 2002) (Rome Statute), Art. 58(1).

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

doubt).[20] Given the lower standard of proof for these preliminary decisions, the jurisprudence in ICL also fails to account for advances in technology that make data manipulation, including of photographs and videos, all the easier.[21] Neither has it dealt with the impact of potential human biases.

The absence of robust jurisprudence concerning these issues is largely a reflection of the fact that international law institutions have not yet had a relevant case progress sufficiently for this understanding of social media, big data and ML to be necessary. But this absence is also an opportunity for these institutions and their investigators, analysts and prosecutors to develop rules and practices which benefit from the experience of domestic law enforcement bodies in dealing with ML and big data. The main challenge is to develop rules addressing the potential role of common human biases. To date, little scholarly work or attention has been paid to understanding these biases, their potential impact on international criminal investigations and the potential legal consequences that might arise under ICL or IHL. This article seeks to fill that scholarly gap. The first part of the article summarizes the most prevalent human biases that impact ML models and big data analytics, and the potential impact that these biases have. The second part looks at the potential legal consequences of these biases under IHL and ICL, using the Rome Statute of the ICC (Rome Statute) as a framework for analyzing those consequences.

Common biases in machine learning and big data analytics

Data sets often contain biases which have the potential to unfairly disadvantage certain groups or to over-focus on certain activities to the detriment of others, and ML models or big data analytics trained on such data sets can inherit these biases.[22] The following section discusses human biases that most commonly appear in data sets used for ML models and thus are most likely to impact ICL investigations and IHL considerations: implicit bias, selection bias, reporting bias, group attribution bias and automation bias.[23] For each, the article discusses how these biases can impact ML models or big data analytics, particularly in the context of international criminal investigations or with regard to IHL violations.

20 *Ibid*., Art. 66(3).

21 All that is required is that the interpretation of the evidence advanced by the Prosecution is a reasonable one. ICC, *Prosecutor v. Omar Hassan Ahmad Al Bashir*, Case No. ICC-02/05-01/09, Decision on the Prosecution's Application for a Warrant of Arrest against Omar Hassan Ahmad Al Bashir (Pre-Trial Chamber I), 4 March 2009, paras 32–34.

22 UN Institute for Disarmament Research, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies*, 2018, p. 3, available at: https://bit.ly/3nPmiTX.

23 The following list contains just a small selection of biases that are often uncovered in ML data sets. It is not intended to be exhaustive. Wikipedia's catalogue of cognitive biases enumerates over 100 different types of human bias that can affect our judgement and, in turn, ML models; see Wikipedia, "List of Cognitive Biases", available at: https://en.wikipedia.org/wiki/List_of_cognitive_biases. See also Forensic Science Regulator, *Cognitive Bias Effects Relevant to Forensic Science Investigations*, 4 April 2018, available at: https://bit.ly/3bNOQe9.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Implicit biases

Implicit biases occur when assumptions are made based on one's own mental models and personal experiences that do not necessarily apply more generally. These biases often have discriminatory elements, such as implicit racial or gender preferences. In 2018, for example, Amazon found that algorithms used to screen résumés to identify candidates were trained on data containing implicit biases against female applicants, resulting in the algorithm penalizing résumés that included the word "women", as in "women's chess club captain".[24]

Implicit biases can take numerous forms. A common implicit bias is confirmation bias, where individuals or model builders unconsciously process data in ways that affirm pre-existing beliefs and hypotheses.[25] Within the context of international investigations, confirmation bias can cause investigators or prosecutors to miss the exculpatory quality of evidence or to discount its value, which can lead to a failure to disclose or collect that data.[26] Similarly, in the pressurized theatre of war, confirmation bias can cause combatants to mistake civilian persons or objects for military objectives, such as when the USS *Vincennes* mistakenly downed an Iranian commercial aeroplane in 1988 due to the belief that the plane's behaviour resembled that of an F-14 warplane.[27]

Other implicit biases closely associated with confirmation bias are selective information processing, belief perseverance and the avoidance of cognitive dissonance. All three can cause prosecutors, investigators, military personnel and analysts to ignore valuable information that conflicts with their pre-existing case theory. Selective information processing causes people to overvalue information that is consistent with their pre-existing theories and to undervalue information that challenges those theories.[28] Belief perseverance is a term used to describe people's tendency to continue to adhere to a theory even after the evidence underlying the theory is disproved.[29] Finally, the desire to avoid cognitive dissonance can cause people to adjust their beliefs in order to maintain existing self-perceptions.[30] As reflected by one commentator, these biases can, in the context of criminal prosecutions, drastically impact a prosecutor's decision-making, including on investigative and charging decisions, presumptions of guilt or innocence, or the disclosure of exculpatory evidence:

24 Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women", *Reuters*, 10 October 2018, available at: https://reut.rs/2HItB0B.

25 Sonia K. Katyal, "Private Accountability in the Age of Artificial Intelligence", *UCLA Law Review*, Vol. 66, No. 1, 2019, p. 79; Kate E. Bloch, "Harnessing Virtual Reality to Prevent Prosecutorial Misconduct", *Georgetown Journal of Legal Ethics*, Vol. 32, No. 1, 2019, p. 5.

26 Alafair S. Burke, "Improving Prosecutorial Decision Making: Some Lessons of Cognitive Science", *William & Mary Law Review*, Vol. 47, No. 5, 2006, pp. 1603–1604.

27 Peter Margulies, "The Other Side of Autonomous Weapons: Using Artificial Intelligence to Enhance IHL Compliance", in Ronald T. P. Alcala and Eric Talbot Jensen (eds), *The Impact of Emerging Technologies on the Law of Armed Conflict*, Oxford University Press, Oxford, 2019, pp. 147, 158–159.

28 A. S. Burke, above note 26, pp. 1594, 1596–1599; Alafair S. Burke, "Commentary: *Brady*'s Brainteaser: The Accidental Prosecutor and Cognitive Bias", *Case Western Reserve Law Review*, Vol. 57, No. 3, 2007, p. 578.

29 A. S. Burke, above note 26, pp. 1594, 1599–1601.

30 *Ibid*., pp. 1594, 1601–1602.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

> In the context of prosecutorial decision making, the biasing theory is the prosecutor's belief that the defendant is guilty. Once that belief is formed, confirmation bias causes her to seek information that confirms the theory of guilt; selective information processing causes her to trust information tending to confirm the theory of guilt and distrust potentially exculpatory evidence; and belief perseverance causes her to adhere to the theory of guilt even when the evidence initially supporting that theory is undermined.[31]

Implicit biases are a particular problem in international criminal investigations since by the time most investigations are initiated, significant reporting of what are presented as international crimes has typically been done by news agencies, NGOs or UN agencies. For instance, the ICC's opening of an investigation into crimes committed against the Rohingya people of Myanmar occurred in November 2019,[32] years after those crimes began in 2016 and following numerous human rights reports by UN agencies and NGOs documenting their commission.[33] ICC analysts relied upon those reports when requesting authorization to open an investigation, and Office of the Prosecutor (OTP) investigators will likely continue relying on those reports for generating leads and establishing a case theory.[34] At the same time, however, such reports can, and will, have a tendency to colour an investigator's working opinion of how crimes occurred or who committed them. Similar concerns were most recently expressed by the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions in her investigation into the death of Mr Jamal Khashoggi:

> By the time the inquiry was initiated, much had already been reported about the killing and the likely responsibilities of various individuals. The risks of confirmation bias (the tendency to bolster a hypothesis by seeking evidence consistent with it while disregarding inconsistent evidence) were particularly high.[35]

These circumstances heightened the risk of confirmation bias, particularly when considering the vast amount of information available on social media and other platforms concerning the situation. Incidentally, confirmation bias also extends to the international community in its dealings with the ICC. Because the

31 *Ibid*., p. 1614.

32 ICC, *Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar*, Case No. ICC-01/19, Decision Pursuant to Article 15 of the Rome Statute on the Authorisation of an Investigation into the Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Pre-Trial Chamber III), 14 November 2019.

33 See, for example, Human Rights Council, *Report of the Independent International Fact-finding Mission on Myanmar*, UN Doc. A/HRC/39/64, 27 August 2018; Médecins Sans Frontières, *"No One Was Left": Death and Violence against the Rohingya in Rakhine State, Myanmar*, 9 March 2018, available at: https://bit.ly/3edvEFV.

34 ICC, *Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar*, Case No. ICC-01/19, Request for Authorisation of an Investigation Pursuant to article 15 (Pre-Trial Chamber III), 4 July 2019.

35 Human Rights Council, *Annex to the Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions: Investigation into the Unlawful Death of Mr. Jamal Khashoggi*, UN Doc. A/HRC/41/CRP.1, 19 June 2019, para. 37.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

OTP's *Policy Paper on Preliminary Examinations* requires preliminary examinations to be conducted on the basis of information received by the OTP in combination with open-source material,[36] there is an external (and, conceivably, sometimes internal) expectation that the investigation will correspond to the matters contained in the preliminary examination report.

ML models based on data sets impacted by implicit biases can also have significant IHL repercussions, such as through leading persons to make targeting decisions that mistake civilian objects for military objectives. In the 2015 US attack on the Médecins Sans Frontières (MSF) facility in Kunduz, Afghanistan, for example, both ground and air personnel focused on the presence of an "arch-shaped gate" in the facility's structure as well as a compound that had "an outer perimeter wall with multiple buildings inside of it", which they understood to be the characteristics of a Taliban base of operations.[37] However, such features were also common to all buildings in the region, such that the information should not have been determinative in a targeting decision. Unfortunately, personnel participating in the targeting choice confirmed the MSF facility in their targeting decision as a result of their own pre-existing notions of targetable military objects.

In both international investigations and IHL determinations, ML models based on data sets affected by implicit biases have a high probability of producing skewed analytical findings which may depict criminal correlations, relationships or patterns that do not reflect reality. As discussed further below, these biases can impact disclosure obligations and the Prosecutor's duty to seek the truth, and can potentially exacerbate stereotypes that tend to permeate in user-generated content (UGC). They can also impact targeting decisions, resulting in mistaken attacks on civilians or civilian objects. Further, they may reinforce international biases in targeting that endanger civilians, such as through encoding the policy of "signature strikes" into the algorithm.[38] More than any other form of bias, implicit bias is probably the most dominant in criminal investigations and IHL considerations, and is the one requiring the most serious remediation.

Selection biases

Selection bias occurs when data sets used to train ML models or for analysis are chosen in a way that fails to reflect their real-world distribution.[39] For instance, if

36 ICC, OTP, *Policy Paper on Preliminary Examinations*, November 2013, paras 79, 80, 104, available at: https://bit.ly/3nXQ2y6. See also ICC, *Proposed Programme Budget for 2021 of the International Criminal Court*, ICC-ASP/19/10, 10 September 2020, para. 128, available at: https://bit.ly/2LxHkJZ.

37 US Central Command, "Summary of the Airstrike on the MSF Trauma Center in Kunduz, Afghanistan on October 3, 2015", 29 April 2016, p. 389; Matthew Rosenburg, "Pentagon Details Chain of Errors in Strike on Afghan Hospital", *New York Times*, 29 April 2016, available at: https://nyti.ms/3irFBBJ; P. Margulies, above note 27, pp. 149–150.

38 Ben Tarnoff, "Weaponised AI is Coming. Are Algorithmic Forever Wars Our Future?", *The Guardian*, 11 October 2018, available at: https://bit.ly/3qz3hqT.

39 Patrick Ball, "The Bigness of Big Data", in Philip Alston and Sarah Knuckey (eds), *The Transformation of Human Rights Fact-Finding*, Oxford University Press, Oxford, 2015, pp. 425, 436–437.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

one's goal is to create a model that can operate security cameras but that model is trained only on night-time data, a selection bias has been introduced into the model that could skew its performance as it relates to daytime conditions.

Selection biases can manifest in different ways and take different forms. Coverage bias is a form of selection bias that emerges when the data set being relied upon is incomplete, such that the sample set from which conclusions are being drawn fails to represent the targeted population.[40] More pointedly, the big data population is not *the* population. For instance, if criminal analysts use an ML model to identify patterns but fail to include data pertaining to crimes committed against children, the model would suffer from coverage bias and would likely fail to detect patterns of, or in, such crimes. Non-response bias or participation bias is a form of selection bias that occurs when users from certain groups opt out from participating in the process, such that the data set ends up being unrepresentative due to participation gaps in the data collection process.[41] This bias can be prevalent where marginalized or traditionally under-represented groups distrust the process and are consequently less likely to participate in it.[42] This also happens to be a common issue in internal armed conflicts or other situations of violence, where a lack of trust with institutions and authorities can severely impact participation by vulnerable and victimized communities.[43]

Sampling bias is a form of selection bias that occurs when the data are not collected randomly from the target group, and only samples from a particular sub-part of the target group are collected.[44] The findings or models, as a result, skew in favour of the part of the population that is sampled. Finally, event size bias is a form of selection bias that refers to the probability "that a given event is reported, related to the size of the event: big events are likely to be known, small events are less likely to be known".[45] In this sense, events that are more prominently pronounced in the data set get more favourable treatment than those that are equally egregious but less pronounced. For instance, killings committed in broad daylight and disseminated widely on social media would influence ML models more than acts of sexual violence that might get less or no public attention.

40 Joann Stonier, "Fighting AI Bias – Digital Rights Are Human Rights", *Forbes*, 19 March 2020, available at: https://bit.ly/35FGKzH.

41 Hanna Tolonen, Miika Honkala, Jaakko Reinikainen, Tommi Härkänen and Pia Mäkelä, "Adjusting for Non-Response in the Finnish Drinking Habits Survey", *Scandinavian Journal of Public Health*, Vol. 47, No. 4, 2019, p. 470.

42 Andrea F. de Winter, Albertine J. Oldehinkel, René Veenstra, J. Agnes Brunnekreef, Frank C. Verhulst and Johan Ormel, "Evaluation of Non-Response Bias in Mental Health Determinants and Outcomes in a Large Sample of Pre-Adolescents", *European Journal of Epidemiology*, Vol. 20, No. 2, 2005.

43 Sam Whitt, "Institutions and Ethnic Trust: Evidence from Bosnia", *Europe-Asia Studies*, Vol. 62, No. 2, 2010.

44 Andrew D. Selbst, "Disparate Impact in Big Data Policing", *Georgia Law Review*, Vol. 52, No. 1, 2017, pp. 109, 134–135.

45 Megan Price and Patrick Ball, "Big Data, Selection Bias, and the Statistical Patterns of Mortality in Conflict", *SAIS Review*, Vol. 36, No. 1, 2014, p. 11.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Selection bias can be particularly problematic when investigating mass atrocities or human rights abuses. As noted by Jay Aronson:

> In the case of human rights abuses, the number of victims is almost always small relative to the entire population, and they are often marginalized in some way (e.g., limited access to networked devices and the Internet due to poverty or rural location), making it more likely that the convenience sample of big data is likely to miss at least some, if not many, instances of such abuses.[46]

As a result, without proactively ensuring that the data set is representative of crimes committed against typically under-represented groups, ML models and other big data analytics risk not accounting for such crimes altogether.

Selection bias also manifests when the demographic composition of the workforce engaged in analyzing and inputting the data is unrepresentative. This fact was alluded to in the recent civil rights audit of Facebook, which noted that "[a] key part of driving fairness in algorithms [is] ensuring companies are focused on increasing the diversity of the people working on and developing FB's algorithm".[47] Applying this to ICL or IHL evaluations, if OTP investigators or analysts on the ICC's Afghanistan investigation are comprised solely of persons who are not intimately familiar with Afghan culture and language(s), the investigation will almost inevitably build conclusions that are biased. An investigative team composed purely of English-speakers with no in-country experience in Afghanistan and no understanding of Pashto or Dari is more likely to focus on evidence relating to the liability of international forces such as those of the United States – since such evidence is in English and is thus more familiar and more accessible – than on incidents that are documented purely in Pashto or Dari. This is despite the fact that the clear majority of civilian casualties in Afghanistan are committed by internal forces.[48] The converse is also true in that a team consisting solely of persons of Afghan origin might lead to a different selection bias, and any ML model or analysis based on that data set would be equally biased as it would be a reflection of the team working on the investigation.

The risk of selection bias is that if data do not reflect the real distribution of events, an ML model using that data for training will learn and enforce that bias. In this way, selection bias can have an impact on whether, for instance, patterns or the scale of events are being properly reflected by the data. These are matters of central importance to a number of legal determinations under IHL, including whether attacks are of sufficient intensity, length and frequency to qualify as a non-international armed conflict,[49] and ICL, such as whether the *chapeau*

46 Jay D. Aronson, "Mobile Phones, Social Media, and Big Data in Human Rights Fact-Finding: Possibilities, Challenges, and Limitations", in P. Alston and S. Knuckey (eds), above note 39, pp. 441, 447.

47 Facebook, above note 14, p. 80.

48 UN Assistance Mission in Afghanistan, *Afghanistan: Protection of Civilians in Armed Conflict, 2019*, 2020, pp. 5–6, available at: https://bit.ly/3e8ObmQ (noting that "Anti-Government Elements continued to cause the majority (62 per cent) of civilian casualties in 2019").

49 ICTY, *Prosecutor v. Duško Tadić*, Case No. IT-94-1-A, Decision on the Defence Motion for Interlocutory Appeal on Jurisdiction (Appeals Chamber), 2 October 1995, para. 70.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

requirements for crimes against humanity (i.e., widespread or systematic attack directed against any civilian population[50]) are met.

Reporting biases

Reporting bias occurs when the frequency with which people write about actions, outcomes or properties is not a reflection of their real-world distribution or the degree to which a property is characteristic of a class of individuals.[51] Reporting bias differs from selection bias in that it focuses on the lack of representativeness in the available data, as opposed to the manner in which data is obtained. As noted by Patrick Ball and Megan Price, "[w]hereas selection bias focuses on how the data collection process identifies events to sample, reporting bias describes how some points become hidden, while others become visible, as a result of the actions and decisions of the witnesses and interviewees".[52] Reporting bias can arise when people focus on documenting circumstances that are to them unusual or especially memorable, assuming that the ordinary can "go without saying". It can also arise because "[e]asily available data tends to be reported and analyzed more often, leading to a reporting bias because harder to find information may never make it into the dataset".[53] For instance, information from Dari or Pashto sources in the Afghanistan preliminary examination is significantly harder to find – and thus less cited in the OTP's Article 15 request to authorize an investigation – than information from English sources.[54]

Reporting bias is a major big data problem compared to other biases. ML models for predictive policing, for instance, are based on where crimes are previously reported, not where they are known to have occurred. If crimes committed by one group are reported with greater frequency than those committed by others, ML models for predicting crime will clearly be biased against the former group.[55] In the context of UGC, reporting biases often result in ML predictions being skewed towards the more extreme points of the spectrum.[56] For example, in 2008, eBay documented that 99% of its user feedback was positive. This does not mean, however, that eBay has achieved great success in terms of its user experience; rather, it is more likely that eBay users are more

50 Rome Statute, above note 19, Art. 7.

51 Eirini Ntoutsi *et al.*, "Bias in Data-Driven Artificial Intelligence Systems – An Introductory Survey", *Data Mining and Knowledge Discovery*, 2019, p. 4, available at: https://bit.ly/3sCECmT. See also Jonathan Gordon and Benjamin Van Durme, "Reporting Bias and Knowledge Acquisition", *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, 2013, p. 25, available at: https://bit.ly/2LXoD2a (analyzing generally how reporting bias functions in artificial intelligence).

52 M. Price and P. Ball, above note 45, n. 4.

53 S. K. Katyal, above note 25, p. 72.

54 ICC, *Situation in the Islamic Republic of Afghanistan*, Case No. ICC-02/17, Request for Authorisation of an Investigation Pursuant to Article 15 (Pre-Trial Chamber II), 20 November 2017.

55 Randy Rieland, "Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?", *Smithsonian Magazine*, 5 March 2018, available at: https://bit.ly/2HHg2Pf.

56 Hongyu Chen, Zhiqiang Zheng and Yasin Ceran, "De-Biasing the Reporting Bias in Social Media Analytics", *Production and Operations Management*, Vol. 25, No. 5, 2015, p. 849.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

reluctant to express their negative experiences as compared to their positive ones.[57] As a result, the reported content was intrinsically biased.

To the extent that ML models or other big data analytics datasets are based on UGC (as many are), these biases can also manifest from the demographic composition of those putting information online.[58] This is because "[b]ig data tends to focus more on the 'haves' and less on the 'have-nots'".[59] As explained by Ricardo Baeza-Yates, "[a]ccessing and using the Internet correlates with educational, economic, and technological bias, as well as other characteristics, causing a ripple effect of bias in Web content and links".[60] While the number of active Facebook users (for example) is massive, not everyone uses Facebook. Similarly, while Twitter is a household name, the number of people who actively tweet is still relatively small and highly selective (about 22% of the US population, of which 10% produce 80% of all tweets).[61] In other words, there needs to be a distinction between the *producers* of social media and the *consumers* of such media – the former may not be representative of the latter, and neither may be representative of the general population. Studying Twitter posts, for example, may be closer to studying members of a certain social class than the general population.

Let us assume, for instance, that OTP investigators create an ML model to determine which crimes to prioritize in the Afghanistan investigation based on UGC, since in-country access is difficult or unsafe. Investigative decisions would assuredly be biased in light of the linguistic bias that impacts online content. In particular, it is estimated that over 30% of all websites on the Internet are in English, while the percentage of native English-speakers in the world is only around 5%. Less than 0.1% of the content on the Internet is in Pashto and 3% in Dari, the dominant languages in Afghanistan.[62]

An additional disadvantage of internet content is the limited access available to some in the general population.[63] Surveys of online content, by design, exclude the entire non-internet population and those who, for whatever reason, do not place content online. In the context of UGC, the clearest circumstance in which reporting bias comes into play is in relation to data connected with mobile phones. Despite the apparent ubiquity of mobile devices

57 Chrysanthos Dellarocas and Charles A. Wood, "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias", *Management Science*, Vol. 54, No. 3, 2008, p. 460.
58 Berkeley Protocol on Digital Open Source Investigations, HR/PUB/20/2, 2020 (Berkeley Protocol), pp. 11, 46, 55.
59 Mick P. Couper, "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys", *Survey Research Methods*, Vol. 7, No. 3, 2013, pp. 145, 147.
60 Ricardo Baeza-Yates, "Bias on the Web", *Communications of the ACM*, Vol. 61, No. 6, 2018, p. 54.
61 Stefan Wojcik and Adam Hughes, "Sizing Up Twitter Users", Pew Research Center, 24 April 2019, available at: https://pewrsr.ch/38TfNeD.
62 Holly Young, "The Digital Language Divide", *The Guardian*, available at: https://bit.ly/2Kn116q.; Web Technology Surveys, "Usage Statistics of Persian for Websites", available at: https://bit.ly/2YN4DCk; Web Technology Surveys, "Usage Statistics of Pushto, Pashto for Websites", available at: https://bit.ly/3oNtVuz.
63 Jill A. Dever, Ann Rafferty and Richard Valliant, "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?", *Survey Research Methods*, Vol. 2, No. 2, 2008, p. 47.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

in some parts of the world, not everyone has a mobile phone, and furthermore, not everyone has or uses a smartphone – Sub-Saharan Africa, for example, has high rates of mobile phone ownership but has the lowest rate of smartphone ownership of any geographic region.[64] A 2018 Pew Research Center survey found that, "[s]imilar to internet use, smartphone ownership varies by age and educational attainment in every country surveyed".[65] Furthermore, there is a significant divide by individual income when it comes to smartphone ownership, and gender can also be a divide, as in many countries women are far less likely to own a smartphone than men.[66]

To go back to the Afghanistan hypothetical, one political analyst found that of the total number of insurgent attacks in 2008 where there was at least one casualty, less than 30% made it into the international news and "[t]hus, with media-based data for Afghanistan, we capture less than a third of the violence that actually occurs on the ground".[67] A cause of this variance was the fact that reporting of violence in the country "was crucially enhanced by cellphone coverage", such that mobile coverage largely determined whether certain incidents were captured in the media.[68] Existing research has also established that US troop presence has a strong effect on whether a country receives news coverage and information is disseminated in the media.[69] For instance, because of the involvement of so many Western international forces, there is more information about small-scale events and casualties in Afghanistan than in the Central African Republic, which rarely makes the international headlines despite the years-long war taking place in the country.[70]

From the standpoint of investigating mass atrocities or IHL violations, reporting biases can manifest in a number of ways. Crimes involving numerous victims or committed in particularly egregious ways are especially memorable; conversely, supposedly less serious crimes (i.e., those with a small number of victims) are less likely to be widely reported, particularly in the course of a violent and lengthy armed conflict. This means that open-source information concerning such events has the potential to paint a misleading picture. As noted by the US Federal Trade Commission in a report on big data, "while big data may be highly effective in showing correlations, it is axiomatic that correlation is not causation. Indeed, with large enough datasets, one can generally find some meaningless correlations."[71]

64 Laura Silver and Courtney Johnson, "Majorities in Sub-Saharan Africa Own Mobile Phones, but Smartphone Adoption Is Modest", Pew Research Center, 9 October 2018, available at: https://pewrsr.ch/3nVR6mj.

65 Jacob Poushter, Caldwell Bishop and Hanyu Chwe, "Smartphone Ownership on the Rise in Emerging Economies", Pew Research Center, 19 June 2018, available at: https://pewrsr.ch/2Ncgkjr.

66 Pew Research Center, "Mobile Fact Sheet", 12 June 2019, available at: https://pewrsr.ch/2LMq9El.

67 Nils B. Weidmann, "A Closer Look at Reporting Bias in Conflict Event Data", *American Journal of Political Science*, Vol. 60, No. 1, 2015, p. 211.

68 *Ibid.*, p. 217.

69 Timothy M. Jones, Peter Van Aelst and Rens Vliegenthart, "Foreign Nation Visibility in U.S. News Coverage: A Longitudinal Analysis (1950–2006)", *Communication Research*, Vol. 40, No. 3, 2013, p. 417.

70 N. B. Weidmann, above note 67, p. 216 and Appendix D.

71 US Federal Trade Commission, *Big Data: A Tool for Inclusion or Exclusion?*, 2016, p. 9, available at: https://bit.ly/31Or102. See also Martin Frické, "Big Data and Its Epistemology", *Journal of the Association for Information Science and Technology*, Vol. 66, No. 4, 2015, p. 659.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

An ML model designed to identify crime patterns or patterns in attacks by insurgents could, based on reporting biases in the data set, generate misleading results as to how common the crimes or attacks are, or their shared characteristics. As astutely noted by Ball and Price,

> killings in urban areas may be nearly always reported, while killings in rural areas are rarely documented. Thus, the probability of an event being reported depends on where the event happened. Consequently, analysis done directly from this data will suggest that violence is primarily urban.[72]

Relatedly, such biases can present a false understanding as to the potential pattern of crimes or attacks – an important element for assessing whether crimes were committed as part of a plan or policy,[73] whether the accused's conduct was intentional and non-coincidental,[74] or whether a given act has a nexus with or to an armed conflict.

Where reporting biases are potentially most damaging is in detecting and investigating traditionally under-reported crimes or IHL violations like sexual and gender-based crimes (SGBC).[75] As noted in a report by the UN Secretary-General, conflict-related sexual violence is routinely under-reported as a result of "the intimidation and stigmatization of survivors, as well as restrictions on access for United Nations staff".[76] Under-reporting of sexual violence has impacted at least three situations currently under investigation by the ICC, namely Afghanistan,[77] the Central African Republic[78] and Libya.[79] In her *Policy Paper on*

72 M. Price and P. Ball, above note 45, pp. 10–11.

73 See, for example, ICTY, *Prosecutor v. Nikola Šainović et al.*, Case No. IT-05-87-A, Judgment (Appeals Chamber), 23 January 2014, paras 614–634 (upholding the Trial Chamber's finding that a "discernible pattern" of forcible transfer evidenced the existence of a common plan to displace the Kosovo Albanian population).

74 See, for example, *ibid.*, paras 988, 1784; *Prosecutor v. Jean-Pierre Bemba Gombo et al.*, Case No. ICC-01/05-01/13, Judgment Pursuant to Article 74 of the Statute (Trial Chamber VII), 19 October 2016, paras 702, 707 (noting the pattern of the accused's conduct for the purposes of assessing their intent to commit the crime).

75 Berkeley Protocol, above note 58, p. 57.

76 UN Secretary-General, *Conflict-Related Sexual Violence: Report of the Secretary-General*, UN Doc. S/2019/280, 29 March 2019 (UNSG Report on Sexual Violence), para. 11. See also World Health Organization, *Global and Regional Estimates of Violence against Women: Prevalence and Health Effects of Intimate Partner Violence and Non-Partner Sexual Violence*, 20 October 2013, available at: https://bit.ly/3oXrFlp; Iness Ba and Rajinder S. Bophal, "Physical, Mental and Social Consequences in Civilians Who Have Experienced War-Related Sexual Violence: A Systematic Review (1981–2014)", *Public Health*, Vol. 142, 10 September 2016; Gerald Schneider, Lilli Banholzer and Laura Albarracin, "Ordered Rape: A Principal–Agent Analysis of Wartime Sexual Violence in the DR Congo", *Violence Against Women*, Vol. 21, No. 11, 2015; Tia Palermo, Jennifer Bleck and Amber Peterman, "Tip of the Iceberg: Reporting and Gender-Based Violence in Developing Countries", *American Journal of Epidemiology*, Vol. 179, No. 5, 2014.

77 UNSG Report on Sexual Violence, above note 76, paras 31–34.

78 *Ibid.*, paras 35–39. See also UN Panel of Experts on the Central African Republic, *Final Report of the Panel of Experts on the Central African Republic Extended Pursuant to Security Council Resolution 2399 (2018)*, UN Doc. S/2018/1119, 14 December 2018, paras 164–167; Phuong N. Pham, Mychelle Balthazard and Patrick Vinck, "Assessment of Efforts to Hold Perpetrators of Conflict-related Sexual Violence Accountable in Central African Republic", *Journal of International Criminal Justice*, Vol. 18, No. 2, 2020, pp. 394–395.

79 UNSG Report on Sexual Violence, above note 76, paras 54–59.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Sexual and Gender-Based Crimes, the ICC Prosecutor recognized the "specific challenges" associated with investigation of SGBC, including "under- or nonreporting of sexual violence owing to societal, cultural, or religious factors" and "the associated lack of readily available evidence".[80] Chronic under-reporting of SGBC means that data sets on international crimes, especially those available online, will naturally mislead in relation to the prevalence of SGBC during a particular conflict, impacting all ML models or big data analyses of that data. As a result, ML and big data analytics have the potential of aggravating the biases described above.

Group attribution bias

Group attribution bias is a tendency to impute what is true of a few individuals to an entire group to which they belong. For instance, imagine that an ML model is created to identify the most suitable candidates for a position with the OTP. In creating this model, the designers assume that the "best" candidates are individuals with a doctorate degree from a Western European university and internship experience with the ICC, purely because some successful employees possess those traits. The resulting model would suffer from group attribution bias by discounting persons who might be equally or more qualified but lack those experiences.

Two key manifestations of group attribution bias are in-group bias and out-group bias. In-group bias refers to the tendency to respond more positively to members of a group to which you also belong, or to individuals who possess characteristics that you also share.[81] In contrast, out-group bias relates to a tendency to stereotype individual members of a group to which you do not belong, or to see their characteristics as more uniform.[82] Relatedly, we also recognize variation among members of our own group with greater subtlety than members of other groups.[83]

From an IHL standpoint, group attribution biases would have to be accounted for where ML is used to identify or predict whether a person is a combatant for targeting purposes. In 2009, for instance, researchers at Norwich University, a senior military college, conducted a study in which male military cadets made rapid decisions to shoot when images of guns briefly appeared on a

80 ICC, OTP, *Policy Paper on Sexual and Gender-Based Crimes*, June 2014, available at: https://bit.ly/3in5nHk (OTP Policy Paper on SGBC), para. 50.

81 S. K. Katyal, above note 25, pp. 80–81, citing Michael J. Bernstein, Steven G. Young and Kurt Hugenberg, "The Cross-Category Effect: Mere Social Categorization Is Sufficient to Elicit an Own-Group Bias in Face Recognition", *Psychological Science*, Vol. 18, No. 8, 2007.

82 S. K. Katyal, above note 25, p. 81, citing S. Alex Haslam, Penny J. Oakes and John C. Turner, "Social Identity, Self-Categorization, and the Perceived Homogeneity of Ingroups and Outgroups: The Interaction Between Social Motivation and Cognition", in Richard M. Sorrentino and Edward T. Higgins (eds), *Handbook of Motivation and Cognition: The Interpersonal Context*, Vol. 3, Guilford Press, New York, 1996.

83 Donald M. Taylor and Janet R. Doria, "Self-Serving and Group-Serving Bias in Attribution", *Journal of Social Psychology*, Vol. 113, No. 2, 1981.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

computer screen.[84] Cadets reacted more quickly and accurately when guns were primed by images of Middle Eastern males wearing traditional clothing, and also made more false-positive errors when pictures of tools were primed by these images. The reason for this was that the cadets, who had grown up in a post-September 11 world where US armed activities had focused on Iraq and Afghanistan, had developed stereotypes towards Middle Eastern males, particularly those wearing traditional robes and turbans, as being associated with terrorists or enemy combatants.[85] ML models respond the same way. Models that are developed using data sets which exclusively focus on one group are significantly more likely to result in targeting errors that stereotype against members of that group.

These biases have also been found to manifest in the judicial decisions of highly polarized societies. For instance, a study of judicial decisions in Israeli small claims courts between 2000 and 2004, where the assignment of a case to an Arab or Jewish judge was effectively random, found that judges were between 17% and 20% more likely to accept a plaintiff's claim when the plaintiff was of the judge's same ethnicity.[86] The same study concluded that the rate of in-group bias was higher in areas recently afflicted by acts of terrorism, and that "[i]n areas which experienced relatively little ethnic strife in the recent past, the bias is substantially lower".[87] A similar study of over 100 judges in the United States revealed that judges there similarly harboured in-group biases.[88] Judges with strong preferences for white defendants gave harsher judgments to black defendants.[89] Alternatively, judges with strong preferences for black defendants were more lenient with black defendants.[90] Other recent research has shown that judges consistently exhibit negative in-group biases: when a black judge rules on a black defendant or a white judge rules on a white defendant, the sentences are 14% longer than when ruling on a defendant of an out-group.[91]

It is not difficult to think of circumstances where group attribution biases could impact matters at the ICC. As recently indicated, for instance, by the Independent Expert Review, "[m]any of the Experts' interlocutors, including Judges themselves, mentioned the extensive 'attachment' of individual Judges to

84 Kevin K. Fleming, Carole L. Bandy and Matthew O. Kimble, "Decisions to Shoot in a Weapon Identification Task: The Influence of Cultural Stereotypes and Perceived Threat on False Positive Errors", *Social Neuroscience*, Vol. 5, No. 2, 2010.

85 *Ibid.*, pp. 206, 219. See also B. Keith Payne and Joshua Correll, "Race, Weapons, and the Perception of Threat", in Bertram Gawronski (ed.), *Advances in Experimental Social Psychology*, Vol. 62, Elsevier, Amsterdam, 2020, Chap. 1.

86 Moses Shayo and Asaf Zussman, "Judicial Ingroup Bias in the Shadow of Terrorism", *Quarterly Journal of Economics*, Vol. 126, No. 3, 2011, p. 1447.

87 *Ibid.*, p. 1483.

88 Jeffrey J. Rachlinski, Sheri Lynn Johnson, Andrew J. Wistrich and Chris Guthrie, "Does Unconscious Bias Affect Trial Judges?", *Notre Dame Law Review*, Vol. 84, No. 3, 2009, pp. 1225–1226.

89 *Ibid.*, p. 1223.

90 *Ibid.* But see p. 1223 (showing that when race is explicitly manipulated, judges show the capacity to treat defendants comparably).

91 Jeff Guo, "Researchers Have Discovered a New and Surprising Racial Bias in the Criminal Justice System", *Washington Post*, 24 February 2016, available at: https://wapo.st/37Nz0hR; Briggs Depew, Ozkan Eren and Naci Mocan, "Judges, Juveniles and In-Group Bias", *Journal of Law and Economics*, Vol. 60, No. 2, 2017.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

their domestic legal systems, whether common law or civil law, as one of the reasons for inconsistent practices between Chambers".[92] In some circumstances ICC judges have gone so far as to render decisions that align with their national legal tradition, even where it seems to depart from the Court's legal texts. For example, in a university lecture, Judge Marc Perrin de Brichambaut, a French national, noted that he and fellow civil law judges on his bench chose not to grant any interlocutory appeals in the *Bemba et al.* case, despite such appeals being permitted under the Rome Statute. Brichambaut reasoned that interlocutory appeals are typically not permitted in civil law countries: "[s]o, we were civil lawyers in *Bemba and others*. We said interlocutory appeals shouldn't even exist, we will ignore it."[93]

Brichambaut's comments could be considered as in-group attribution bias, to the extent that he was willing to view more positively procedural rules emanating from his own legal tradition (civil law) when in conflict with others, or even a plain reading of the Rome Statute. Were an ML model to be designed that predicts judicial decision-making, as they have for the purpose of predicting decisions of the European Court of Human Rights,[94] group attribution biases would necessarily have to be accounted for.

Automation bias

Automation bias refers to the human tendency to favour results generated by automated or computer systems over those generated by non-automated systems, irrespective of the error rates of each.[95] As noted by one commentator, "[a]utomation bias effectively turns a computer program's suggested answer into a trusted final decision".[96] This bias largely arises in the ML context where differences exist between the actual goals being pursued, on the one hand, and the machine's understanding of those goals, and of any relevant constraints, on the other. If the algorithm fails to consider social, cultural or political factors, among others, or the user fails to recognize inherent limitations in the algorithm, automation bias can result.

92 *Independent Expert Review of the International Criminal Court and the Rome Statute System: Final Report*, 30 September 2020 (IER Report), para. 632, available at: https://bit.ly/2XSkA9Z.

93 Marc Perrin de Brichambaut, "ICC Statute Article 68", Peking University Law School, Beijing, 17 May 2017, p. 9, available at: https://bit.ly/35SIYg4.

94 Masha Medvedeva, Michel Vols and Martijn Wieling, "Using Machine Learning to Predict Decisions of the European Court of Human Rights", *Artificial Intelligence and Law*, Vol. 28, 2020; Conor O'Sullivan and Joeran Beel, "Predicting the Outcome of Judicial Decisions Made by the European Court of Human Rights", *27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, 2019, available at: https://bit.ly/3nXEBGO.

95 Linda J. Skitka, Kathleen Mosier, Mark Burdick and Bonnie Rosenblatt, "Automation Bias and Errors: Are Crews Better Than Individuals?", *International Journal of Aviation Psychology*, Vol. 10, No. 1, 2000, p. 86; Ric Simmons, "Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System", *UC Davis Law Review*, Vol. 52, No. 2, 2018, pp. 1109–1110; Mary L. Cummings, "Automation and Accountability in Decision Support System Interface Design", *Journal of Technology Studies*, Vol. 32, No. 1, 2006, p. 25.

96 Danielle K. Citron, "Technological Due Process", *Washington University Law Review*, Vol. 85, No. 6, 2008, p. 1272.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Automation bias is a significant concern in the realm of IHL assessments. In the context of autonomous weapons systems and "decision support" systems used for targeting, for example, it can result in humans placing too much confidence in the operation of those systems, even going so far as to shift "moral reasonability and accountability to the machine as a perceived legitimate authority".[97] Automation bias can also result in a parallel concern from an IHL perspective that the ML model is more likely to be correct in identifying and targeting combatants than the human operator. The tendency to trust the machine and not intervene, even when the machine appears to have made an error, is greater the more complex the ML model becomes, as there is a proclivity to trust the machine's greater sophistication.[98]

In the realm of law enforcement, automation bias can cause jurists and investigators to accept conclusions derived from ML models or other big data analytics, versus those that keep humans in the loop, without accounting for potential biases in those conclusions. Automation bias has an additional effect of creating the assumption that techniques for searching and extracting data online, or through social media, will result in identifying credible evidence quicker and more efficiently than traditional investigative techniques. To date, there is no empirical evidence to support that proposition. In the United States, for instance, ProPublica examined Northpointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, a criminal risk assessment system used in sentencing and parole hearings across the country.[99] ProPublica's research found that the COMPAS algorithm skewed towards labelling black defendants as "high risk" and white defendants as "low risk". The Wisconsin Supreme Court, however, approved the COMPAS algorithm in *State v. Loomis* while offering no real due process protections,[100] which many commentators have concluded to be a result of that court's automation bias.[101]

Automation bias has yet to become a prominent concern at the ICC or other ICL institutions since the Court has, to this author's knowledge, avoided employing automated systems or ML in decision-making. But as international institutions consider incorporating new technologies, including ML, into their practices, one can foresee similar issues arising unless proper safeguards and controls are implemented.

97 ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, Geneva, 3 April 2018, p. 14, available at: https://bit.ly/3ioj3C5.

98 Chantal Grut, "The Challenge of Autonomous Lethal Robotics to International Humanitarian Law", *Journal of Conflict and Security Law*, Vol. 18, No. 1, 2013, p. 19. See also Shin-Shin Hua, "Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control", *Georgetown Journal of International Law*, Vol. 51, No. 1, 2019, p. 141.

99 Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks", *ProPublica*, 23 May 2016, available at: https://bit.ly/39GHiHK.

100 Wisconsin Supreme Court, *State v. Loomis*, 881 N.W.2d 749, 13 July 2016, pp. 770–771.

101 Aleš Završnik, "Criminal Justice, Artificial Intelligence Systems, and Human Rights", *ERA Forum*, Vol. 20, No. 4, 2020; Katherine Freeman, "Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis", *North Carolina Journal of Law & Technology*, Vol. 18, No. 5, 2016, pp. 97–98.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Legal implications of biases under IHL

As illustrated above, ML models can be used in varied ways during armed conflict. ML might be integrated into sophisticated autonomous weaponry such as counter-rocket, artillery, and mortar (C-RAM) systems, which have some autonomy in detecting, tracking, selecting and attacking targets.[102] Further, ML models can be used to assist human decision-making concerning where and when to launch attacks. The US Joint Artificial Intelligence Center, for example, intends to use ML models to improve situational awareness and decision-making, increase the safety of equipment and implement predictive maintenance and supply.[103] Such uses come with potential IHL implications; the following sections highlight two of these.

Weapons review

A combatant's right to choose their means and methods of warfare is limited by a number of IHL rules, including treaties that prohibit the use of specific weapons.[104] Complementing these rules are IHL principles concerning new weapons, means and methods of warfare; these principles are aimed at preventing the use of emerging technologies in war that would violate international law and ensuring that the lawfulness of such technologies is determined before they are developed, acquired or otherwise incorporated into a State's arsenal.[105] Article 36 of Additional Protocol I (AP I), in particular, requires that in the "study, development, acquisition or adoption of a new weapon, means or method of warfare", all States Parties must determine whether employment of the weapon would be prohibited by AP I or "by any other rule of international law" applicable to the State in question.

Article 36 would seem to apply to weapons systems that use ML models to better enable targeting systems to distinguish between civilians, allied soldiers and enemy combatants, or even to inform decision-making for military attacks, as such systems would almost certainly qualify as a "methods" or "means" of warfare. As noted in the Commentary to AP I, both terms are broadly intended to "include weapons in the widest sense, as well as the way in which they are used".[106] The term "means of warfare" "generally refers to the weapons being

102 ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons. Expert Meeting*, Geneva, March 2016, p. 10, available at: https://bit.ly/35VHscW.

103 DoD, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*, 2018, p. 7, available at: https://bit.ly/2LKOSZN.

104 AP I includes provisions imposing limits on the use of weapons, means and methods of warfare and protecting civilians from the effects of hostilities. See Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), in particular Part III, Section I, and Part IV, Section I, Chaps I–IV.

105 ICRC, "A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977", *International Review of the Red Cross*, Vol. 88, No. 864, 2006 (ICRC New Weapons Guide), pp. 932–933.

106 Yves Sandoz, Christophe Swinarski and Bruno Zimmermann (eds), *Commentary on the Additional Protocols*, ICRC, Geneva, 1987 (ICRC Commentary on APs), para. 1402.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

used";[107] in contrast, the term "methods of warfare" relates to "how weapons are used".[108] In this regard, the material scope of Article 36 would cover not only "new" weapons in the technical sense, but also, as noted by the ICRC, "the ways in which these weapons are to be used pursuant to military doctrine, tactics, rules of engagement, operating procedures and counter measures", as well as modifications to existing weapons that alter their functions, including weapons that have already passed legal review.[109] Indeed, the Commentary to Article 36 makes it a particular point to emphasize the development of new weapons or systems leading "to the automation of the battlefield in which the soldier plays an increasingly less important role", and the concern "that if man does not master technology, but allows it to master him, he will be destroyed by technology".[110]

What is less clear is whether the ambit of Article 36 is broad enough to cover military decision support systems that incorporate ML models but which may not be directly involved in targeting cycles. This is relevant to circumstances where ML, like that which is intended by the US Joint Artificial Intelligence Center, is used to improve situational awareness, implement predictive maintenance and supply or increase the safety of equipment. Fundamentally, the question is whether such systems could be classified as a "method of warfare". This is unlikely to be the case, since even under the broad definition of what constitutes a "method of warfare", a nexus is still required with the use or operation of a weapon.[111] A case-by-case assessment would need to be undertaken as to whether the ML model impacts "the ways in which" weapons are to be used. For circumstances in which the model is limited to maintenance and supplies, it is unlikely to do so. Conversely, ML models used for military tactics, rules of engagement or operating procedures would appear to fall squarely within the scope of Article 36.

Article 36 likely requires States to account for common human biases during the "study, development, acquisition or adoption" of ML models insofar as they impact the normal and expected use of a weapon. The reason for this is that biases in ML models incorporated into weapons systems or informing military decision-making can impact specific and general prohibitions on weapons, means and methods of warfare under treaty and customary international law. For instance, ML models used in targeting systems that fail to account for human biases can run afoul of the customary international law prohibition on the use of means and methods of warfare which are of a nature to cause superfluous injury or unnecessary suffering[112] and the customary

107 *Ibid.*, para. 1957. See also Michael N. Schmitt, *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Cambridge University Press, Cambridge, 2013, Rule 41(b).
108 ICRC Commentary on APs, above note 106, para. 1957.
109 ICRC New Weapons Guide, above note 105, pp. 937–938.
110 ICRC Commentary on APs, above note 106, para. 1476.
111 ICRC New Weapons Guide, above note 105, pp. 937–938.
112 Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, Vol. 1: *Rules*, Cambridge University Press, Cambridge, 2005 (ICRC Customary Law Study), Rule 70, p. 237, available at: https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

international law prohibition on the use of weapons which are by nature indiscriminate.[113] Indeed, some NGOs have expressed the concern that "[o]nce developed, fully autonomous weapons would likely proliferate to irresponsible states or non-state armed groups, giving them machines that could be programmed to indiscriminately kill their own civilians or enemy populations".[114]

Finally, Article 36 likely requires States to account for the potential impact that biases may have on the rules of distinction and proportionality. While these rules are primarily determined in the field on a case-by-case basis, they are also relevant to new weapons assessments "to the extent that the characteristics, expected use and foreseeable effects of the weapon allow the reviewing authority to determine whether or not the weapon will be capable of being used lawfully in certain foreseeable situations or under certain conditions".[115] The rule of distinction requires that all parties distinguish between civilians and combatants, and between civilian objects and military objectives.[116] The principle of proportionality prohibits attacks against military objectives which are "expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated".[117] It also encompasses an affirmative duty that all parties to a conflict take feasible precautions for the protection of civilians and civilian objects.[118] This includes taking precautions in assessing the risk to civilians, and in selecting the appropriate timing and means of an attack.[119]

In circumstances, as discussed above, where ML models are impacted by a group attribution bias, there is a serious potential that targeting systems relying on those models may confuse civilians for combatants due to stereotypes arising from the data sets used to train them. As one example, the United States and most European Union countries are increasingly employing a collateral damage estimate methodology (CDEM) for determining the potential incidental injury to civilians and damage to civilian objects during an attack on a lawful target.[120] The CDEM parameters are largely classified, but they do appear to use

113 *Ibid*., Rule 71, p. 244; see also Rule 11, p. 37. And see International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, 8 July 1996, *ICJ Reports 1996*, paras 78, 95.

114 Bonnie Docherty, "Mind the Gap: The Lack of Accountability for Killer Robots", *Human Rights Watch*, 9 April 2015, p. 7, available at: https://tinyurl.com/16fvbit4.

115 ICRC New Weapons Guide, above note 105, p. 943.

116 AP I, Art. 51(4)(b); ICTY, *Prosecutor v. Dragomir Milošević*, Case No. IT-98-29/1-A, Judgment (Appeals Chamber), 12 November 2009, para. 53; ICTY, *Prosecutor v. Stanislav Galić*, Case No. IT-98-29-A, Judgment (Appeals Chamber), 30 November 2006, para. 190; ICTY, *Prosecutor v. Tihomir Blaškić*, Case No. IT-95-14-A, Judgment (Appeals Chamber), 29 July 2004, para. 109.

117 AP I, Arts 51(5)(b), 57(2)(iii); ICTY, *Galić*, above note 116, para. 190.

118 AP I, Art. 57(4); see also Art. 57(2).

119 Jean-François Quéguiner, "Precaution Under the Law Governing the Conduct of Hostilities", *International Review of the Red Cross*, Vol. 88, No. 864, 2006, pp. 793, 797–808.

120 Michael N. Schmitt, "Targeting and International Humanitarian Law in Afghanistan", *International Law Studies*, Vol. 85, No. 1, 2009, p. 311. See also ICRC, *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*, Geneva, March 2014, p. 83, available at: https://bit.ly/3c7h1F1; Maura Riley, "Killer Instinct: Lethal Autonomous Weapons in the Modern Battle Landscape", *Texas Law Review*, Vol. 95, 2017, pp. 33–34, available at: https://bit.ly/3iFAsGp.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

computer-assisted modelling, and it is reasonable to expect that they may benefit from ML modelling in future iterations, if not already.[121] If those ML models fail to address biases that might categorize certain civilians as combatants or civilian objects as military objectives due to biased data sets, the CDEM could underestimate the extent of the potential damage and thereby result in increased harm to civilians and civilian objects.

Given the known potential that human biases have for impacting ML models, these IHL rules require that persons designing such ML models account for common human biases in their study, development, acquisition or adoption. This includes requiring that persons using ML models in weapons systems and in decision-making for military attacks ensure that such models are capable of selecting weapons, methods or means that minimize civilian casualties. For instance, with regard to the IHL rules on targeting, the weapon, means or method of warfare should have the capacity to comply with the principles of distinction, proportionality and precaution within the specific context of an operation and given the specific biases in the data used to train the ML program. Practically, this means that a person may be obligated to test and monitor the operation of the weapon system or programming in order to reduce the impact of such biases in any decision-making. To this point, the ICRC has taken the notable view that ML models will almost invariably create "inherent unpredictability, lack of explainability and bias" in the design and use of any system into which they are incorporated.[122] This raises the possibility that no amount of testing or monitoring is sufficient to ensure that an ML-based system will pass a weapons review. Ultimately, legal reviews cannot replace the need for discussions concerning the application of IHL to varied military uses of ML. Assuming that the legal review process can never be sufficiently robust to properly account for the potential impact of biases in ML models used for military purposes (particularly given existing technical limitations), the question will always arise as to whether such systems should be precluded outright, or whether additional constraints, such as human monitoring, are necessary to protect against unanticipated harms.

Finally, while Article 36 of AP I does not specify the modality through which these legal reviews are to occur, at a minimum it requires that the State Party set up a formal procedure, which, as noted by the ICRC, "implies that there be a standing mechanism ready to carry out reviews of new weapons whenever these are being studied, developed, acquired or adopted".[123] For instance, in relation to autonomous weapons, DoD Directive 3000.09 requires the establishment of rigorous standards for designs, testing, and the training of personnel; ensuring senior-level lines of review before development and fielding; and only permitting the use of such weapons that is consistent with

121 M. N. Schmitt, above note 120.

122 ICRC, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, Geneva, August 2019, p. 3, available at: https://bit.ly/3a8787w.

123 ICRC New Weapons Guide, above note 105, p. 949.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

their design, testing, certification, operator training and doctrine.[124] While the United States is not party to AP I, the standards set forth in Directive 3000.09 provide a useful reference as to how formal processes for testing can be exacted under Article 36. This is particularly the case since only a handful of States are known to have systematic approaches to the legal review of new weapons,[125] and of those, none (at least publicly) appear to have systems in place to test for human biases that may impact ML models used in weapons systems.

Principle of individual responsibility

A second issue relates to accountability and the principle of individual responsibility. IHL establishes a set of principles and rules that guide the methods and means of warfare, including individual criminal responsibility for war crimes.[126] As noted by the ICRC, "[t]he rules of international humanitarian law are addressed to humans. It is humans that comply with and implement the law, and it is humans who will be held accountable for violations."[127]

The use of ML models in targeting assessment structures clearly has the potential to have far-reaching implications and would ultimately have an effect on the issue of accountability. There are also different implications should ML be used in decision support versus autonomous weapons. The latter could create additional legal concerns given human obligations to make certain judgements in applying IHL rules. This raises potential accountability concerns for weapons or decisions that fully rely on ML models without human intervention. Unfortunately, there is no consensus on how to resolve these concerns.

The ICRC and a number of States, for instance, have concluded that human control must always be present, to prevent any accountability gaps: "combatants have a unique obligation to make the judgements required of them by the [IHL] rules governing the conduct of hostilities, and this responsibility cannot be transferred to a machine, a piece of software or an algorithm".[128] The same conclusion was reached by all High Contracting Parties to the Convention on Certain Conventional Weapons.[129] Conversely, a number of States and commentators have taken the view that such accountability concerns are not

124 DoD, *Autonomy in Weapon Systems: Directive 3000.09*, 21 November 2012, available at: https://bit.ly/2XVRDtW; DoD, *Law of War Manual*, Washington, DC, December 2016 (DoD Law of War Manual), § 6.5.9.4, available at: https://bit.ly/3sFrrBJ.

125 ICRC New Weapons Guide, above note 105, pp. 931, 934; James D. Fry, "Contextualized Legal Reviews for the Methods and Means of Warfare: Cave Combat and International Humanitarian Law", *Columbia Journal of Transnational Law*, Vol. 44, No. 2, 2006, pp. 453, 473–479.

126 ICRC Customary Law Study, above note 112, Rule 151.

127 ICRC, above note 5, p. 7. See also DoD Law of War Manual, above note 124, § 6.5.9.3

128 ICRC, above note 5. See also Eric Talbot Jensen, "The (Erroneous) Requirement for Human Judgment (and Error) in the Law of Armed Conflict", *International Law Studies*, Vol. 96, No. 1, 2020, pp. 37–42 (summarizing the views of several States on why human control is necessary).

129 *Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2019/3, 25 September 2019, Annex IV, para. (b).

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

particularly significant, and that it is sufficient that there is some "appropriate level of human judgment" in the deployment of autonomous weapons, without specifying where that judgement need necessarily be exercised.[130] Seemingly, these States are overly relying on Article 36, which ensures that humans are not absolved of responsibility for the use of ML models in violation of IHL, or in fact any rule of international law. As the Commentary to Article 36 notes, if the measures prescribed in Article 36 "are not taken, the State will be responsible in *any* case for any wrongful damage ensuing".[131]

It is submitted, however, that this perception relies too heavily on the weapons review process. While Article 36 creates safeguards before new weapons are employed, it lacks sufficient robustness when those methods or means of warfare are actually deployed in armed conflict. For instance, Article 36 does not specify how a review of the legality of weapons, means and methods of warfare is to be carried out. Further, it does not appear to create liability or responsibility for any unforeseen consequences of a new weapon. This is particularly problematic in the context of ML. ML systems are, "by definition, unpredictable" in the sense that an ML system is constantly learning and adapting based on the data that it reviews, coupled with the fact that "the machine has no understanding, in a human sense, of the nature or concept" of the objects that it observes.[132] As a result, an ML model may entirely meet all of the IHL requirements at the weapons review stage, but then "fail" or "malfunction" in a manner that still leads to civilian harm when employed.[133] In such circumstances, an accountability gap could occur, since individual liability under ICL requires at least an "awareness that a circumstance exists or a consequence will occur in the ordinary course of events".[134]

One counter to this position would be that human biases in ML models are inevitable (as discussed in the preceding section), rendering such biases and their consequences entirely foreseeable. However, many judges at international courts have demonstrated reluctance to convict individuals based on such broad and generalized arguments, and without evidence that the individual was aware of prior, specific circumstances in which those consequences occurred.[135] In the *Bemba* case, for instance, a majority of the Appeals Chamber acquitted the accused.[136] In a separate opinion, two of the judges comprising the majority determined that the information relating to crimes that the accused's subordinates had committed was not "sufficiently specific" at the time. They reasoned that "[i]t is not enough for the commander to be aware of a generic risk

130 E. T. Jensen, above note 128, pp. 42–44.
131 ICRC Commentary on APs, above note 106, para. 1466 (emphasis added).
132 ICRC, above note 102, p. 13. See also S.-S. Hua, above note 98, pp. 128–129.
133 S.-S. Hua, above note 98, pp. 128–129.
134 Rome Statute, above note 19, Art. 30.
135 See, for example, ICTY, *Prosecutor v. Milan Milutinović et al.*, Case No. IT-05-87-T, Judgment (Trial Chamber), 26 February 2009, para. 933.
136 ICC, *Prosecutor v. Jean-Pierre Bemba Gombo*, Case No. ICC-01/05-01/08, Judgment on the appeal of Mr Jean-Pierre Bemba Gombo against Trial Chamber III's "Judgment Pursuant to Article 74 of the Statute" (Appeals Chamber), 8 June 2018.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

that his or her troops may commit unspecified crimes" since "[s]uch a risk is inherent in every military operation".[137] One can see individuals responsible for deploying an ML model which "accidentally" targets civilians due to biases in the model making the same argument and pointing to the absence of any specific prior "mistakes" in the testing phase as a defence for why they lacked knowledge, in the legal sense, of the crime. Such concerns should countenance against any suggestion that the weapons review process is a significant enough legal burden for the purposes of ensuring legal accountability.

Overall, given the implications that biases can have for fundamental IHL norms, it is essential that more robust policy-making surrounding the issue be pursued by both States and international organizations. A central part of that policy-making should be to ensure that there are sufficient enough mechanisms and modalities for accountability where ML models result in breaches of IHL.

Legal implications of biases under ICL

ML models and big data analytics impacted by any of the above-mentioned biases can have a real impact on the investigation of international crimes and judicial proceedings. They risk presenting a misleading image as to the circumstances on the ground, perpetuating negative stereotypes or other racial or gender biases and even obfuscating exculpatory information. In these ways, ML models that include biases can perpetuate those biases in a way that is self-fulfilling. Through these effects, the biases described above can have genuine legal consequences by impacting the obligations owed by the Prosecutor and the rights of accused persons and victims. Using the Rome Statute as a framework, the next section looks at some of those consequences, by focusing on the potential impact these biases can have on the Prosecutor's duty to establish the truth and investigate exculpatory evidence, on admissibility considerations for evidence and on the Prosecutor's disclosure obligations.

Impact on the Prosecutor's duty to establish the truth

One special feature of the Rome Statute is that it places upon the Prosecutor an obligation "to establish the truth" and to "extend the investigation to cover all facts and evidence relevant to an assessment of whether there is criminal responsibility under this Statute".[138] To this end, the Prosecutor is required to "investigate incriminating and exonerating circumstances equally".[139] Another manifestation of the same philosophy can be seen in Article 81(1)(b) of the Rome Statute, which allows the Prosecutor to appeal a conviction on behalf of an

137 ICC, *Prosecutor v. Jean-Pierre Bemba Gombo*, Case No. ICC-01/05-01/08, Separate Opinion of Judge Christine Van den Wyngaert and Judge Howard Morrison (Appeals Chamber), 8 June 2018, para. 44.
138 Rome Statute, above note 19, Art. 54(1).
139 *Ibid.*

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

accused person. As noted by one commentator, these features of the Rome Statute transform the Prosecutor into an "officer of justice rather than a partisan advocate".[140]

Critically, these obligations create a statutory duty that the Prosecutor's investigation be sufficiently expansive and neutral such as to "establish the truth" and ensure the collection of information that might negate guilt. As noted by the Appeals Chamber in the Afghanistan situation,

> to obtain a full picture of the relevant facts, their potential legal characterisation as specific crimes under the jurisdiction of the Court, and the responsibility of the various actors that may be involved, the Prosecutor must carry out an investigation into the situation as a whole.[141]

This responsibility is also articulated in the OTP's policy papers and codes. For instance, the OTP's *Policy Paper on Sexual and Gender-Based Crimes* recognizes that the Office "will investigate both incriminating and exonerating circumstances relating to sexual and gender-based crimes in a fair and impartial manner to establish the truth".[142] Similarly, the OTP Code of Conduct elaborates that to meet these obligations, members of the OTP must "consider all relevant circumstances when assessing evidence, irrespective of whether they are to the advantage or the disadvantage of the prosecution".[143]

As they relate to big data investigations, these obligations arguably require the Prosecutor to undertake measures to counter potential biases in any ML model or big data analysis. Prosecutors at ICL institutions have historically been criticized for mishandling exculpatory evidence;[144] these critiques are likely to be heightened if ML models and big data are relied upon without sufficient safeguards against biases that might circumvent the collection or disclosure of exculpatory evidence. As illustrated above, if the ML model is based on data that is biased in some way, then decisions that are derived from that data can systematically disadvantage individuals who happen to be over- or under-represented in the data set. Similarly, if the methodologies used to survey big data sources do not account for potential investigative biases, it is very likely that the investigative process will

140 Claus Kress, "The Procedural Law of the International Criminal Court in Outline: Anatomy of a Unique Compromise", *Journal of International Criminal Justice*, Vol. 1, No. 3, 2003, p. 608.

141 ICC, *Situation in the Islamic Republic of Afghanistan*, Case No. ICC-02/17, Judgment on the Appeal against the Decision on the Authorisation of an Investigation into the Situation in the Islamic Republic of Afghanistan (Appeals Chamber), 5 March 2020, para. 60. See also ICC, *Prosecutor v. Thomas Lubanga Dyilo*, Case No. ICC-01/04-01/06, Judgment on the Prosecutor's Appeal against the Decision of Pre-Trial Chamber I Entitled "Decision Establishing General Principles Governing Applications to Restrict Disclosure Pursuant to Rule 81(2) and (4) of the Rules of Procedure and Evidence" (Appeals Chamber), 12 October 2006, para. 52.

142 OTP Policy Paper on SGBC, above note 80, para. 48.

143 ICC, OTP, *Code of Conduct for the Office of the Prosecutor*, 5 September 2013, para. 49(b), available at: https://bit.ly/3itiSoU.

144 See, for example, ICTY, *Prosecutor v. Radovan Karadžić*, Case No. IT-95-5/18-T, Decision on Accused's Ninety-Fourth Disclosure Violation Motion (Trial Chamber), 13 October 2014; ICC, *Prosecutor v. Alfred Yekatom and Patrice-Edouard Ngaïssona*, Case No. ICC-01/14-01/18, Decision on the Yekatom Defence Request Concerning Disclosure Violation (Trial Chamber V), 18 January 2021.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

miss exculpatory information, or information pertaining to other crimes – such as SGBC – that are typically unrepresented in big data sets. These consequences clearly impact the Prosecutor's obligations under Article 54(1).

The danger of ignoring or missing exculpatory evidence is particularly acute. As noted by one commentator when evaluating ML models employed by local law enforcement in the United States, "[w]hile prosecutors' offices, like the rest of the professional world, are beginning to embrace a data-driven future, [big data collection systems] have not been engineered to identify exculpatory or impeaching evidence for the defense".[145] The reason for this is that analytical tools deployed for investigations are typically geared towards proving guilt. Most criminal analysts, for instance, now use data management systems to visualize networks, perform social network analysis and view geospatial or temporal relations to help uncover hidden connections and patterns in data, among other things. Analysts at the International Criminal Tribunal for the former Yugoslavia (ICTY) relied heavily upon Analyst's Notebook, ZyFIND and CaseMap, which enabled them to organize and categorize information with analytical notes or tags in a single database and then filter or re-organize that information in various ways to assist the analytical process.[146] Similarly at the ICC, OTP analysts use a Fact Analysis Database in support of investigations to collate and integrate all sources of evidence about relevant groups, locations, persons and other entities.[147]

While none of these tools mimic the sophistication of ML models or analytical databases employed by some countries or private corporations, they do provide insight into some of the issues that may arise. In particular, these systems are typically designed to identify relationships pointing to a person's guilt. For instance, criminal analysts, including those at the ICC, now more frequently collate communications data from call data records, emails, social media communications and other forms of communication to detect social networks, which in turn can be used to infer organizational relationships, including hierarchies.[148] That analysis can be invaluable for identifying links between senior military or political figures and the actors on the ground. But again, that analysis is largely aimed at detecting criminal networks and the role of specific individuals within them – it is not aimed at negating them, as this could create a colourable claim under Article 54(1) that the Prosecutor failed to "investigate incriminating and exonerating circumstances equally".[149]

145 Andrew G. Ferguson, "Big Data Prosecution and *Brady*", *UCLA Law Review*, Vol. 67, No. 1, 2020, p. 184.
146 Richard A. Wilson and Matthew Gillett, *The Hartford Guidelines on Speech Crimes in International Criminal Law*, 2018, para. 265, available at: https://bit.ly/2M0To6e.
147 ICC, *Proposed Programme Budget for 2018 of the International Criminal Court*, ICC-ASP/16/10, 11 September 2017, para. 330, available at: https://bit.ly/3itlDXi.
148 Nema Milaninia, "Using Mobile Phone Data to Investigate Mass Atrocities and the Human Rights Considerations", *UCLA Journal of International Law and Foreign Affairs*, Vol. 24, No. 2, 2020, pp. 283–286.
149 Elizabeth E. Joh, "The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing", *Harvard Law and Policy Review*, Vol. 10, No. 1, 2016, p. 25; Jennifer A. Johnson, John David Reitzel, Bryan F. Norwood, David M. McCoy, D. Brian Cummings and Renee R. Tate, "Social Network Analysis: A Systematic Approach for Investigating", *FBI Law Enforcement Bulletin*, 5 March 2013, available at: https://bit.ly/35SFlH6.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

The sheer volume of information that exists has also increased the probability that investigators or prosecutors will miss or ignore exculpatory evidence due to implicit biases, which in turn can impact the outcome of any ML model or big data analysis. The volume of information currently available is beyond anything prosecutors and investigators have ever had to handle before. This is especially true in relation to international crimes, which often involve many incidents spread across thousands of actors, numerous years and entire territories.[150] Complicating the issue further is the fact that investigators and prosecutors are under significant pressure to complete successful prosecutions, rather than truth-finding.[151]

It is unsurprising that in such circumstances, latent biases exacerbate these problems. Investigators and prosecutors may have witness evidence pinpointing the accused's location at a crime scene, but videos and photos shared on different social media accounts might cast doubt on the veracity of those claims. Evidence pointing toward guilt, such as threatening messages, may be commingled with exculpatory evidence, such as time-stamped pictures far from the crime scene. Given the time and effort required to extract and sift through that expanse of information, the investigative workload naturally lends itself towards prioritizing the collection of incriminating evidence over exculpatory evidence. Finally, the growing volume of data increases the difficulty for any prosecutor or investigator to better understand the relationship between different pieces of evidence, including whether some pieces undermine the credibility of witnesses or contradict their accounts. With more data, and more complex relationships between witnesses, places and groups, the question as to whether one piece of information is material to an accused's liability is even more difficult to discern.

Finally, biases in ML models and big data analytics have the potential of being impacted by under-reported crimes or those that carry serious stigmatization. Murders committed in broad daylight and captured on mobile phones might be the low-hanging fruit that investigators hold onto in lieu of the sexual assaults that take place behind closed doors.[152] Reporting and selection biases have the serious effect of devaluing under-reported crimes, especially SGBC. Models that are based on that data set carry over those biases and have the potential to hinder the Prosecutor's ability to detect, investigate and "seek the truth" for marginalized crimes. For instance, researchers at Harvard Medical School created a risk model to prevent sexual assault among female US army soldiers. In doing so, they relied upon administrative reports and surveys of sexual assault victimizations, while making adjustments to the data set to account for sexual assaults that are unreported by the victim or in any survey. They concluded that "no more than 29.3% of all sexual assaults experienced by these women were reported to authorities, no more than 34.2% were self-reported in

150 IER Report, above note 92, para. 479.

151 ICC, OTP, *Regulations of the Office of the Prosecutor*, ICC-BD/05-01-09, 23 April 2009, Regulation 8; ICC, *Proposed Programme Budget for 2020 of the International Criminal Court*, ICC-ASP/18/10, 25 July 2019, para. 278, available at: https://bit.ly/39HayOs.

152 N. Milaninia, above note 148, p. 297.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

the survey, and no more than 46.5% were reported either to authorities or in the survey".[153] Without similar types of adjustments in data sets pertaining to under-reported international crimes, any analytical result will naturally be skewed towards more prominently reported events, undermining the Prosecutor's duty under Article 54 to establish the truth.

Admissibility considerations for evidence

Another area where biased data sets in ML models and big data can have a discernible impact is on evidentiary considerations. Article 69(4) of the Rome Statute provides that the ICC may rule on the relevance or admissibility of any evidence. In doing so, it allows the Court to take into account a non-exhaustive list of factors, including "the probative value of the evidence and any prejudice that such evidence may cause to a fair trial or to a fair evaluation of the testimony of a witness".

This provision leaves the door wide open for judges to assess biases that could have impacted the collection or analysis of information when determining the impact on a fair evaluation of the witness's evidence or even the trial. It also allows judges to factor or assess those biases when determining what weight to afford the evidence, or whether to admit it at all. For instance, judges could conclude that the failure of OTP investigators to factor in biases when collecting evidence on Facebook or Twitter could have prevented them from reviewing or producing information that would undermine a witness's account.

Where biases are most likely to have an impact is on the weight to be afforded to expert testimony relating to ML outcomes or outcomes derived from large datasets, such as crime patterns. The weight of an expert's testimony is only as strong as the information he or she relies upon when drawing conclusions. If the expert's conclusions rely upon large data sets or an algorithm that has excluded or failed to account for relevant evidence, the expert opinion concerning that analysis is questionable. This type of inquiry is not new – at the domestic level, for instance, courts are already familiar with challenges to data collection methods and evaluating whether they have produced a biased sample that would reduce the data's relevance to the issue in question. This is particularly true in relation to opinions and conclusions stemming from evidence databases like Analyst's Notebook. The late US Supreme Court Justice Ruth Bader Ginsburg recognized that "[t]he risk of error stemming from … databases is not slim", noting issues with the National Crime Information Center, terror watch lists and public employment databases.[154]

Finally, biases in ML models could also impact the evidentiary scheme chosen by judges in a particular case if the judges are aware of and understand

153 Amy E. Street *et al.*, "Developing a Risk Model to Target High-risk Preventive Interventions for Sexual Assault Victimization among Female U.S. Army Soldiers", *Clinical Psychological Science*, Vol. 4, No. 6, 2016.

154 US Supreme Court, *Herring v. United States*, 555 U.S. 135, Justice Ginsburg Dissenting, 14 January 2009, p. 155.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

these biases. The modalities through which evidence can be assessed are left quite broad by the Rome Statute. The ICC Appeals Chamber has confirmed that Article 69(4) of the Statute gives Trial Chambers the discretion on whether to rule on the admissibility of each piece of evidence upon its tender during the course of proceedings (the so-called "admissions" approach), or to reserve that determination for the end of the trial after all the evidence has been tendered and heard (the "submissions" approach).[155] Increasingly, a number of the Trial Chambers have adopted the latter approach, under which the Trial Chamber defers any ruling on the relevance, probative value and potential prejudice of any piece of evidence until the end of the trial and when it begins deliberating the judgment pursuant to Article 74(2) of the Rome Statute. At that point the Trial Chamber considers all the standard evidentiary criteria for each item of evidence "submitted" during trial.[156] In a practical sense, the Trial Chamber choosing this approach effectively permits the "submission" of all evidence during the course of trial, even when there are indications that the information is inauthentic or unreliable.

In circumstances where the Prosecution intends to rely upon big data, or ML models and analytics based on big data sources, it is submitted that a Trial Chamber should use the "submissions" approach rather than making admissibility determinations during the course of the trial. The reason for this is that given the breadth of potential credibility markers with the potential to undermine the weight or relevance of a witness's statement or other piece of evidence, judges may be more inclined to view the full pool of information available to them before dismissing any singular piece too quickly. This cautiousness is sensible when one considers the potential impact that biases can have on the investigative process and, more pointedly, on the potential exclusion of exculpatory material.

155 ICC, *Prosecutor v. Jean-Pierre Bemba Gombo et al.*, Case No. ICC-01/05-01/13, Judgment on the Appeals of Mr Jean-Pierre Bemba Gombo, Mr Aimé Kilolo Musamba, Mr Jean-Jacques Mangenda Kabongo, Mr Fidèle Babala Wandu and Mr Narcisse Arido against the Decision of Trial Chamber VII Entitled "Judgment Pursuant to Article 74 of the Statute" (Appeals Chamber), 8 March 2018, paras 576–601; ICC, *Prosecutor v. Jean-Pierre Bemba Gombo*, Case No. ICC-01/05-01/08, Judgment on the Appeals of Mr Jean-Pierre Bemba Gombo and the Prosecutor against the Decision of Trial Chamber III Entitled "Decision on the Admission into Evidence of Materials Contained in the Prosecution's List of Evidence" (Appeals Chamber), 3 May 2011, para. 37.

156 ICC, *Prosecutor v. Alfred Yekatom and Patrice-Edouard Ngaïssona*, Case No. ICC-01/14-01/18, Initial Directions on the Conduct of the Proceedings (Trial Chamber V), 26 August 2020, paras 52–53; ICC, *Prosecutor v. Al Hassan Ag Abdoul Aziz Ag Mohamed Ag Mahmoud*, Case No. ICC-01/12-01/18, Annex A to the Decision on the Conduct of Proceedings (Trial Chamber X), 6 May 2020, paras 30–31; ICC, *Prosecutor v. Dominic Ongwen*, Case No. ICC-02/04-01/15, Initial Directions on the Conduct of the Proceedings (Trial Chamber IX), 13 July 2016, paras 24–25; ICC, *Prosecutor v. Laurent Gbagbo and Charles Blé Goudé*, Case No. ICC-02/11-01/15, Decision on the Submission and Admission of Evidence (Trial Chamber I), 29 January 2016; ICC, *Prosecutor v. Jean-Pierre Bemba Gombo et al.*, Case No. ICC-01/05-01/13, Decision on Prosecution Requests for Admission of Documentary Evidence (ICC-01/05-01/13-1013-Red, ICC-01/05-01/13-1113-Red, ICC-01/05-01/13-1170-Conf) (Trial Chamber VII), 24 September 2015, paras 10–13.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Disclosure obligations

A final area where biases in ML models and big data analytics could have legal implications under ICL is disclosure. The Rome Statute has a robust disclosure regime. The Prosecutor has the duty to disclose to the Defence "as soon as practicable", and on a continuous basis, all evidence in his or her possession or control which he or she believes shows or tends to show the innocence of the person or mitigate the guilt of the person, or which may affect the credibility of the prosecution evidence (Article 67(2) of the Rome Statute), or is material to the preparation of the defence (Rule 77 of the ICC Rules of Procedure and Evidence). There are potentially three ways that biases in ML models and big data analytics can affect the Prosecutor's disclosure obligations.

First, the methodology used by the Prosecution to collect information from big data sources, or as a result of any ML model, could be subject to disclosure under Rule 77 of the ICC Rules of Procedure and Evidence. This would presume, however, that the Defence is actually aware of some bias that might have impacted the Prosecution's analysis or collection. The reason for this is that Rule 77 does not require disclosure on the basis that information could be hypothetically material. The rule does not entitle the Defence to embark on a speculative "fishing expedition" to obtain information;[157] rather, it requires the Defence to make a *prima facie* showing of materiality.[158] To do so, the Defence would need to provide a particularized explanation of the materiality of the requested items with sufficient specificity.[159]

Assuming the Defence is able to establish such specificity, the information concerning the OTP's ML model or methodological approach in evaluating big data is arguably "material", though there is no ICC jurisprudence on this to date. In *Bemba et al.*, however, the Trial Chamber concluded that Requests for Assistance (RFAs) – letters that are sent to States requesting the acquisition of evidential material – were material to the Defence's preparation and disclosable under Rule 77 of the Rules of Procedure and Evidence because they were "intrinsically linked to the admissibility of the evidence relied upon by the Prosecution", namely intercepts and call data records of the defendants' criminal communications.[160]

157 ICC, *Prosecutor v. Thomas Lubanga Dyilo*, Case No. ICC-01/04-01/06, Decision on the Defence Request for Unrestricted Access to the Entire File of the Situation in the Democratic Republic of the Congo (Pre-Trial Chamber I), 17 May 2006, pp. 2–3 (rejecting the Defence's request for access to the entire file of the DRC situation, noting the Prosecution's submission that the request constituted a "fishing expedition" and did not identify the legitimate forensic purpose for the request). See also ICTY, *Prosecutor v. Dragomir Milošević*, Case No. IT-98-29/l-A, Decision on Motion Seeking Disclosure of Rule 68 Material (Trial Chamber I), 7 September 2012, para. 5.

158 ICC, *Prosecutor v. Jean-Pierre Bemba Gombo*, Case No. ICC-01/05-01/08, Decision on Defence Requests for Disclosure (Trial Chamber III), 2 July 2014, para. 29.

159 ICC, *Prosecutor v. Jean-Pierre Bemba Gombo et al.*, Case No. ICC-01/05-01/13, Decision on Mangenda Defence Request for Cooperation (Trial Chamber VII), 14 August 2015, para. 11; ICC, *Prosecutor v. Saif Al-Islam Gaddafi and Abdullah Al-Senussi*, Case No. ICC-01/11-01/1, Corrigendum to Decision on the "Defence Request for an Order of Disclosure", (Pre-Trial Chamber I), 1 August 2013, para. 40.

160 ICC, *Prosecutor v. Jean-Pierre Bemba Gombo et al.*, Case No. ICC-01/05-01/13, Decision on Defence Requests for Prosecution Requests for Assistance, Domestic Records and Audio Recordings of Interviews (Trial Chamber VII), 10 September 2015, para. 13.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

The Chamber reasoned that because the Defence intended to challenge the RFAs for being disproportionate or based on misleading information, it was "imperative that the Defence be able to test the reliability of the procedures employed in collecting the evidence against them".[161] This reasoning was consistent with a prior determination by the Chamber wherein it concluded that "material which enables the defence to assess the legality of evidence which the Prosecution intends to rely upon at trial is relevant to the preparation of the defence".[162]

The same reasoning could, arguably, be applied to the process by which OTP investigators employ ML models in their analysis or collection of big data. Given the potential impact of implicit biases, as detailed above, disclosure of that information or documents reflecting those processes could arguably be seen as being "intrinsically linked" to the admissibility of the evidence derived from those processes. In such cases, they could be subject to disclosure under Rule 77 in the same way that RFAs were in the *Bemba et al.* case. This assumes, however, that the disclosure of such information is not prohibited by any other applicable rule of law, such as Rule 81 (1) of the Rules of Procedure and Evidence, which expressly protects OTP reports and other internal documents from disclosure (see below).

Second, these biases could have an impact on the Prosecutor's obligation to disclose exculpatory material in accordance with Article 67(2) of the Rome Statute. That provision requires the Prosecution to disclose evidence "in the Prosecutor's possession or control which he or she believes shows or tends to show" the accused's innocence, mitigates the accused's guilt, or may affect the credibility of prosecution evidence. The ICC's jurisprudence is clear that the Prosecution's disclosure obligation under Article 67(2) does not require it to proactively collect exculpatory material, but only to produce that which is actually in the Prosecution's possession.[163] To that end, while the Prosecutor has an obligation to proactively search for exculpatory data in accordance with Article 54(1), the failure to do so does not amount to a disclosure violation since that information is not within its possession.

There is little jurisprudence on what it means for information to be in the Prosecutor's "control". This issue could be particularly salient as concerns analytical information and evidence procured through ML models or big data analysis. For instance, to the extent that OTP analysts are capable of generating a report through the use of one of their analytical tools that shows the accused to be disconnected or hierarchically remote from the direct perpetrators of the offence, that information is arguably in the OTP's "control", even if it fails to be in its possession. Even were that to be the case, however, such reports would likely be protected from disclosure under Rule 81(1) of the Rules of Procedure and

161 *Ibid.*

162 ICC, *Bemba Gombo et al.*, above note 159, para. 10.

163 ICC, *Prosecutor v. William Samoei Ruto, Henry Kiprono Kosgey and Joshua Arap Sang*, Case No. ICC-01/09-01/11, Decision on the Defence Requests in Relation to the Victims' Applications for Participation in the Present Case (Pre-Trial Chamber II), 8 July 2011, para. 9; ICC, *Prosecutor v. Callixte Mbarushimana*, Case No. ICC-01/04-01/10, Decision on Issues relating to Disclosure (Pre-Trial Chamber I), 30 March 2011, para. 15.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Evidence.[164] That rule expressly restricts from disclosure "[r]eports, memoranda or other internal documents prepared by a party, its assistants or representatives in connection with the investigation or preparation of the case". Nonetheless, even in those circumstances such reports could still be relevant to the Prosecutor's Article 54(1) obligation, as described above.

Finally, these biases present the possibility of making more onerous a disclosure process at the ICC that has already shown itself to be difficult to navigate. Several compounding issues have served to make disclosure particularly problematic at the ICC, and this in turn has been *the* factor in impacting the duration of proceedings.[165] These issues include the following:

- ICL investigations are typically broad in scope, covering hundreds, if not thousands, of potentially criminal incidents spread across years. As a result, the evidentiary record is particularly voluminous, often containing hundreds of thousands of pages of documents (many handwritten), hundreds of hours of audio and video materials, and thousands of pictures.[166]
- A number of ICC Chambers appear to have broadened the Prosecution's disclosure obligations beyond what is statutorily required. For instance, while Rule 77 limits the Prosecution's disclosure obligations to information that is "*material* to the preparation of the defence or … intended for use by the Prosecutor as evidence", some judges have interpreted that requirement to include anything that is of "*prima facie* relevance" to the Defence.[167] The consequence of this interpretation has been to put at risk of disclosure any and all information, despite Rule 77's clear terms. Indeed, at least one Defence team has sought, albeit unsuccessfully, to apply Rule 77 to any information that could "assist the Defence to make an informed decision as to whether to submit a request to admit additional evidence on appeal, and to then prepare that potential request".[168]
- Divergent standards on disclosure and redactions by different Chambers mean that the time-consuming process of readying evidence for disclosure only begins at a late stage.[169] For instance, in *Yekatom and Ngaïssona*, Pre-Trial Chamber II

164 ICC, *Prosecutor v. Thomas Lubanga Dyilo*, Case No. ICC-01/04-01/06, Redacted Decision on the Prosecution's Disclosure Obligations Arising Out of an Issue Concerning Witness DRC-OTP-WWWW-0031 (Trial Chamber I), 20 January 2011, para. 16.

165 IER Report, above note 92, para. 481.

166 See, for example, ICC, *Prosecutor v. Alfred Yekatom and Patrice-Edouard Ngaïssona*, Case No. ICC-01/14-01/18, Prosecution's Request to Vary the Decision on Disclosure and Related Matters (ICC-01/14-01/18-64-Red) (Pre-Trial Chamber II), 20 March 2019, para. 7.

167 ICC, *Prosecutor v. Abdallah Banda Abakaer Nourain and Saleh Mohammed Jerbo Jamus*, Case No. ICC-02/05-03/09 OA 4, Judgment on the Appeal of Mr Abdallah Banda Abakaer Nourain and Mr Saleh Mohammed Jerbo Jamus against the Decision of Trial Chamber IV of 23 January 2013 Entitled "Decision on the Defence's Request for Disclosure of Documents in the Possession of the Office of the Prosecutor" (Appeals Chamber), 28 August 2013, para. 42.

168 ICC, *Prosecutor v. Jean-Pierre Bemba Gombo et al.*, Case No. ICC-01/05-01/13, Defence Request for Leave to Reply to the Prosecution's Response to Bemba's "Consolidated Request for Disclosure and Judicial Assistance", ICC-01/05-01/13-2236-Conf, 6 October 2017, ICC-01/05-01/13-2236-Conf-Corr, 10 October 2017 (Appeals Chamber), 12 October 2017, para. 10.

169 IER Report, above note 92, para. 480 ("It was submitted that during the confirmation stage the Prosecutor does not commence redaction and disclosure until the Chamber first adopts a redaction protocol").

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

permitted the Prosecution to redact sensitive information in the evidence, and indicate in a chart whether the item contained exculpatory, incriminatory or Rule 77 material.[170] A year later, in the *Ali Kushayb* case, the same bench of judges departed from this approach (and from normal practice) and required the Prosecution to "mark the relevant sections of documents, statements and transcripts as [exculpatory], [incriminatory], [or Rule 77 material], or other or provide the relevant information by indicating page and paragraph numbers in a dedicated metadata field"; this added an immense amount of complexity and manual work to the disclosure process.[171] The unpredictability of disclosure practices, even by the same judges, means that the Prosecution can only begin applying redactions once a redactions and disclosure protocol is put into place, or risk having to do the work all over again if the Chamber decides to adopt a new approach – even if before the same bench.

- The ICC's technology is relatively antiquated,[172] meaning that the process of reviewing and demarcating items for disclosure is largely manual despite the Court's chronic lack of personnel and budgetary limitations.

All of these issues are exacerbated by ML and big data biases. The obligation to "investigate incriminating and exonerating circumstances equally" and to seek the truth, if done correctly, means that the Prosecution will hopefully expand its collection practices to account for any biases and to include information that might traditionally be missed. More data means more information to review, as the obligation to turn over exculpatory and impeachment evidence means that prosecutors need to search through all digital media in their possession, including forensic examination reports. It also means more sources for data. As highlighted by the Independent Expert Review in a recent auditing of the ICC, "dealing with disclosure has become increasingly difficult with the proliferation of material relating to events that are the subject of the Court's trials".[173] This will only become more difficult in our "big data" world, and as the ICC considers relying more on ML models and big data analytics. Confronting these issues now, with more practical and consistent rules of disclosure and the incorporation of more sophisticated technologies, becomes all the more essential.

170 ICC, *Prosecutor v. Alfred Yekatom and Patrice-Edouard Ngaïssona*, Case No. ICC-01/14-01/18, Public Redacted Version of "Decision on Disclosure and Related Matters" (Pre-Trial Chamber II), 23 January 2019. See also ICC, *Prosecutor v. Alfred Yekatom and Patrice-Edouard Ngaïssona*, Case No. ICC-01/14-01/18, Prosecution's Communication of the Disclosure of Evidence (Pre-Trial Chamber II), 31 July 2019.

171 ICC, *Prosecutor v. Ali Muhammad Ali Abd-Al-Rahman ("Ali Kushayb")*, Case No. ICC-02/05-01/20, Second Order on Disclosure and Related Matters (Pre-Trial Chamber II), 2 October 2020, para. 24; ICC, *Prosecutor v. Ali Muhammad Ali Abd-Al-Rahman ("Ali Kushayb")*, Case No. ICC-02/05-01/20, Prosecution's Third Progress Report on the Evidence Review, Translation and Disclosure Process (Pre-Trial Chamber II), 9 October 2020, para. 25 (noting "that this order will substantially increase the time required for the primary and secondary review of items for disclosure, especially in relation to lengthy documents, such as interview transcripts").

172 IER Report, above note 92, paras 577–584.

173 *Ibid*., para. 479.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Conclusion

US Supreme Court Justice Anthony Kennedy once explained that "[b]ias is easy to attribute to others and difficult to discern in oneself".[174] Unconscious bias manifests in judgments and behaviours towards others that we are not aware of. In organizations, these biases can take hold in the form of systems, structures, policies and practices, making the cycle difficult to break. The same is true in ML models and big data analytics that draw upon and make conclusions based on information which may be impacted by common human biases.

That said, the increased use of ML and big data analytics can be incredibly beneficial. In the realm of IHL, ML models have the potential to make weapons, targeting systems and military decisions more informed and more likely to reduce the prospect of civilian casualties. A properly developed ML model, as discussed above with CDEM, has the potential to calculate collateral damage to civilians at a much more advanced level than humans, thereby reducing the prospect of unnecessary harm to civilians and civilian property. Similarly, within the ICL context, many of the ICC's problems – inconsistent judicial making, the slow pace of review for disclosure, inefficient investigations, ineffective investigations due to lack of in-country access –are ones for which ML models have answers and, in some cases, may be the only solution. Indeed, ML models and data analytics could also serve as a "double net" in catching the very human biases that might otherwise influence them, serving as a safeguard against those biases.

For these reasons, it would be wise to start grappling with the issues that could arise when ML and big data analytics become more prevalent, the most important being the potential impact of common human biases. This article seeks to assist in the development of those rules and practices by addressing the challenge of biases in ML models and big data analytics, highlighting the potential issues that could arise and the attendant legal implications. From an IHL standpoint, this means more robust recommendations by international actors, like the ICRC, on acknowledging, assessing and seeking to mitigate the effect of these biases and in considering constraints or methods in the use of ML models in weapons systems or military decision-making. It also means the development and integration of rules and recommendations in military manuals (similar to DoD Directive 3000.09, discussed above), which are routinely audited and tested. Equally, ICL institutions must more proactively engage with issues in order to ready themselves for the challenges of addressing biases in ML and big data. That includes the need for ongoing judicial education, and the integration of principles and practices in the OTP's operations manual.

Without these proactive steps, ML models and big data analytics, however attractive they are and however well-intentioned the institutions using them might be, will likely perpetuate common human biases and result in the victimization of traditionally vulnerable and under-represented communities. In this regard, legal implications clearly arise, including fundamental principles under ICL and IHL.

174 US Supreme Court, *Williams v. Pennsylvania*, 136 S. Ct. 1899, 9 June 2016, p. 1905.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press

Institutions that are typically reticent towards public scrutiny must learn to feel comfortable being open to being tested, audited, criticized and examined. A process of examination and re-examination is likely the only way to ensure that a machine-driven world that is built on our fragilities and flaws is made somewhat fair, and somewhat just.

https://doi.org/10.1017/S1816383121000096 Published online by Cambridge University Press