Volume 102 Number 913

International Review of the Red Cross

Humanitarian debate: Law, policy, action

Digital technologies and war

ICRC

International
Review_
of the Red Cross

Aim and scope

Established in 1869, the *International Review of the Red Cross* is a peer-reviewed journal published by the ICRC and Cambridge University Press. Its aim is to promote reflection on humanitarian law, policy and action in armed conflict and other situations of collective armed violence. A specialized journal in humanitarian law, it endeavours to promote knowledge, critical analysis and development of the law, and contribute to the prevention of violations of rules protecting fundamental rights and values. The *Review* offers a forum for discussion on contemporary humanitarian action as well as analysis of the causes and characteristics of conflicts so as to give a clearer insight into the humanitarian problems they generate. Finally, the *Review* informs its readership on questions pertaining to the International Red Cross and Red Crescent Movement and in particular on the activities and policies of the ICRC.

Editorial Team
Editor-in-Chief: Bruno Demeyere
Thematic Editor: Saman Rejali
Editorial Team: Sai Venkatesh and Ash Stanley-Ryan
Book review editor: Jamie A. Williamson
Special thanks: Neil Davison, Laurent Gisel, Netta Goussac, Massimo Marrelli, Ellen Policinski, Mark David Silverman, Delphine van Solinge

International Review of the Red Cross
19, Avenue de la Paix, CH 1202 Geneva
CH - 1202 Geneva
t +41 22 734 60 01
e-mail: review@icrc.org

Cover Photo: In eastern Ghouta rubble, a father looks for his son. Credit: REUTERS/ Bassam Khabieh

CONTENTS: Volume 102 Number 913

Artificial intelligence, autonomous weapon systems and their governance

Cyber operations and warfare

Selected articles

Books and articles

Reports and documents

International Review of the Red Cross (2020), **102** (913), 1–22.
Digital technologies and war
doi:10.1017/S1816383121000114

IRRC

EDITORIAL

THE ROLE OF DIGITAL TECHNOLOGIES IN HUMANITARIAN LAW, POLICY AND ACTION: CHARTING A PATH FORWARD

Saman Rejali and Yannick Heiniger*

Why does "Digital Technologies and War" deserve the thematic focus of an entire issue of the *International Review of the Red Cross*? Contributors to this issue highlight two overarching reasons.

First, digitalization is a priority topic for humanitarian organizations like the International Committee of the Red Cross (ICRC)[1] because it is rapidly shaping how humanitarian operations and assistance activities are carried out, impacting how the humanitarian sector is serving affected populations.

In this respect, one group of contributors to this issue (with their pieces appearing under the headers "Humanitarian Action in the Digital Age" and "Business and Digital Technologies in Humanitarian Crises") analyze how the use of digital technologies for delivering humanitarian relief brings forth both unparalleled opportunities and risks. They highlight that some digital technologies, including those which ease communication and service delivery, are vital tools for the humanitarian sector. Those technologies help to ensure that the humanitarian sector brings solutions to crisis contexts and continues to serve affected populations. For example, increased connectivity and digital access can empower affected people in armed conflicts and other situations of violence to connect with others via messaging applications and social media platforms, to find information online and to express their needs via rapid feedback mechanisms, serving as active agents working with humanitarians.[2] Furthermore, digitally rooted contextual analyses, crisis mapping, and digitalized services can allow humanitarians to more efficiently serve affected people, and to predict and respond to humanitarian crises.

On the other hand, these contributors also stress how there are certain risks and considerations that go hand in hand with the opportunities brought forth in using digital technologies for humanitarian relief efforts. Accounting for and mitigating these risks is how the humanitarian sector can ensure that it is well prepared as it embarks on digital transformation processes. One dominant risk factor is data protection and privacy: collecting data from affected populations

* Saman Rejali is a Law and Policy Adviser at the ICRC and was Thematic Editor at the ICRC for this issue of the *Review* on "Digital Technologies and War". Yannick Heiniger is Deputy CEO of Swissnex San Francisco and was previously Partnerships Manager within the Office of the Director for Digital Transformation and Data at the ICRC.

puts an undeniable onus on humanitarian organizations to ensure that affected people's data is not misused and does not put them in harm's way, contrary to the purpose for which it was collected. Moreover, the ways in which data and information are disseminated are shaping how conflicts and other situations of violence unfold, in contexts such as Myanmar.[3] Misinformation, disinformation and hate speech (MDH), and the "new realities" presented through deepfakes (using machine learning to generate synthetic video, audio and text content),[4] are among the new risks that have come together with the widespread use of social media platforms and other online dissemination tools. As such, one section of this issue is dedicated to the role of businesses – more specifically, technology companies, which have been increasingly involved in supporting the work of humanitarian organizations, while concurrently being indirectly involved in the conduct of hostilities through the way their technologies are used.

There is also a second reason why digitalization matters for the humanitarian sector. Digital technologies (or "new" technologies) are used in armed conflicts as a means and method of warfare, governed by international humanitarian law (IHL). The uses of these new technologies have humanitarian consequences as they are giving rise to unprecedented means and methods of warfare. For example, as Gisel, Rodenhäuser and Dörmann point out in their article for this issue, cyber operations against electricity grids, health-care systems, nuclear facilities or other critical infrastructure could cause "significant human harm", with catastrophic humanitarian consequences.[5] Besides cyber threats, autonomous weapon systems (AWSs), including those enabled by artificial intelligence (AI), also raise humanitarian, legal and ethical concerns, as they select and apply force to targets without human intervention, meaning that the user does not know the specific target that will be struck, nor where or when.[6] In the third and fourth parts of this issue – focusing on AI and AWSs, and cyber operations and warfare, respectively – the contributors take different stances and

1 In the case of ICRC, "Embracing the Digital Transformation" is the fifth pillar of the organization's Institutional Strategy for 2019–22, available at: www.icrc.org/en/publication/4354-icrc-strategy-2019-2022 (all internet references were accessed in January 2021).

2 A look at the cover of this issue reflects this reality: in Syria, where emergency and long-term needs are great following years of violence, a father holds a smartphone in his hand, a photo of his son displayed on-screen, surrounded by the destruction created by the armed conflict. See "In Eastern Ghouta Rubble, a Father Looks for His Son", *Reuters*, 4 March 2018, available at: www.reuters.com/article/us-mideast-crisis-syria-ghouta-victims-idUSKBN1GG0EJ.

3 Alexandra Stevenson, "Facebook Admits It Was Used to Incite Violence in Myanmar", *New York Times*, 6 November 2018, available at: www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

4 Aengus Collins, *Forged Authenticity: Governing Deepfake Risks*, EPFL International Risk Governance Center, 2019, available at: https://infoscience.epfl.ch/record/273296?ln=en.

5 See Laurent Gisel, Tilman Rodenhäuser and Knut Dörmann, "Twenty Years On: International Humanitarian Law and the Protection of Civilians against the Effects of Cyber Operations during Armed Conflicts", in this issue of the *Review*.

6 See Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*, SIPRI and ICRC, June 2020, covered in the "Reports and Documents" section of this issue of the *Review*; Frank Sauer, "Stepping Back from the Brink: Why Regulating Autonomy in Weapons Systems is Difficult, Yet Imperative and Feasible", in this issue of the *Review*.

offer varying perspectives exploring how digital technologies are used in warfare, assessing the application of IHL to cases where digital technologies are used for destructive purposes.

"New" technologies, however, are ever evolving with the latest advancements. Just as the telegraph – now all but gone – was a "communication game changer"[7] two centuries back, some of the current "new" technologies will one day no longer be relevant, let alone "new", and perhaps the risks and opportunities relating to data protection and MDH will be moot. However, there are a series of "timeless" themes relating to digital technologies and humanitarian law, policy and action, which will stand the test of time, and are highlighted in the discussion that follows.

We[8] believe that one key common thread across these contributions is that of trust: principled humanitarian action is rooted in trust, and humanitarians have a responsibility to gain the trust of the affected populations they aim to serve.[9] While digital technologies offer unparalleled opportunities for providing humanitarian relief, they must be used ethically and responsibly in order to minimize the risks outlined in these pages. It is only by so doing that humanitarians can hope to gain the trust of the affected populations to whom they are accountable.

Along with trust comes ethics: operating in a way that does justice to the people we serve, ensuring that the benefits derived from digital technologies outweigh their risks, and ensuring that we work *with* affected people and do not decide for them on the issues that shape their lives. Ethical frameworks also apply to the means and methods of warfare. For example, when it comes to the uses of AI, Pizzi, Romanoff and Engelhardt[10] of UN Pulse, the UN Secretary-General's initiative on big data and AI, point to how ethical frameworks are necessary when regulating AI, but are not always sufficient in organizational structures where an "ethics-first approach" often does not go hand in hand with robust accountability mechanisms.

The authors featured in this issue also highlight the ethical considerations and potential inclusivity barriers to "humanitarian innovation". Humanitarian innovation has the possibility to open up new ways for us to serve affected people, but if innovative projects don't take into account data and personal information protection measures, and if they're created without truly being centred around and inclusive of affected people, then the risks they pose may outweigh the benefits.[11] In such cases, the "product" can outpace the due

7 Jimmy Stamp, "How the Telegraph Went from Semaphore to Communication Game Changer", *Smithsonian Magazine*, 11 October 2013, available at: www.smithsonianmag.com/arts-culture/how-the-telegraph-went-from-semaphore-to-communication-game-changer-1403433/.

8 The term "we" in this paper refers solely to the authors of this editorial and not to the ICRC or the humanitarian sector. The views expressed in this editorial reflect solely those of the authors and not those of the ICRC nor Swissnex San Francisco.

9 Hugo Slim, "Trust Me – I'm a Humanitarian", *Humanitarian Law and Policy Blog*, 24 October 2019, available at: https://blogs.icrc.org/law-and-policy/2019/10/24/trust-humanitarian/.

10 See Michael Pizzi, Mila Romanoff and Tim Engelhardt, "AI for Humanitarian Action: Human Rights and Ethics", appearing in this issue of the *Review*.

11 See ICRC, *Symposium Report: Digital Risks in Armed Conflicts*, Geneva, October 2019, covered in the "Reports and Documents" section of this issue of the *Review*.

diligence required to ensure that digital technologies cause more benefit than harm to affected populations. Indeed, this is a point echoed by Sandvik and Lohne,[12] who clearly identify that the problem is that "affected populations are often not present in innovation processes – they are neither properly consulted nor invited to participate". This potentially produces "new digital harms, whether these occur through (in)visibilizing the suffering of particular groups or individuals, creating undesirable consequences, or introducing new risks".

In what follows, we will first highlight how the contributions in this issue explore the benefits and risks of digital technologies used for humanitarian action, identifying the opportunities and mitigating measures to be taken as paths forward if we want to embark on a digitalization of humanitarian action, taking into account the growing role of the private sector. Thereafter, we will provide an overview of how digital technologies can be used as a means and method of warfare in armed conflict, making reference to contributions that explore themes of cyber operations and the application of IHL, as well as AWSs, machine learning and AI. This analysis is framed by outlining who we are, as two millennials with expertise on the theme at hand co-authoring this editorial, and concludes by addressing some contextual elements that impacted the publication of this issue of the *Review* and reflecting on the overall takeaway from the issue.

Millennial views on digital technologies and war

With this editorial, our purpose is to offer a cross-cutting glimpse into the diverse ideas that you will encounter throughout this issue of the *Review*. As two millennials who have been working on the broad theme of technologies in humanitarian action for a few years, we are admittedly part of the first generation of "digital natives".[13] We are supposed to be at ease with using and integrating new digital technologies into our daily lives. This is surely true when it comes to many aspects of our social lives: Facebook, Twitter, LinkedIn, TikTok and the likes are consuming a significant amount of our time, and digital interactions are perceived by many as an essential complement to physical ones.

As the ICRC's 2020 *Millennials on War* report highlights,[14] the perspective of millennials on the potential for digital technologies to help the people affected by war is quite positive. By exploring the impact and implications of digital technologies in armed conflicts and other situations of violence, our desire is that this issue of the *Review* will provide a "reality check" into the world that we, as millennials, are contributing to creating, highlighting how our actions in the

12 See Kristin Bergtora Sandvik and Kjersti Lohne, "The Struggle against Sexual Violence in Conflict: Investigating the Digital Turn", in this issue of the *Review*.

13 Berkman Klein Center for Internet and Society at Harvard University (BKC), "Digital Natives", available at: https://cyber.harvard.edu/research/youthandmedia/digitalnatives. Digital natives are defined by the BKC as "a generation 'born digital' – those who grow up immersed in digital technologies, for whom a life fully integrated with digital devices is the norm".

14 ICRC, *Millennials on War*, Geneva, 2020, available at: www.icrc.org/en/millennials-on-war.

humanitarian sphere have consequences for the affected people whom we aim to serve. We also recognize that we have our own biases as millennials. "Digital natives" tend to have a different relationship with the principles that are at the core of humanitarian practice – neutrality, impartiality, and independence in humanitarian action (NIIHA).[15] Digital technologies and algorithms are playing an important role in how we view the world.

When it comes to the principle of humanity, we not only recognize that "suffering is universal and requires a response"[16] – we also take an "activist" stance. We use the various social media channels at our disposal to galvanize action, both online and offline.[17] We are writing as two co-authors who have grown up mainly in the global North, and we recognize that our experiences are not universal to all millennials but are rather those of a subset of "global citizens" in the hubs of Geneva, London, New York, Toronto and Paris (among others).[18] Our commitment to making a difference and not being *indifferent* to suffering means that we have stuck it out through a financial crisis (and a second one at the time of writing); we have carried out internship after internship, one short-term contract after another, foregoing the stable lives of our parents[19] for the global lives we lead,[20] in order to land on our feet and pursue our vocation in the humanitarian sector. We are not easily deterred, which is what the world needs, as we are grappling with the risks that digital technologies can pose for humanitarian action and how they can be misused in warfare. Our world is not exclusively State-driven – it is multipolar, with "an ever-increasing number" of non-State armed groups,[21] using digital technologies as a means of advancing their aims in armed conflicts.[22] Moreover, we've grown up with social media, but

15 In comparison, the Fundamental Principles of the International Red Cross and Red Crescent Movement (the Movement) include seven principles: humanity, impartiality, neutrality, independence, voluntary service, unity and universality. Jérémie Labbé and Pascal Daudin, "Applying the Humanitarian Principles : Reflecting on the Experience of the International Committee of the Red Cross", *International Review of the Red Cross*, Vol. 97, No. 897/898, 2016; Office of the UN High Commissioner for Refugees, "Humanitarian Principles", available at: https://emergency.unhcr.org/entry/44765/humanitarian-principles; ICRC, "The Fundamental Principles of the Red Cross and Red Crescent Movement", August 2015, available at: www.icrc.org/sites/default/files/topic/file_plus_list/4046-the_fundamental_principles_of_the_international_red_cross_and_red_crescent_movement.pdf.

16 ICRC, above note 15.

17 Emily Logan, "Millennial Activism: Tapping into a Network of Millennial Donors", available at: https://csic.georgetown.edu/magazine/millennial-activism-tapping-network-millennial-donors/.

18 April Rinne, "What Is Global Citizenship?", 9 November 2017, available at: www.weforum.org/agenda/2017/11/what-is-global-citizenship/.

19 Janet Adamy, "Millennials Slammed by Second Financial Crisis Fall Even Further Behind", *Wall Street Journal*, 9 August 2020, available at: www.wsj.com/articles/millennials-covid-financial-crisis-fall-behind-jobless-11596811470.

20 BKC, above note 13.

21 Jelena Nikolic, Tristan Ferraro and Thomas de Saint Maurice, "Aggregated Intensity: Classifying Coalitions of Non-State Armed Groups", *Humanitarian Law and Policy Blog*, 7 October 2020, available at: https://blogs.icrc.org/law-and-policy/2020/10/07/aggregated-intensity-classifying-coalitions-non-state-armed-groups/.

22 Delphine van Solinge, "Digital Risks for Populations in Armed Conflict: Five Key Gaps the Humanitarian Sector should Address", *Humanitarian Law and Policy Blog*, 12 June 2019, available at: https://blogs.icrc.org/law-and-policy/2019/06/12/digital-risks-populations-armed-conflict-five-key-gaps-humanitarian-sector/.

have increasingly grown critical of it as well.[23] This is particularly the case when looking specifically at global technology companies, and at how the recent coronavirus crisis has strengthened their hold on key aspects of our society, exacerbating the daily realities of many, including in humanitarian contexts.

As millennials, we advocate for an understanding of the principle of impartiality as being arguably the driving force towards genuine inclusion and diversity in the humanitarian sector.[24] As such, we are cognisant that while digital technologies can allow us to more easily connect with and reach out to affected people, they can also cause a digital divide, resulting in intersectional inequities in access to digital technologies and their associated benefits, and thereby putting certain affected populations at a disadvantage.[25] This digital divide is highlighted cogently by Jo Burton in her article for this issue of the *Review*, through the example of the digitalization of cash. As she notes,

> the increase in digital payments may deepen the "digital divide" …. Any person can use cash if they can get their hands on it, and providing the goods and services they need to pay for are available. … However, to use digital payments, the recipient will require a level of digital and financial literacy. It is estimated that only one in three adults globally shows an understanding of basic financial concepts, and that there are lower standards of financial literacy amongst women and the poor.[26]

As Burton's analysis highlights, intersectional inequity – including financial and gender-based inequity – deepens the digital divide. However, we believe that if we take action that embodies the impartiality principle, we can address the systemic inequities that hinder humanitarian response.

With regard to the neutrality principle, as millennials we have come to realize, through our first-hand experience with many a misinformation and disinformation campaign on social media, that because of the way certain people make use of digital technologies, such technologies are not necessarily neutral. This is clearly illustrated by the use of digital technologies for destructive means during armed conflicts and other situations of violence.

Altogether, our millennial views on NIIHA shape how we view, analyze and work with digital technologies in humanitarian crises and in the context of

23 Nick Statt, "Facebook's US User Base Declined by 15 Million since 2017, According to Survey", *The Verge*, 6 March 2019, available at: www.theverge.com/2019/3/6/18253274/facebook-users-decline-15-million-people-united-states-privacy-scandals; Jack Nicas, Mike Isaac and Sheera Frenkel, "Millions Flock to Telegram and Signal as Fears Grow over Big Tech", *New York Times*, 13 January 2021, available at: www.nytimes.com/2021/01/13/technology/telegram-signal-apps-big-tech.html.

24 Saman Rejali, "Race, Equity, and Neo-Colonial Legacies: Identifying Paths Forward for Principled Humanitarian Action", *Humanitarian Law and Policy Blog*, 16 July 2020, available at: https://blogs.icrc.org/law-and-policy/2020/07/16/race-equity-neo-colonial-legacies-humanitarian/.

25 Barnaby Willitts-King, John Bryant and Kerrie Holloway, *The Humanitarian "Digital Divide"*, Humanitarian Policy Group Working Paper, Overseas Development Institute, London, November 2019, p. 15; Lina Gurung, "The Digital Divide: An Inquiry from Feminist Perspectives", *Dhaulagiri Journal of Sociology and Anthropology*, Vol. 12, 2018.

26 See Jo Burton, "'Doing No Harm' in the Digital Age: What the Digitalization of Cash Means for Humanitarian Action", in this issue of the *Review*.

evaluating the application of IHL to the means and methods of warfare. The outlook on digital technologies provided in these pages gave us a unique opportunity to (re) think about and broaden our reflections on digital technologies and war, and the broader purpose that digital technologies serve for humanitarian action. We hope this will be the case for every reader, beyond generational boundaries.

Digital technologies and humanitarian action: The risks

A collection of the contributions in this issue highlight how digital technologies can be used without full awareness of what they will trigger. Thus, digital technologies pose certain risks within conflicts – societal, economic, political and cognitive – which must be accounted for in humanitarian assistance and operational activities. The most problematic issue, perhaps, is that benefit/risk assessments are often carried out solely by humanitarians, and not *with* affected people. Furthermore, those most affected by such risks are individuals and communities in crisis contexts.

There are three main vectors for risk[27] identified with respect to humanitarian action: (1) digital surveillance, monitoring and intrusion; (2) MDH; and (3) the misuse and mishandling of data and personal information.

Digital surveillance, monitoring and intrusion

The risks associated with digital surveillance, monitoring and intrusion can come from various sources, including big data analyses, machine learning models, misuse of data by authorities, and as a consequence of people's online presence and activities. As Gazi and Gazis[28] point out in their contribution to this issue, big data and open data analyses not only entail privacy risks but may also produce biased results. The latter is due to the fact that big data and open data

> often lack demographic information that is crucial for epidemiological research, such as age and sex. [Also], this data represents only a limited portion of the population – i.e., excluding marginalized and under-represented groups such as infants, illiterate persons, the elderly, indigenous communities and people with disabilities – while potentially under-representing some developing countries where digital access is not widespread.

This is particularly problematic for humanitarian assistance and protection activities, since big data and open data analytics can lead humanitarians to inadvertently ignore the marginalized people, standing at several intersections of

27 According to the ICRC's *Digital Risks in Armed Conflicts* Symposium Report, digital risks "include (often unintended) side-effects of digital data experimentation, privacy violations, and the mishandling of sensitive information that accompanies the humanitarian sector's efforts to deploy emerging technologies in already fragile contexts". ICRC, above note 11.

28 See Theodora Gazi and Alexandros Gazis, "Humanitarian Aid in the Age of COVID-19: A Review of Big Data Crisis Analytics and the General Data Protection Regulation", in this issue of the *Review*.

inequity, whom they aim to serve. This is reaffirmed by Milaninia[29] in his analysis, which illustrates how machine learning models and big data analytics are "highly susceptible to common human biases" and can thereby "accelerate existing racial, political or gender inequalities" and potentially paint "a misleading and distorted picture of the facts on the ground".

Similarly, Pizzi, Romanoff and Engelhardt[30] illustrate that a lack of quality data increases the risks that an AI system will produce unfair outcomes, as

> AI systems can reveal sensitive insights into individuals' whereabouts, social networks, political affiliations, sexual preferences and more, all based on data that people voluntarily post online (such as the text and photos that users post to social media) or incidentally produce from their digital devices (such as GPS or cell-site location data).

Once this data is collected, it is highly susceptible to misuse, if necessary data protection measures are not taken. Most dangerously, through their online behaviour, affected populations can unknowingly be subjecting themselves to potential offline harm, including but not limited to being surveilled and profiled in crisis contexts,[31] and facing the threat of violence, hate crimes and/or discrimination.[32] A real case scenario of such surveillance is provided in the ICRC's Symposium Report on *Digital Risks in Armed Conflicts*, featured in the "Reports and Documents" section of this issue, whereby Syrian refugees' mobile devices were compromised through a malware attack. Other examples show surveillance occurring by humanitarians themselves as they use technologies to better respond to needs – for example, via drone usage for mapping and risk assessment purposes.[33] Here, the aforementioned risks of surveillance particularly apply as the drones may gather information from contexts where affected populations live, without their consent and/or knowledge. More sophisticated illustrations of surveillance, monitoring and intrusion can be found, for instance, in the article by Siatitsa,[34] which discusses such issues in relation to facial recognition.

Misinformation, disinformation and hate speech

Speaking about MDH in the Q&A conducted for this issue of the *Review*,[35] Delphine van Solinge unpacks how through MDH, information can be manipulated and

29 See Nema Milaninia, "Biases in Machine Learning Models and Big Data Analytics: The International Criminal and Humanitarian Law Implications", in this issue of the *Review*.
30 M. Pizzi, M. Romanoff and T. Engelhardt, above note 10.
31 J. Burton, above note 26.
32 ICRC, above note 11.
33 Faine Greenwood, "Data Colonialism, Surveillance Capitalism and Drones", in Doug Specht (ed.), *Mapping Crisis: Participation, Datafication and Humanitarianism in the Age of Digital Mapping*, University of London Press, London, 2020.
34 See Ilia Siatitsa, "Freedom of Assembly under Attack: General and Indiscriminate Surveillance and Interference with Internet Communications", in this issue of the *Review*.
35 See "Q&A: Humanitarian Operations, the Spread of Harmful Information and Data Protection", in this issue of the *Review*.

spread using digital technologies – particularly amidst the coronavirus pandemic, when populations are more reliant on digital communication technologies. An example of a disinformation tactic is the creation of "deepfakes", using machine learning to generate fake video and audio content.[36] In crisis contexts such as Myanmar, South Sudan and Ethiopia,[37] MDH are disseminated via social media platforms, and public opinion is manipulated based on false or incomplete information, exacerbating the humanitarian crises at hand. The use of technologies by mass publics gives increasing power to the companies behind social media, messaging and search platforms, including in armed conflicts and other situations of violence. We have seen recently how big technology firms have had reverberating global effects in being arbiters between freedom of speech, on the one hand, and social media accounts being used for the spreading of harmful information (i.e., MDH), on the other. This is more the case now, during the coronavirus pandemic, as affected populations are more reliant than ever before on such platforms to receive information and communicate with each other.

The misuse and mishandling of data and personal information

When it comes to the misuse and mishandling of data, the concept of "technocolonialism", coined by Mirca Madianou,[38] serves as an excellent guiding light for what can go wrong, even with the best of intentions, if we strive for digital innovation and aggregate biometric data in humanitarian crises without putting in place the necessary data protection practices and digitally tailored protection frameworks. Indeed, technologies integrate and reinforce the value systems, cultures and world views of their builders. Uninhibited digital innovation and data practices can further ingrain the colonially rooted power asymmetries between humanitarian actors and affected populations.[39]

This is reflected in the occurrence of "surveillance capitalism", described by Zuboff as "data from humans used to turn a profit, at the expense of the people themselves".[40] In the context of humanitarian crises, this means that data from affected populations can be not only collected but also used for profit. As this collection of data often happens without the knowledge of the affected person, Zuboff draws a parallel with colonial practices of extraction without permission. In this respect, Sandvik and Lohne note how the ramifications of such uninhibited gathering of data, and the recording of affected populations'

36 Harvard Kennedy School, Belfer Centre for Science and Information Affairs, "Tech Factsheets for Policymakers: Deepfakes", Spring 2020, available at: www.belfercenter.org/sites/default/files/2020-10/tappfactsheets/Deepfakes.pdf.
37 "Q&A", above note 35.
38 Mirca Madianou, "Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises", *Social Media + Society*, Vol. 5, No. 3, 2019, available at: https://journals.sagepub.com/doi/full/10.1177/2056305119863146.
39 *Ibid.*
40 Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future*, PublicAffairs, New York, 2019.

information on digital clouds, can create "digital bodies" with gendered ramifications for how we address conflict-related sexual violence.[41]

The after-effects of what happens in the humanitarian sector when data protection measures are not implemented properly are highlighted by Massimo Marelli in this issue's Q&A,[42] and by Burton, who applies the "do no (digital) harm" principle of data protection to the digitalization of cash. Burton highlights how metadata – data that provides information about other data – can have grave consequences for humanitarian crises and can be used for military gain, particularly when we hear influential people such as General Hayden, former director of the US National Security Agency and Central Intelligence Agency, quoted by Burton, saying: "We kill people based on metadata."[43]

The issue of what happens to affected populations' data once collected by humanitarian organizations is critical. Large technology companies have on several occasions shared users' data with governments, which can pose security risks to citizens living in armed conflicts or other situations of violence.[44] It is worth noting that even when affected people's data is not shared, the stored data may be hacked into or stolen if not well protected by humanitarian organizations.[45] This data protection risk is highlighted in the ICRC's Symposium Report[46] as well: "Humanitarian organizations collect, store, share, and analyze data that is attractive to parties to armed conflict. … As a result, humanitarian organizations are exposed to a growing wave of digital attacks and cyber espionage, and have become highly prized targets."

Digital technologies and humanitarian action: Mitigating measures and benefits

To account for these digital risks, which have societal, economic, political and cognitive consequences for affected populations and humanitarian crises, there are several active steps that can be taken, including (1) fostering digital literacy, (2) strengthening data protection practices and creating the right safeguards for the adoption of digital technologies, and (3) adopting suitable humanitarian policies, ensuring humanitarians continue to put people at the centre of their work.

Fostering digital literacy

Digital literacy is not just a "nice to have" component for humanitarian organizations; it is a crucial necessity for affected populations. This important observation emerges

41 K. B. Sandvik and K. Lohne, above note 12.
42 "Q&A", above note 35.
43 J. Burton, above note 26.
44 *Ibid.*; F. Greenwood, above note 33.
45 The importance of data protection measures in humanitarian action is highlighted in Christopher Kuner and Massimo Marelli (eds), *Handbook on Data Protection in Humanitarian Action*, 2nd ed., ICRC and Brussels Privacy Hub, Geneva, June 2020, covered in the "Reports and Documents" section of this issue of the *Review*.
46 ICRC, above note 11.

several times throughout this issue of the *Review*. Van Solinge, for instance, advocates for increasing people's resilience to misinformation and disinformation by "promoting digital literacy, critical thinking and … humanitarian values".[47] Sandvik and Lohne, for their part, highlight how digital literacy must go "beyond technical competence to include awareness and perceptions about technology, law, rights and risk".[48] These elements are also crucial for humanitarian organizations themselves. As an example of ongoing initiatives aimed at giving humanitarian decision-makers and lawyers crucial digital literacy skills, the ICRC has now partnered with the Swiss Federal Institute of Technology Lausanne (Ecole Polytechnique Fédérale de Lausanne, EPFL) on a series of collaborative initiatives,[49] including one to create a five-day introductory course on information and communication technology fundamentals.[50] Along the same lines, the International Federation of Red Cross and Red Crescent Societies (IFRC) has developed a Data Playbook Project, aiming to "improve data literacy across teams, sectors, the IFRC Secretariat, and within National Societies".[51] While these concrete projects focus on humanitarian actors themselves, they are a first step into a new territory for many humanitarian organizations, and call for similar initiatives at the field level focusing on affected populations and their digital literacy.

Strengthening data protection practices

Alongside digital literacy skills, appropriate data protection practices can ensure that unwanted access to affected populations' data – through surveillance, monitoring or breach of digital storage solutions – is prevented, serving as another mitigating measure for digital risk. In this respect, for example, Massimo Marelli highlights how the ICRC has

> now adopted several specific safeguards in [its] Rules on Personal Data Protection, adopted in 2015, which are designed to reduce the risk of unauthorized use or access to personal data by applying data protection standards and requirements to data processing throughout the organization. Where new technologies or riskier data processing operations are considered by the ICRC, a Data Protection Impact Assessment must be conducted to identify and mitigate the risks of harm. The Rules also require the ICRC to follow a "data protection by design" approach to minimize the collection of

47 "Q&A", above note 35.
48 K. B. Sandvik and K. Lohne, above note 12.
49 The EPFL and ETH Zürich are joining forces with the ICRC to explore innovative solutions to today's humanitarian crises, through the HAC initiative: see EPFL, "Science and Technology for Humanitarian Action Challenges (HAC)", available at: www.epfl.ch/research/services/fund-research/funding-opportunities/research-funding/science-and-technology-for-humanitarian-action-challenges-hac/. See also EPFL, "EPFL, ETH Zurich and the ICRC Team Up to Bolster Humanitarian Aid", 10 December 2020, available at: https://actu.epfl.ch/news/epfl-eth-zurich-and-the-icrc-team-up-to-bolster-hu/.
50 EPFL, "Executive Training: Foundations of Information and Communication Technologies", available at: www.c4dt.org/event/fict-executive-course/.
51 IFRC, "Discover the Data Playbook Beta Project", 18 October 2018, available at: https://media.ifrc.org/ifrc/2018/10/18/discover-data-playbook-beta-project/.

personal data to that which is necessary for the operation and ensure that data subjects' rights are respected.[52]

Humanitarian policy as an enabler for the responsible use of technologies

Among the resources available to humanitarians in this process of balancing opportunities and risks, policy emerges as a unique enabler. As we look for organizations engaging in digital transformation processes while ensuring that digital risks are mitigated, a few interesting examples come to mind, which complement what you will find in this issue of the *Review*. One of the key resolutions of the 33rd International Conference of the Red Cross and Red Crescent (International Conference) in 2019 was Resolution 4 on "Restoring Family Links while Respecting Privacy, Including as it Relates to Personal Data Protection".[53] This resolution calls on States and the International Red Cross and Red Crescent Movement (the Movement) to respect numerous privacy and data protection stipulations when processing the information of affected populations. In particular, it "urges States and the Movement to cooperate to ensure that personal data is not requested or used for purposes incompatible with the humanitarian nature of the work of the Movement."[54] The data protection stipulations that underpin this resolution are embodied within the *Restoring Family Links Code of Conduct on Data Protection*.[55] This code of conduct sets out the minimum principles, commitments and procedures that personnel of the ICRC, National Societies and the IFRC must comply with when processing personal data within the framework of Restoring Family Links activities. Such documents can ensure that humanitarians have a common understanding of the inherent risks and common measures needed to make sure technologies work in a way that reinforces the protection of the sensitive data of individuals in conflict zones.

In fact, the 2019 International Conference also yielded a Movement-wide Digital Pledge on "Strengthening National Digital and Data Capacities for Humanitarian Action",[56] whereby the Movement committed to an action plan by the end of 2023 to (1) foster partnerships in this respect; (2) convene regarding these issues; and commit to (3) digital literacy, (4) digital inclusion, (5) data protection and (6) digital responsibility. This example is another illustration of the importance of embracing a principled digital transformation journey and aligning visions around the measures needed to mitigate the adverse effects of

52 "Q&A", above note 35.

53 ICRC, "Restoring Family Links while Respecting Privacy, Including as it Relates to Personal Data Protection", 33IC/19/R4, Resolution 4 adopted at the 33rd International Conference of the Red Cross and Red Crescent, Geneva, 9–12 December 2019, available at: https://rcrcconference.org/app/uploads/2019/12/33IC-R4-RFL-_CLEAN_ADOPTED_en.pdf.

54 *Ibid.*, para. 11.

55 ICRC, *Restoring Family Links Code of Conduct on Data Protection*, Geneva, November 2015, available at: www.icrc.org/en/document/rfl-code-conduct.

56 "Strengthening National Digital and Data Capacities for Humanitarian Action", Digital Pledge, 2019 International Conference, available at: https://tinyurl.com/110x3pmp.

digital technologies. It also shows how the Movement can lead by example in translating the Movement-wide Fundamental Principles[57] to the use of such technologies.

Still in the domain of humanitarian policy, in 2019 the ICRC also produced a policy stance on its use of biometric technologies,[58] which are used in forensics and the restoration of family links. Given the sensitivity of creating a permanent record for individuals who may not want to be identifiable forever, this policy facilitates a responsible use of the technology by the organization and addresses the data protection challenges this poses. Overall, these multiple initiatives illustrate well the role that humanitarian policy can play in creating an actionable framework enabling the principled use of new digital technologies.

Benefits of digital technologies for humanitarian action

While digital technologies pose certain aforementioned risks, they also bring unparalleled benefits for how humanitarian operational and assistance activities are carried out. This is exemplified in this issue of the *Review* through the "Voices and Perspectives" section, which features testimonies collected from affected populations.

This section presents direct quotes from individuals whose lives have changed for the better because of the digitally driven initiatives heralded by the ICRC. One testimony focuses on the ICRC's Trace the Face platform,[59] which is an "online photo gallery with thousands of pictures of people looking for their family." It was through this site that Matty, based in Abidjan, was able to find his uncle, whom he'd had no news of since the outbreak of the 2010–11 crisis in Ivory Coast.[60]

Similarly highlighting the positive potentials for digital technologies in humanitarian crises, the other featured testimony is from Zawadi, who talks about how she was able to connect with her husband's family through the Electronic Red Cross Messaging initiative, which is a collaborative pilot project between the ICRC, the Congolese Red Cross and the Rwandan Red Cross.[61] The pilot project began in November 2018 and uses digital Red Cross messages to re-establish links between separated family members. As a part of the project, Red Cross volunteers roam the villages of the eastern Democratic Republic of the Congo and Rwanda with digital tablets connected to the Internet. The project shows great promise as it has improved one of the oldest services at the ICRC, the Red Cross messaging system, managing to facilitate the work of restoring

57 ICRC, above note 15.
58 ICRC, "The ICRC Biometrics Policy", 16 October 2019, available at: www.icrc.org/en/document/icrc-biometrics-policy.
59 ICRC, "Trace the Face – Migrants in Europe", available at: https://familylinks.icrc.org/europe/en/pages/publish-your-photo.aspx.
60 See "How Humanitarian Technologies Impact the Lives of Affected Populations", in the "Voices and Perspectives" section of this issue of the *Review*.
61 *Ibid.*

links between families in a faster and more efficient manner than before. Such initiatives, as seen through the testimonies of affected people, serve as examples of what is possible when humanitarian innovation merges with digital technologies with the aim of alleviating human suffering in armed conflicts and other situations of violence. Building on this momentum, the ICRC is piloting Red Safe,[62] a digital platform allowing affected populations to access a variety of services digitally.

In her Q&A with the *Review*, Delphine van Solinge also highlights how humanitarian practitioners have "made use of the enhanced situational awareness and actionable information afforded by the digital age". For example, as she points out, human rights defenders and humanitarian practitioners have

> employed remote sensing tools for augmenting conflict early warning capacities and documenting human rights abuses. They have leveraged mobile data solutions for tracking the conditions, profiles and routes of transit of migrant and refugee populations; exploited metadata from call detail records to understand the spread of infectious diseases; harvested social media for sentiment analysis and rumour tracking in fragile contexts; and of course, they've deployed aerial robotics for surveillance of damaged locations and monitoring critical infrastructure.
>
> In the case of COVID-19, digital tools, artificial intelligence and "big data" analysis are being used in various contexts to support health-based responses. They can help us collect, analyze and transmit critical information in order to organize health resources and capabilities, accelerate medical logistical and procurement chains or manage the public safety and security dimensions of confinement.

While Gazi and Gazis[63] analyze the aforementioned risks of using big data in their contribution to this issue of the *Review*, they also highlight the potential benefits of big data for humanitarian action. They note how in the context of disaster management, big data can help with responding to migration crises, epidemics and natural disasters, as well as epidemic surveillance and response. A notable example they put forward is that of Ushahidi, a software application used to improve humanitarian relief efforts. Through using the platform, researchers in Kenya

> analyzed the geographic mobile phone records of nearly 15 million individuals between June 2008 and June 2009 in order to measure human mobility in low-income settings in Kenya and understand the spread of malaria and infectious diseases. The Kenyan phone company Safaricom provided de-identified information to researchers, who then modelled users' travel patterns. Researchers estimated the probability of residents and visitors being infected in each area by cross-checking their journeys with the malaria prevalence map provided by the government.

62 See ICRC, "ICRC'S Activities in Favour of Migrants in Southern Africa", 2020, p. 5, available at: www.icrc.org/en/download/file/147853/icrcs_activities_in_favour_of_migrants_in_southern_africa_newsletter.pdf.

63 T. Gazi and A. Gazis, above note 28.

The use of such data to track infectious diseases has great potential, especially during the current age of COVID-19. Granted, as Gazi and Gazis stress, any such harvesting of data poses "re-identification risks based on persons' unique activity patterns. For this reason, when de-identified personal data are used for analysis purposes, anonymization procedures typically alter the original data slightly (causing a loss of data utility) in order to protect individuals' identities".

Also highlighting some of the benefits of digital technologies for monitoring compliance with IHL, as well as for other purposes, Milaninia[64] discusses how machine learning is being used positively, "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".

These contributors thus balance the benefits and risks of digital technologies for humanitarian action and beyond, noting how necessary it is that humanitarian practitioners make use of digital technologies while taking the appropriate mitigation measures.

Humanitarians engaging with the technology sector

Another interesting avenue emerging from this issue on digital technologies relates to the interactions between humanitarians and the actors who create these technologies. We have mentioned how technologies are used in cyber operations and their potential humanitarian consequences, a topic developed further in the "Cyber Operations and Warfare" section of this issue and covered in the next part of this editorial. Technologies are not developed in a vacuum – they are products and solutions developed by specific companies. In this context, how can humanitarian actors, who have built on decades of proven experience in "humanitarian diplomacy",[65] better interact with the technology sector? A few examples come to mind, illustrating both the substance of a possible dialogue and the forms it can take.

In this issue, Massimo Marelli highlights how specific digital infrastructures and solutions are needed to ensure that the organization remains in full control of the sensitive data it manages – for instance, through the creation of a digital humanitarian space along the model of a "sovereign cloud" or a "digital embassy".[66] The development of new technologies ensuring that the data collected by the ICRC under its mandate are and remain at all times under its exclusive control indicates one productive area for enhanced dialogue and

64 N. Milaninia, above note 29.

65 ICRC, "Humanitarian Diplomacy", available at: www.icrc.org/en/what-we-do/humanitarian-diplomacy-and-communication.

66 See Massimo Marelli, "Hacking Humanitarians: Defining the Cyber Perimeter and Developing a Cyber Security Strategy for International Humanitarian Organizations in Digital Transformation", in this issue of the *Review*.

concrete collaboration between humanitarians, the technology sector, governments and academics.[67]

In this regard, the ICRC has established a presence in the Bay Area of the United States as a step towards building a sustained dialogue with global tech companies on the way digital tools can impact people affected by armed conflicts and other situations of violence.[68] The main thrust of engagement is bringing the ICRC's operational and legal expertise, and its contextual knowledge, to a dialogue centred on the drive to ensure responsible use of technologies – which includes the core principles of privacy and trust – in humanitarian settings. This requires mutual learning about how technology may, on the one hand, be of help to populations in conflict zones, and, on the other, create different risks and harms for these individuals, communities and societies. On that basis, efforts are being made to ensure that the technology that the ICRC uses, and that affected populations use (and are exposed to), is as effective and safe as possible. This may entail the co-creation of new tools and services (such as those previously mentioned that are not easily available "off the shelf" from commercial suppliers), as well as humanitarian diplomacy to convince different stakeholders to support the ICRC, its approach and its (legal, policy and/or ethical) recommendations.

At the same time, a privileged dialogue with the technology sector is also key to better understanding its intellectual bedrock. At the heart of the growing interest from technology firms towards collaboration with humanitarians is the conviction not only that digital technologies are a source of good, but also that they can help humanitarian actors to meet the needs of affected populations efficiently and with lasting positive effects. Nevertheless, the interaction can easily be marred by "technological determinism"[69] and a "hero-preneuership culture".[70] It has been argued that these two concepts are closely connected to the Bay Area of the United States,[71] where two widespread beliefs are that technology is "leading to good outcomes for everyone" and that "new kinds of technologies should be deployed as quickly as possible, even if we lack a general idea of how the technology works, or what the societal impact will be".[72] This is one of the dimensions at play when Jo Burton, quoting Nathaniel Raymond, encourages humanitarians to avoid a "blind embrace of the potential 'promises of Silicon

67 EPFL, "EPFL, ETH Zurich and the ICRC Leverage Science and Technology to Address Humanitarian Challenges", 10 December 2020, available at: https://essentialtech.center/engineering-humanitarian-aid-awards-six-epfl-ethz-icrc-projects/.

68 Sean Capitain, "The Red Cross Presses Silicon Valley to Fight Cyberwarfare", *Fast Company*, 10 October 2017, available at: https://www.fastcompany.com/40476581/red-cross-could-silicon-valley-limit-cyberwarfare-if-governments-wont.

69 John Naughton "Think the Giants of Silicon Valley Have Your Best Interests at Heart? Think Again", *The Guardian*, 21 October 2018, available at: www.theguardian.com/commentisfree/2018/oct/21/think-the-giants-of-silicon-valley-have-your-best-interestsat-heart-think-again.

70 Daniela Papi-Thornton, "Tackling Heropreneurship", *Stanford Social Innovation Review*, 23 February 2016, available at: https://ssir.org/articles/entry/tackling_heropreneurship#.

71 Jasmine Sun, "Silicon Valley's Saviorism Problem", *The Stanford Daily*, 16 February 2018, available at: www.stanforddaily.com/2018/02/16/silicon-valleys-saviorism-problem/.

72 J. Naughton, above note 69.

Valley'", as its tendency to reduce complicated problems to a technological solution is surely at odds with the complexity faced in conflict zones. These assumptions can have massive implications when considering the adoption of technologies in humanitarian action and calls for a critical engagement with the technology sector and its companies and workforce, anywhere technology is developed.

The pilot nature of the engagement in the Bay Area illustrates the potential for similar dialogues in other hubs where digital technologies are being developed, potentially shaping future cyber operations. While the ICRC has recently strengthened its technology engagements in Japan,[73] there is surely room for similar interactions in other technology hubs around the world.

Multilateralism and the development of international law

The growing influence of the private sector also has implications on multilateralism and the development of law, a topic explored in more depth in this issue. Given the fast rise of the technology sector and the widespread use of digital technologies, Ambassador Amandeep S. Gill, in his contribution "The Changing Role of Multilateral Forums in Regulating Armed Conflict in the Digital Age",[74] identifies the structural issues that make it difficult for multilateral forums to discuss fast-moving digital issues and respond in time with required norms and policy measures. For Gill,

> [w]hile private companies and civil society have had an important agenda-setting and opinion-shaping role in some discussions, they take a secondary position to more powerful State and inter-State actors. This power asymmetry sits uneasily with the digital technology reality. For example, digital platforms such as Facebook, Alipay and WhatsApp may have more users ("virtual residents") than the populations of most countries; they operate quasi-global infrastructures, act as cross-border "content policemen" and have market capitalizations that dwarf other sectors and most national GDPs.

Gill's article highlights that "[i]f norms related to digital technologies are to have an impact, the digital industry has to be a part of the discussion on policy responses and has to cooperate with State-actors for their implementation".

Such a claim is also relevant for the humanitarian sector, particularly when it comes to IHL and its development. Given the complexities of how technologies work, how fast they evolve and the fact that their capabilities remain largely unknown, the international community and humanitarians must find new ways to ensure that new technologies used as means and methods of warfare are compliant with IHL.

73 NEC, "NEC and ICRC: A Blueprint for Ethical Technology Partnerships between the Private and Humanitarian Sectors", 11 November 2020, available at: www.nec.com/en/global/sdgs/innovators/project/article02.html.

74 See Amandeep S. Gill, "The Changing Role of Multilateral Forums in Regulating Armed Conflict in the Digital Age", in this issue of the *Review*.

Digital technologies and the means and methods of warfare

The second half of this issue of the *Review* shifts the focus from how digital technologies can be used for humanitarian relief, assessing the risks and benefits of their use, to how new technologies can used for destructive purposes in armed conflicts.

In this regard, Frank Sauer's contribution[75] engages with the implications of not regulating autonomous weapons systems, including those that rely on AI. Unpacking the costs of non-regulation, Sauer makes a robust case for why regulating AWSs is difficult, but nevertheless quite imperative on ethical, legal and policy grounds. As Sauer argues, it is essential that autonomy in weapons systems is regulated, by "codifying a legally binding obligation to retain meaningful human control over the use of force".

New applications of sensors and software, especially AI and machine-learning systems, also have broader implications for decision-making in armed conflict. Pizzi, Romanoff and Engelhardt argue that AI and machine learning "can be extremely powerful, generating analytical and predictive insights that increasingly outstrip human capabilities. They are therefore liable to be used as replacements for human decision-making, especially when analysis needs to be done rapidly or at scale, with human overseers often overlooking their risks and the potential for serious harms to individuals or groups of individuals that are already vulnerable."[76] The ICRC position paper "Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach", updated for this issue of the *Review* and appearing in the "Reports and Documents" section, is more cautious, stressing that "AI and machine-learning systems remain tools that must be used to serve human actors, and augment human decision-makers, not replace them". It argues for an approach that foregrounds human legal and ethical obligations in order "to preserve human control and judgement in applications of AI and machine learning for tasks and in decisions that may have serious consequences for people's lives, especially where these tasks and decisions pose risks to life, and where they are governed by specific rules of international humanitarian law".[77] Both papers highlight the technical limitations of AI that bring legal questions; Pizzi, Romanoff and Engelhardt note how AI

> creates challenges for transparency and oversight, since designers and implementers are often unable to "peer into" AI systems and understand how and why a decision was made. This so-called "black box" problem can preclude effective accountability in cases where these systems cause harm, such as when an AI system makes or supports a decision that has discriminatory impact.[78]

75 F. Sauer, above note 6.
76 M. Pizzi, M. Romanoff and T. Engelhardt, above note 10.
77 See ICRC, "Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach", in this issue of the *Review*.
78 M. Pizzi, M. Romanoff and T. Engelhardt, above note 10.

New digital technologies are also set to influence cyber operations and cyber warfare. The adoption of new digital technologies by parties to an armed conflict has a direct effect on the means and methods of warfare itself, and consequently on the application and interpretation of IHL in that case. As Laurent Gisel, Tilman Rodenhäuser and Knut Dörmann point out in their contribution to this issue:[79]

> [T]he use of cyber operations during armed conflict has become a reality of armed conflicts and is likely to be more prominent in the future. This development raises a number of concerns in today's ever more cyber-reliant societies, in which malicious cyber operations risk causing significant disruption and harm to humans. … The international community, societies, and each of us individually are increasingly relying on digital tools. This trend – which may be accelerated further by the COVID-19 pandemic spreading at the time of writing this article – increases our dependency on the uninterrupted functioning of these technologies, and thus increases our vulnerability to cyber operations.

The latter point is supported by recent findings by Freedom House,[80] which highlight how governments around the world have exploited the pandemic in order to expand their domestic surveillance capabilities, using, for instance, contact tracing apps that collect private information. Similarly, the CyberPeace Institute[81] has raised its voice regarding the growing and alarming number of cyber attacks. This phenomenon takes a particular shape when it comes to health-care infrastructures, because, as Gisel, Rodenhäuser and Dörmann point out, "The health-care sector seems particularly vulnerable to cyber attacks. The sector is moving towards increased digitization and interconnectivity, which increases its digital dependency and its attack surface."[82] These trends are also highlighted in the piece written by Zhixiong Huang and Yaohui Ying.[83] The authors cogently offer a different perspective on the application of the principle of distinction to the cyber context, by injecting the positions of Chinese officials and the views of Chinese scholars into the debate. They highlight how certain elements of distinction – such as uniforms and distinguishing marks – are either impractical or unworkable in the cyber sphere. While the principle of distinction remains relevant, the authors argue that it should be interpreted in a manner appropriate for the cyber realm.

79 L. Gisel, T. Rodenhäuser and K. Dörmann, above note 5.

80 Adrian Shahbaz and Allie Funk, "The Pandemic's Digital Shadow", Freedom House, 2020, available at: https://freedomhouse.org/report/freedom-net/2020/pandemics-digital-shadow.

81 CyberPeace Institute, "A Call to All Governments: Work Together Now to Stop Cyberattacks on the Healthcare Sector", 26 May 2020, available at: https://cyberpeaceinstitute.org/call-for-government/.

82 In light of this worrying trend, the ICRC has joined world leaders in calling to stop attacks against health-care infrastructure, particularly since these attacks could endanger the lives of vulnerable civilians. See "Call by Global Leaders: Work Together Now to Stop Cyberattacks on the Healthcare Sector", *Humanitarian Law and Policy Blog*, 26 May 2020, available at: https://blogs.icrc.org/law-and-policy/2020/05/26/call-global-leaders-stop-cyberattacks-healthcare/.

83 See Zhixiong Huang and Yaohui Ying, "The Application of the Principle of Distinction in the Cyber Context: A Chinese Perspective", in this issue of the *Review*.

Together, these contributions bring together not only diverse profiles of authors, but also diverse and multidisciplinary sets of analyses that enrich the ability of this issue of the *Review* to address how such digital technologies are, in times of armed conflict, regulated by IHL.

Thematic scope of this issue on "Digital Technologies and War"

As highlighted above, the contents of this issue of the *Review* touch on the dual uses for digital technologies: (1) for humanitarian action and relief – weighing the risks and benefits – aimed at assisting and protecting affected populations in armed conflicts and other situations of violence; and (2) for use in the conduct of warfare in armed conflicts. The contributions to this issue also account for the growing role of the private sector – especially big tech – in providing the platforms that are used for the dissemination of MDH and which shape how information is shared during crisis contexts.

In putting together this issue, the *Review* was cognisant that we are just beginning to unpack the patterns and trends in how digital technologies will affect the world. Thus, while this thematic issue on "Digital Technologies and War" opens up the black box to how digital technologies shape and are being shaped by armed conflicts and other situations of violence, it is not exhaustive. In other words, our current understanding of existing and emerging technologies is still increasing, bringing to light new challenges and opportunities around the digital technologies that we use, embrace, and at times, fear.

Gender, diversity and inclusion in the *Review*

An essential parameter for the production of this issue was gender parity and the inclusion of diverse profiles and views. Gender gaps in the technology sector are well known, with women comprising less than 35% of the workforce in the sector.[84] In terms of diversity, most of the largest technology companies[85] are populated by a near-homogenous group of young white males,[86] coming out of prestigious US universities,[87] with little or no training in humanities, ethics or

84 Sam Daley, "Women In Tech Statistics for 2020 (and How We Can Do Better)", *Built In*, 13 March 2020, available at: https://builtin.com/women-tech/women-in-tech-workplace-statistics.

85 Jonathan Ponciano, "The Largest Technology Companies in 2019: Apple Reigns as Smartphones Slip and Cloud Services Thrive", *Forbes*, 15 May 2019, available at: www.forbes.com/sites/jonathanponciano/2019/05/15/worlds-largest-tech-companies-2019/.

86 Shelly Banjo and Dina Bass, "On Diversity, Silicon Valley Failed to Think Different", *Bloomberg Businessweek*, 3 August 2020, available at: www.bloomberg.com/news/articles/2020-08-03/silicon-valley-didn-t-inherit-discrimination-but-replicated-it-anyway.

87 Avery Hartmans, "These 25 Universities Produce the Most Tech Employees", *Business Insider*, 2 May 2017, available at: www.businessinsider.com/top-colleges-for-working-in-silicon-valley-2017-5.

international relations.[88] Moreover, it has been argued that there is gender and race discrimination evident in digital technologies themselves,[89] and in digital data, which carry structural biases, as they represent and amplify the societal discriminations and power relations that exist. The *Review* team's aim was to at least break with this trend in terms of the profiles of authors featured in this issue. Yet, we faced certain hurdles in this quest; as we were starting to produce this thematic issue during the late winter of 2019, the coronavirus hit us like no other pandemic in the last century.[90]

On the *Review*'s end, we witnessed the gendered effects of this crisis on our authorship composition. Numerous female authors whom we'd actively solicited dropped out of submitting their manuscripts to the journal. This is a trend that has been observed across the academic publishing sector: many female academics and authors have faced the double burden of housework and professional work at higher rates than their male counterparts. As more than a few of our female authors dropped out and female authors disproportionately turned down our invitations for submission to the journal, the *Review* prolonged its publication timeline to ensure that in the end, the final product was not dominated by one demographic. As our final selection evidences, while we unfortunately have not managed to reach perfect gender parity in authorship, the gender parity gap stands at 0.82 (female to male contributors) – a ratio that the *Review* is firmly and actively committed to closing with future issues.[91] Similarly, our quest for more diversity in our publications continues – most recently, for example, the *Review* has welcomed its new Editorial Board for the 2021–26 term, comprised of a diverse group of nineteen experts from around the world.[92]

The diversity element of the issue comes not just in the form of the backgrounds of the authors but is also enhanced by the cross-disciplinary and multidisciplinary perspectives put forward by the contributors. These multidisciplinary approaches are increasingly fundamental when it comes to understanding how different practitioners, organizations and countries are accounting for and mitigating the adverse effects of digital technologies in humanitarian crises and grappling with the unprecedented ways in which digital technologies are used as means and methods of warfare.

88 Victor Lukerson, "The Ethical Dilemma Facing Silicon Valley's Next Generation", *The Ringer*, 6 February 2019, available at: www.theringer.com/tech/2019/2/6/18212421/stanford-students-tech-backlash-silicon-valley-next-generation.

89 See, for example, Karen Hao, "An AI Saw a Cropped Photo of AOC. It Autocompleted Her Wearing a Bikini", *MIT Technology Review*, 29 January 2021, available at: www.technologyreview.com/2021/01/29/1017065/ai-image-generation-is-racist-sexist/; Ryan Steed and Aylin Caliskan, "Image Representations Learned with Unsupervised Pre-Training Contain Human-Like Biases", Carnegie Mellon University, 2021, available at: https://arxiv.org/pdf/2010.15052.pdf.

90 Eskild Petersen *et al.*, "Comparing SARS-CoV-2 with SARS-CoV and Influenza Pandemics", *The Lancet Infectious Diseases*, 3 July 2020, available at: www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30484-9/fulltext.

91 As a reference, see the new composition of the *Review*'s Editorial Board, available at: https://international-review.icrc.org/about/editorial-board.

92 *Ibid.*

The way forward

A key insight brought forth across many of the pieces in this issue of the *Review* is the need to assess and mitigate the risks of integrating new digital technologies into humanitarian work while undertaking digital transformation processes. As much as these tools bring certain benefits, they also pose irreversible risks, and there is a "dark side" to the use of digital technologies in humanitarian crises. There is also another aspect to this theme: as digital technologies evolve and are used in armed conflicts and other situations of violence, there is an ever-present need to ensure that IHL is respected. Yet, the question remains: where do we go from here as humanitarians when it comes to digital technologies and war?

The director of digital transformation and data at the ICRC, Balthasar Staehelin, perfectly reflects on the future of digital technologies and war by asking: "Is data the 'new oil' – or the new asbestos? Will we live in a world of the Internet or of a splinter-net?"[93] As Staehelin highlights, however, whatever the answer may be, in the coming years and decades, "the ICRC will do its utmost to adapt to the exponentially growing impact of digital transformation with and for the populations it seeks to serve in war zones around the world. Their continued trust in the ICRC will tell us whether we have succeeded in responsibly leveraging the enormous potential of evolving digital technologies for their good."

In line with this key point made by Staehelin, most of the diverse issues covered in this issue of the *Review* come down to one key requirement of humanitarian action: trust. All throughout this issue, trust indeed emerges as the backbone of digital transformation in the humanitarian system – yet there is no quick fix to create it. Humanitarian organizations craft their humanitarian access on the basis of the trust that they create with local communities and authorities on a daily basis. By the same logic, the way the bonds of trust are affected by digital technologies should be considered with particular attention and care by all stakeholders working towards the use of "technology for good".

As we look forward, embracing digital technologies should not only be about adopting new technologies, but about ensuring that such technologies reinforce the bonds of trust that we as humanitarians build with affected populations, offering them new options to ensure their needs are met. To this end, we hope that this issue of the *Review* will inspire the design of innovative digital transformation strategies centred around and in collaboration with the people they are meant to serve.

93 Quote submitted by Balthasar Staehlin to the authors of this editorial.

International Review of the Red Cross (2020), **102** (913), 23–26.
Digital technologies and war
doi:10.1017/S181638312100014X

IRRC

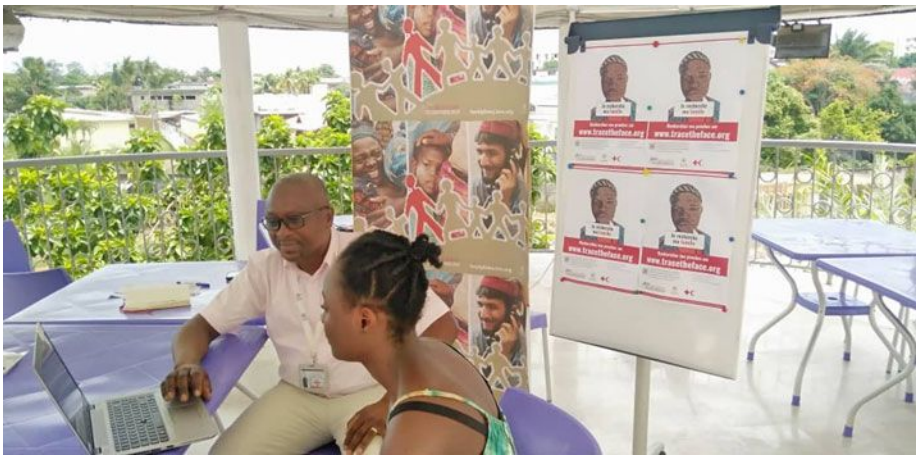
VOICES AND PERSPECTIVES

Testimonies: How humanitarian technologies impact the lives of affected populations

Digital technologies are changing the very processes we use to serve affected people. In this issue, the Review *has chosen to profile the testimonies of two affected people, based in Côte d'Ivoire and the Democratic Republic of the Congo (DRC), who expressed their opinions on two digitally driven projects by the International Committee of the Red Cross (ICRC); these projects respectively facilitate restoring and maintaining family links.*[1] *Being of the view that affected people should speak for themselves and that their testimonies should not be cut down or "reworked" to fit the purposes of humanitarians and their outputs, in this section the* Review *has directly translated the quotes we received from field delegations, word for word, sentence for sentence, making no changes except for the redaction of information to ensure the extent of anonymity requested. By keeping the framing of the quotes to a minimum, we aim to ensure that the* Review *serves as a platform for the voices of the featured affected people.*

The ICRC's Trace the Face digital platform

The ICRC's Trace the Face digital platform, mentioned by Matty in the following testimony, is a "photo-gallery of people looking for their lost relatives".[2] It is through this site that Matty, based in Abidjan, was able to find her uncle, of whom she'd had no news since the outbreak of the 2010–11 Ivorian crisis.



Matty, in a video call with her uncle; she had not had any news of him since 2011. This call took place on 20 April 2020, after Matty had located her uncle using the Trace the Face site. Photo: ICRC.

The testimony of Matty, based in Côte d'Ivoire, who found her uncle after nine years, using the Trace the Face site

I live with my friends. …[3] My mother and father are no longer alive. … My mother died in 2002, following the first crisis. … I was taken in by my uncle, my mother's brother. When I was a baby, my uncle told me that his sister, my mother, had entrusted him to take care of me, and so he took care of me before the crisis separated us in 2011. When the crisis started, everyone was running. This is how he ended up fleeing and I was left behind. I lived in the streets. People picked me up. My friends told me that I could come and live with them. … I didn't want to clutter them too much. …

One day, I was talking to a friend and I was telling him about my situation. He told me that we could go to the Red Cross to see if they could help me find my uncle. We left for the Red Cross headquarters in Plateau. The gentleman who received us took notes. He listened to my story. He said he was going to call me back. And if he didn't call me back, I should call him back. I went back to see him, as he hadn't reached out. He told me he was busy because of the coronavirus crisis. It had just started here. … So, I was waiting. That's when another friend told me that the Red Cross has a website where I can find my uncle by his picture. We went to the cyber café together to do some research. We searched the list. This is how I recognized my uncle. … I wrote

1 For humanitarian innovation to be effective and accountable to affected people, these efforts go hand in hand with processes to minimize digital risks and include affected populations as part of the process. For more information on digital risks, see "Q&A: Humanitarian Operations, the Spread of Harmful Information and Data Protection" and the "Reports" section in this issue of the *Review*.

2 See ICRC, "What Is Trace the Face?", available at: https://familylinks.icrc.org/europe/en/Pages/home.aspx.

3 Throughout these testimonies, ellipses are used to redact identifying information.

under his photo. And that's how the gentleman from the ICRC came to find me. And we exchanged. Then we kept in touch and that's how one day he told me I was going to talk to my uncle. And that's how we ended up here [at the ICRC regional delegation in Abidjan] to talk. I felt abandoned, but here I am happy. I don't know how to describe my emotion. I am filled with joy.

What do you think of the Trace the Face platform?
Tracing the face is a good thing because it allowed me to find my uncle who I was looking for. Tracing the face is good, but it is not too well known. You have to advertise on television, on radios, and display the photos everywhere. Everyone will see and everyone will be able to speak to those who are looking for their parents. You should also make it easier to use so that those who cannot use a computer can easily access the photos and be able to respond without having to ask others for help, who must know our problem first.

The ICRC's Electronic Red Cross Messaging system

The Electronic Red Cross Messaging (ERCM) system is an innovation pilot launched by the ICRC in Rwanda and the DRC. The principal aim is to test whether transforming the paper-based Red Cross messages into an electronic format using tablets and smartphones can reduce transmission times.

The pilot successfully demonstrated proof of concept and recently concluded with promising results. Discussions are currently under way to determine the feasibility of rolling out the ERCM initiative to additional contexts while considering a range of technical, programmatic and data protection questions.

The testimony of Zawadi, a mother based in Rwanda, who was connected with her family through the ICRC's ERCM system

We had already lost them; we thought they were already dead. We are surprised to see their photos; we are very happy. This is my husband's nephew – he left with his mom when he was 6 years old. Now in the post they say he already has two children. I had seen this Red Cross volunteer pass by; they were writing messages in the villages, so I also decided to write to get news. When the war came, the nationals of Rwanda were asked to return home – we thought they were already dead when they left. Now that we have just received their letter, we are really happy with the Red Cross with this letter writing job. He wrote to me that they are still alive, that they were not dead, that his mother is alive, and that he is already married but that his father was already deceased. They were gone while I was still a girl; now I have seven children, you can imagine. I'm going to write to them and send the photos of my children, and I will ask them to send me the photos of their children. I am very happy. I am also happy with you for this model of writing letters that go far. I am proud of you – you're doing a good job. It is very important to study. If I had the money, I could educate all my children. Everything you do is due to education; if someone is not educated they cannot do this job.

What do you think of the ERCM project?
It is a really good project that allowed me to make sure that the members of my family, who I thought were dead, are still alive. I'm now in contact with them thanks to the response to my electronic Red Cross message, which had my phone number. Now, me and my children can talk to our relatives without a problem and share updates. I wish that the project can continue and be useful to others as it was for me because I think that maybe other people could benefit from the project. With the ERCM, I was able to see the current photos of my relatives – my mother and brothers.

What was the impact of the ERCM project on your life?
The project changed my life. … It felt valuable to have been connected with my family. [After finding our family members], my husband … even promised to organize a visit to Rwanda for the official delivery of my dowry to my family. This could not be possible until we heard news of his family members – his mother and brothers. This project has lifted my spirits; I'm very happy and have even gained weight because of the joy.

International Review of the Red Cross (2020), **102** (913), 27–41.
Digital technologies and war
doi:10.1017/S1816383120000429

IRRC

Q&A: Humanitarian operations, the spread of harmful information and data protection

In conversation with Delphine van Solinge, the ICRC's Protection Advisor on Digital Risks for Populations in Armed Conflict, and Massimo Marelli, Head of the ICRC's Data Protection Office

In this Q&A, the Review *talks to Delphine van Solinge and Massimo Marelli of the International Committee of the Red Cross (ICRC). Van Solinge is the ICRC's focal point for understanding how digital technologies and the spread of harmful information affect populations living in conflict environments, and what this means for humanitarian action. To this end, her portfolio is focused on exploring, on behalf of the ICRC and through partnerships, how to mitigate the risks that digital technologies bring in humanitarian settings and ensure relevant protection responses in the digital age. Marelli is Head of the ICRC's Data Protection Office (DPO). During his tenure with the ICRC, the organization has chartered new pathways for how it can carry out its operational work, while ensuring that the data of the affected people which it serves, as well those of its employees, are well protected.*

During this conversation, van Solinge and Marelli discuss how their areas of work complement and reinforce each other, forming two halves of the same coin with regard to how digital information and data can both be used for positive change and misused

in humanitarian settings. Marelli highlights how humanitarian organizations process, protect and use data and digital information. Van Solinge discusses how through misinformation, disinformation and hate speech, information can be manipulated and spread using digital technologies – particularly in the age of the COVID-19, when populations are more reliant on digital communication technologies. Among the issues they discuss are how digital technologies can be used positively, the ethical considerations that humanitarian organizations should take into account, and the possible paths forward for public–private sector collaborations on this theme.

Keywords: spread of harmful information, misinformation, disinformation, hate speech, social media platforms, COVID-19, data protection, do no harm, humanitarian metadata.

: : : : : : :

What does the "weaponization of information" mean for the ICRC? How is it different from misuses of information during armed conflicts and other situations of violence? How do the "do no harm" principle and the work of the DPO relate to the "weaponization of information"?

Delphine van Solinge: As a foreword, it is important to underline that the term "weaponization of information" presents several shortcomings, in particular from a legal perspective. For instance, there is the question of whether, legally speaking, information can become a weapon – and if so, how and when? For such reasons, the ICRC would instead speak about the spread of harmful information, which covers misinformation, disinformation and hate speech [MDH] and the various different facets of these.

The ICRC has been concerned for some time by instances in which digital information and communication systems are being used in ways that have the potential to put populations of concern – such as internally displaced persons, migrants, detainees and minority groups, as well as humanitarian staff and volunteers – at a new or increased risk of humanitarian consequences. For a lack of a better term, the ICRC is using MDH as an umbrella acronym to refer to such phenomena, acknowledging however that information can be used for other purposes or without increased risk of humanitarian consequences (such as in operations that have consequences solely against adversary forces). In this respect, humanitarian consequences can refer to displacement; death; disappearance; loss or destruction of property; loss of income; physical, mental/psychological and social harm or injury; stigmatization; family separation; or denial of access to services such as education, health, shelter or food. Humanitarian consequences can also include the creation or exacerbation of existing humanitarian needs, namely for shelter, food and non-food items, medical care, psychological and psycho-social support, economic support, access to services, access to timely and locally relevant information, legal advice and support, or access to the Internet.

MDH may take the form of misinformation, disinformation, mal-information, viral rumours, digital hate speech, online propaganda, etc.[1]

The way in which information is being used is usually not enough, in and of itself, to "cause" harm. Rather, the "potential" for harm can be exacerbated when combined with underlying social, cultural and historical dynamics; existing social or political tensions; people's lack of digital literacy or critical thinking when browsing for information online; lack of trusted, accurate sources from which to triangulate information; and so on.

One may ask, how is this different from before? As history shows, there are many examples where information and communication systems have demonstrated their capacity to generate harm, such as in the case of Radio Mille Collines in Rwanda.[2] What has changed is the type of vehicle being used to spread information at a global level. Digital technologies, and social media in particular, have increased the speed, scale and impact with which information can spread and impact different audiences. Increased internet penetration, the availability of smartphones and social media have emerged as powerful tools for sharing information and connecting people, but also for exacerbating violence and conflict, such as in the case of hate speech on Facebook in Myanmar. These new variables have impacted the way in which information may be viewed as a potential means to induce civilian harm.

Massimo Marelli: I would relate what we do in the DPO to what Delphine has just said. We approach our work about data protection in much the same way: it's not just about the data, it's about how the data are used, or possibly misused. The DPO makes sure that the personal data of the ICRC's beneficiaries, interlocutors and staff are well protected, and that the essential bond of trust remains both within and toward the organization. Within this framework, "do no harm" means recognizing that a failure to protect the personal data of the ICRC's beneficiaries, interlocutors and staff could be incredibly harmful, both to the individuals concerned and to the viability of the ICRC's operations. Data protection as a tool to "do no harm in a digital environment"[3] is also linked to its capacity to provide

1 For further reading on these terms, see Peter Singer and Emerson T. Brooking, *LikeWar: The Weaponization of Social Media*, Houghton Mifflin Harcourt, Boston, MA, 2018, available at: www.likewarbook.com; Mark Silverman, "Book Review: *LikeWar: The Weaponization of Social Media*", *International Review of the Red Cross*, Vol. 101, No. 910, 2019, available at: https://international-review.icrc.org/sites/default/files/reviews-pdf/2019-12/irrc_101_910_21.pdf; John Mingers and Craig Standing, "What is Information? Toward a Theory of Information as Objective and Veridical", *Journal of Information Technology*, Vol. 33, No. 3, 2018, available at: https://link.springer.com/article/10.1057/s41265-017-0038-6.

2 Radio Mille Collines, also known as Radio Télévision Libre des Mille Collines, was a Rwandan radio station that spread disinformation and misinformation during its broadcasts between 8 July 1993 and 31 July 1994. The false propaganda that it spread played a dominant role in inciting the 1994 Rwandan Genocide against the Tutsi people in the country. For more information, see Elizabeth Baisley, "Genocide and Constructions of Hutu and Tutsi in Radio Propaganda", *Race and Class*, Vol. 55, No. 3, 2014, available at: https://journals.sagepub.com/doi/abs/10.1177/0306396813509194.

3 ICRC, *The Humanitarian Metadata Problem: "Doing No Harm" in the Digital Era*, 2018, available at: www.icrc.org/en/download/file/85089/the_humanitarian_metadata_problem_-_icrc_and_privacy_international.pdf.

a lens through which we can analyze the data flows generated by the use of technologies and understand what new stakeholders may be involved, the risks that this may generate, and possible ways of mitigating or avoiding those risks. Having said that, data protection brings much more than "do no harm": it is about ensuring the respect of the rights and dignity of affected populations when we process their data, keeping the individual at the centre. It is also about being accountable to affected populations against clear standards.

Can you tell us more about why humanitarian organizations should care about the spread of harmful information and data protection?

Massimo Marelli: Both data protection and combating the spread of harmful information are integral to the ICRC's broader protection mandate. As well, both areas of work are essential if the ICRC is to maintain the trust of affected populations and parties with whom it maintains a confidential dialogue. When it comes to data protection, in addition to serious consequences for the data subjects, a major data breach at the ICRC or any other humanitarian organization could undermine trust in the sector and its ability to access and serve those who need it the most.

The ICRC's use of biometrics in its forensics work and in Restoring Family Links [RFL] is an example of how data protection plays a role in the operational work of a humanitarian organization and speaks to why humanitarian organizations should care about data protection. The ICRC's Biometrics Policy was adopted in August 2019 to address the acute data protection challenges posed by the use of biometric data – fingerprints, facial recognition, DNA, etc. – which is particularly sensitive because once it has been collected, if retained, it creates a permanently identifiable record of an individual.[4] This can be a problem in humanitarian settings, where people may not want to be permanently identifiable, particularly if there is a risk that their information may fall into the wrong hands. The Policy was a response to growing internal interest in the potential that biometrics could bring to the ICRC's operations, and strikes a careful balance between facilitating their responsible use and addressing the inherent data protection risks. The Policy is now helping shape aspects of the digital transformation of the ICRC's Central Tracing Agency, which is developing new tools to enhance our capacity to determine the fate and whereabouts of the missing and, working with partners in the International Red Cross and Red Crescent Movement [the Movement], to restore family links. This includes the possible use of facial recognition technology to match photographs of missing and sought persons, and the use of artificial intelligence to help locate individuals in the databases of the ICRC and its humanitarian partners. Robust adherence to data protection standards will be essential to build trust in the integrity, security and use of these tools.

4 ICRC, "The ICRC Biometrics Policy", 16 October 2019, available at: www.icrc.org/en/document/icrc-biometrics-policy.

Delphine van Solinge: And the same applies when we think about the spread of harmful information, which takes place through acts of misinformation, disinformation, hate speech, etc. The way these practices and dynamics play out has implications for our humanitarian protection work. While these are not new phenomena *per se* and their potential to generate harm is relatively well-established, rapid digitalization – in ICRC operational contexts and beyond – has accelerated the speed with which harmful information can spread and can resonate with and influence different audiences. Rumours are no longer geographically contained; photos and videos can be fabricated quickly, with little overhead cost; and individuals and communities can be identified and targeted.

Myanmar, South Sudan and Ethiopia[5] are a few examples showing that phenomena related to MDH, in particular on social media, are currently and increasingly playing out in contexts struck by violence and war. What does this mean? Basically, it means that humanitarian organizations are more frequently coming face to face with practices and dynamics that exploit information and communication systems, which can be very disruptive. Such practices have the potential to destabilize already fragile environments, increase people's vulnerability and contribute to humanitarian consequences. If we are unable to understand and identify those factors of risk when we design our protection response, we might be missing out on certain harms or provide only a partial response to the needs of affected people.

Another thing that we need to realize is that MDH can directly affect humanitarian organizations' operational response and credibility. What would happen if the ICRC became the target of a coordinated disinformation campaign in a war-torn country? Our credibility might be severely tarnished, thus losing the trust of affected people; we might be denied access to war zones, thus preventing our capacity to bring protection and assistance to the populations on the ground; we might even be attacked.

If we agree that the spread of harmful information can be a vector or a factor that may increase people's vulnerability, contribute to civilian harm or lead to reputational damage or security risks for humanitarian organizations, we may need to have this on our radar.

What are the challenges that the humanitarian sector is facing as a result of the spread of harmful information? Also, we hear a lot about the challenges posed for humanitarian organizations in terms of the collection and protection of humanitarian metadata – can you tell us about the challenges and risks of collecting and protecting this metadata?

Delphine van Solinge: I'll answer your first question. There are many challenges related to the spread of harmful information, but we can look at three of them here.

5 In those contexts, social media platforms such as Facebook and Twitter have been used to spread misinformation, rumours and hateful speech which have exacerbated tensions and led to acts of violence on the ground.

Firstly, MDH threats are non-localized. They are not generated by one group of people sitting in a bunker, with a unique ability to build harmful content or to create new approaches to spreading false information. In the digital age, every connected individual can consume, create and share content, and with it, they can become an actor who is spreading information that is causing harm, without even knowing it. In the same vein, in some instances, you do not need to do much to achieve a lot when it comes to disinforming or misinforming a potentially large audience. With a two-minute video, filmed and edited with a smartphone and posted on social media, you can achieve a huge effect. We have all watched "flash mob" videos, and this is a similar concept: with one message you have everyone dancing in Central Station in Antwerp.[6] Also, the availability of tools and malware at relatively cheap prices makes the work of malicious actors increasingly easier. This can potentially raise a number of legal questions in terms of responsibility, accountability and participation in hostilities.

Related to this, we know also that MDH is not easily detected or verified by conventional means. In the ocean of news and information that bombards us, it takes time and practice to develop an expert eye. Adding to the complexity is the difficulty of measuring the impact of MDH, which can sometimes be diffuse or intangible. For example, how would you measure the erosion of trust? Or, how would you determine the relationship between online activity and real-world consequences? Some non-profit organizations and academics have been more closely studying the different aspects of "weaponization of information", but they rarely focus on countries affected by war and violence. As for the humanitarian community, it is slowly awakening to the potential risks related to MDH and the spread of harmful information, but it struggles to determine how to integrate this new dimension into its work.

Finally, digital communication and MDH bring in a new set of actors with varied roles and responsibilities, such as the media and the private sector. The question of how to engage with these actors in a meaningful dialogue, beyond just promoting our humanitarian activities, is still very new for us.

Massimo Marelli: I think that much like the spread of harmful information, there are a number of challenges when it comes to how humanitarian organizations protect and store metadata. But first, let me start with the terminology we're working with here. Metadata is data about data. This includes information about communications, such as who contacted whom and when, and the digital traces left behind when computer devices, applications and networks interact with one another.

Why is the gathering and protection of metadata important? Well, because metadata reveals a lot about its subjects. Metadata is what online advertising companies use to profile users of the Internet, and what intelligence and security agencies use to identify persons and groups of interest. In a humanitarian context, the challenges and risks of collecting metadata relate to how metadata

6 The referenced flash mob dance can be seen at: www.youtube.com/watch?v=7EYAUazLI9k.

can be used to reveal the location, movements and interactions of humanitarian organizations and affected populations – information which could betray confidentiality and be used for non-humanitarian purposes, for example by parties to a conflict. In some contexts, though, metadata can also be incredibly useful to humanitarian organizations, allowing them to better understand how their services are used and where they are needed most. Metadata can also be used to help determine the fate of missing persons in online environments. The ubiquity of this kind of information is what makes the data protection risks so difficult to manage. In this respect, the ICRC has partnered with Privacy International to help the raise awareness of the "humanitarian metadata problem"[7] and has collaborated with the Brussels Privacy Hub to produce data protection guidance on how to manage these risks.[8]

How do pandemics like COVID-19 affect the cases and prevalence of spread of harmful information and the need to be vigilant about data protection in the humanitarian sector?

Delphine van Solinge: The outbreak of an infectious disease often leads to increased MDH. Ideas around pandemics tap into our deepest and strongest emotions, inciting fear and creating waves of panic. With growing levels of uncertainty and anxiety, people seek answers by turning to information outlets. News on social media spreads faster and wider, and has low levels of curation. In this context, information can be manipulated for economic, ideological or political gain.

In the case of COVID-19, misinformation, viral rumours and disinformation have been observed on many digital platforms and websites. The spread of information has attained such levels that it has been labelled as an "infodemic".[9] Certain unverified content related to infectious diseases, imminent threats and death can override people's rational minds and increase polarization within societies. In some cases, these frameworks of thinking can manifest in the form of extreme, possibly violent behaviours, including direct physical attacks on medical staff or facilities, riots followed by police or military use of force, etc. Misleading or wrong information spread through social media about the location of quarantine sites, for instance, has increased public uncertainty and fear, prompting citizens to attack convoys transferring patients and to block evacuations.

7 ICRC, "Digital Trails Could Endanger People Receiving Humanitarian Aid, ICRC and Privacy International Find", 7 October 2018, available at: www.icrc.org/en/document/digital-trails-could-endanger-people-receiving-humanitarian-aid-icrc-and-privacy.

8 ICRC, *Handbook on Data Protection in Humanitarian Action*, 23 August 2017, available at: www.icrc.org/en/data-protection-humanitarian-action-handbook.

9 For example, see UN Department of Global Communications, "UN Tackles 'Infodemic' of Misinformation and Cybercrime in COVID-19 Crisis", 31 March 2020, available at: www.un.org/en/un-coronavirus-communications-team/un-tackling-'infodemic'-misinformation-and-cybercrime-covid-19; Farah Lalalni and Juraj Majcin, "Inside the Battle to Counteract the COVID-19 'Infodemic'", World Economic Forum, 9 April 2020, available at: www.weforum.org/agenda/2020/04/covid-19-inside-the-battle-to-counteract-the-coronavirus-infodemic/.

The spread of misinformation or disinformation over social media may also be increasing because platforms' content moderators have been sent home, leaving the removal of misleading content to machines. While promising, the ability of artificial intelligence and automated systems to take down misleading content is still insufficient.

Misinformation or disinformation can affect humanitarian organizations, particularly those affiliated or perceived to be affiliated to countries with high levels of COVID-19 infection. Foreign humanitarians can be perceived by the population as carrying the virus. This could have far-reaching consequences in terms of security and operational capacity.

What has the ICRC done so far about MDH in terms of its humanitarian operations, and about applying the "do no harm" principle to its work on data protection?

Delphine van Solinge: In terms of operational work regarding MDH, in December 2018, the ICRC organized a symposium on digital risk in London.[10] This event was aimed at understanding how digital technologies and the ways they are being used affect civilian populations in armed conflicts, and the implications in terms of protection and humanitarian response.

Based on some of the key highlights from the event, the ICRC initiated a programme of work on MDH. The first phase of this programme was conducted in 2019, and its focus was on understanding the key barriers and challenges that prevent ICRC staff from integrating MDH into their analysis and work. This research was carried out in two ICRC delegations: Sri Lanka and Ethiopia. Based on some of the initial needs identified during the research, we have developed a practical guide on MDH for ICRC field staff to help them become more familiar and at ease with this concept.

Last but not least, we are now embarking on the creation of a research network on MDH with academia and interested humanitarian organizations. Our aim is to build evidence-based research around MDH and humanitarian consequences with a view to developing the conceptual foundations that humanitarian organizations will need to address this issue. This will help in defining the implications of MDH for protection and humanitarian work. It will also elucidate how the ICRC and other interested humanitarian organizations can best incorporate MDH into their analyses and responses to affected people.

Massimo Marelli: The work we're doing in terms of data protection similarly focuses on carefully assessing the risks of harm of a specific course of operational action and then taking steps to mitigate those risks, which may include not going ahead with the operation or rethinking operational strategy. With regard to "do no harm", we have now adopted several specific safeguards in the ICRC's Rules

10 The report of the London Symposium on Digital Risks in Armed Conflicts can be found at: www.icrc.org/fr/publication/4403-symposium-report-digital-risks-armed-conflicts.

on Personal Data Protection, adopted in 2015, which are designed to reduce the risk of unauthorized use or access to personal data by applying data protection standards and requirements to data processing throughout the organization.[11] Where new technologies or riskier data processing operations are considered by the ICRC, a Data Protection Impact Assessment must be conducted to identify and mitigate the risks of harm.[12] The Rules also require the ICRC to follow a "data protection by design"[13] approach to minimize the collection of personal data to that which is necessary for the operation and ensure that data subjects' rights are respected.

As a humanitarian community, what can we do to start addressing the challenges arising from the spread of harmful information (MDH) and data protection concerns?

Delphine van Solinge: To start addressing the challenges arising from the spread of harmful information, we need a deeper understanding about MDH and humanitarian consequences, a strategic orientation supported by solid conceptual foundations, and a willingness to engage with new actors in uncharted waters.

However, to do this we need to keep in mind the following. If we start working on solutions in silos and without shared priorities, the impact will be limited, and we will waste energy. MDH is an issue that concerns many of us, and it is a complex problem in which many different dynamics, systems and actors are present and interacting. For this work on MDH, we need to follow a systemic approach.

Finally, we need to develop the conceptual foundations for the spread of harmful information. This includes a clear and thorough understanding of the potential risks and harms that digital information technologies and their uses can introduce for affected populations and humanitarian organizations, as well as how to respond to and mitigate those risks. We therefore need a theory of digitally derived harm, and to connect theory to practice, we need a conceptual framework. Given the breadth of this endeavour, there is a need to partner and join forces with one another – for example, with other organizations or academics and academic institutions.

Massimo Marelli: I agree with Delphine – we need to break down the silos and identify some shared priorities across the humanitarian sector if we want to effectively address the spread of harmful information and data protection concerns. In terms of data protection, the embrace of new digital technologies by the humanitarian sector and the specific risks that these technologies have

11 ICRC, *ICRC Rules on Personal Data Protection*, 2015, available at: www.icrc.org/en/publication/4261-icrc-rules-on-personal-data-protection.
12 *Ibid.*; see also ICRC, "Policy on the Processing of Biometric Data by the ICRC", 28 August 2019, available at: www.icrc.org/en/download/file/106620/icrc_biometrics_policy_adopted_29_august_2019_.pdf
13 ICRC, above note 11.

brought with them have prompted the ICRC to think seriously about what it means to "do no harm in a digital environment". To this end, we have actively formed strategic partnerships with data protection-conscious organizations[14] and service providers, and have engaged in "digital diplomacy" to address some of the specific challenges facing the humanitarian sector. This includes efforts to safeguard the "digital humanitarian space",[15] ensuring that data collected for humanitarian purposes can only be used for those purposes in accordance with the principles of neutrality and independence.

In addition to ensuring that data protection rules are incorporated into partnership agreements at the operational level, the ICRC has, for example, worked closely with the Movement to develop data protection standards for RFL, which are represented in a Movement Code of Conduct.[16] Data protection issues also figured prominently at the 33rd International Conference of the Red Cross and Red Crescent in December 2019,[17] where doing no harm in digital environments was widely discussed in the context of "Shifting Vulnerabilities" and "Trust in Humanitarian Action". The Conference also adopted a groundbreaking resolution on RFL and data protection which recognizes that the acquisition and use of humanitarian data for non-humanitarian purposes undermines trust in humanitarian organizations and threatens their ability to operate. The resolution "urges States and the Movement to cooperate to ensure that personal data is not requested or used for purposes incompatible with the humanitarian nature of the work of the Movement".[18]

14 For example, the ICRC and the Brussels Privacy Hub have collaborated together on the Data Protection in Humanitarian Action Project, aimed at the staff of humanitarian organizations involved in processing personal data as part of humanitarian operations, particularly those in charge of advising on and applying data protection standards. Outputs include the *Handbook on Data Protection in Humanitarian Action*, available at: www.icrc.org/en/data-protection-humanitarian-action-handbook. The ICRC has collaborated and/or consulted with experts from numerous organizations in its data protection work, including but not limited to the Brussels Privacy Hub, the Swiss Data Protection Authority, the European Data Protection Supervisor, the Office of the United Nations High Commissioner for Refugees, the International Organization for Migration, the International Federation of Red Cross and Red Crescent Societies, the United Nations Office for the Coordination of Humanitarian Affairs, Yale University, Privacy International, the French-Speaking Association of Personal Data Protection Authorities, the Swiss Federal Institute of Technology in Lausanne, Doctors Without Borders, and the Senegalese Data Protection Authority.

15 For a more detailed analysis of the ICRC's role in safeguarding the "digital humanitarian space", see Massimo Marelli, "Hacking Humanitarians: Moving Towards a Humanitarian Cybersecurity Strategy", *Humanitarian Law and Policy Blog*, 16 January 2020, available at: https://blogs.icrc.org/law-and-policy/2020/01/16/hacking-humanitarians-cybersecurity-strategy/.

16 International Red Cross and Red Crescent Movement Family Links Network, *Code of Conduct on Data Protection*, November 2015, available at: www.icrc.org/en/download/file/18229/rfl-code-of-conduct.pdf.

17 International Conference of the Red Cross and Red Crescent, "33rd International Conference: At a Glance", available at: https://rcrcconference.org/about/33rd-international-conference/.

18 International Conference of the Red Cross and Red Crescent, "Resolution: Restoring Family Links While Respecting Privacy, Including as It Relates to Personal Data Protection", 33rd International Conference, 9–12 December 2019, available at: https://rcrcconference.org/app/uploads/2019/12/33IC-R4-RFL-_CLEAN_ADOPTED_en.pdf.

What are some examples of how digital information and communication systems can be used positively?

Delphine van Solinge: Digital information technologies offer opportunities to improve the humanitarian responses to affected populations, including, but not limited to, by facilitating two-way communication between humanitarian staff and people affected by crises, or by using innovative ways of capturing and using crisis-related information to inform responses. Human rights defenders and humanitarian practitioners have made use of the enhanced situational awareness and actionable information afforded by the digital age. There are many examples, but I'll note a few now: they've employed remote sensing tools for augmenting conflict early warning capacities and documenting human rights abuses. They have leveraged mobile data solutions for tracking the conditions, profiles and routes of transit of migrant and refugee populations; extracted metadata from call detail records to understand the spread of infectious diseases; harvested social media for sentiment analysis and rumour tracking in fragile contexts; and of course, they've deployed aerial robotics for surveillance of damaged locations and monitoring critical infrastructure.

In the case of COVID-19, digital tools, artificial intelligence[19] and "big data" analysis are being used in various contexts to support health-based responses. They can help us collect, analyze and transmit critical information in order to organize health resources and capabilities, accelerate medical logistical and procurement chains or manage the public safety and security dimensions of confinement.

Digital information technologies can be valuable for exchanging key information and thus for advancing medical and epidemiological research in laboratories. In terms of prevention and awareness, information applications for sharing relevant and accurate information with affected populations have also been widely used during the COVID-19 crisis. They have come with different models, interfaces, content, and levels of privacy compliance and security. Overall, when associated with and validated by official public health agencies, these apps are useful for helping to improve the awareness and level of information of the population, who are then better equipped to take appropriate measures in terms of prevention and wellness.

Overall, there are many different digital tools that are being promoted and discussed in the context of the COVID response. They can play different roles and functions depending on their purpose and design, but also on the moment and places in which they are deployed and used. Thus, while digital technologies can be helpful to the COVID response, the analysis of their relevance, data protection compliance

19 Shana Lynch, "Artificial Intelligence and COVID-19: How Technology Can Understand, Track and Improve Health Outcomes", *Stanford Institute for Human-Centered Artificial Intelligence Blog*, 1 April 2020, available at: https://hai.stanford.edu/blog/artificial-intelligence-and-covid-19-how-technology-can-understand-track-and-improve-health.

and adequacy needs to be systematically pushed further and be based on a case-by-case and contextualized assessment, as their relevance and added value will vary widely.

Massimo Marelli: I think digital information and communication systems can have positive effects if we regulate them properly. For example, data protection laws ensure that digital technologies can be used in ways that bring "digital dignity" or "data dignity", by giving the data subject the information, control and rights that they need to exercise control over how information about them is used. These laws also seek to impose limits on what data controllers can do with our information, to ensure that we are treated fairly and not discriminated against or exploited. By incorporating these laws into our work, we can ensure that digital technologies' positive effects outweigh their negative ones. Based on this understanding, these fundamental principles for data protection are the basis for the ICRC's Rules on Personal Data Protection.[20] Of course, it's easy to say that it can be difficult to apply these principles to humanitarian action because of the circumstances on the ground, or because our beneficiaries "don't care" – but if we are serious about ensuring that digital technologies have an overall positive net effect, and we are serious about respecting the dignity of affected people, doing no harm, being accountable and retaining trust, then this is not good enough. We have no choice but to try to overcome these challenges and ensure that our actions meet up with the principles that define humanitarian work in this area.

Why should the humanitarian sector engage with the private sector's tech industry?

Delphine van Solinge: I'll take this one. The rapid evolution in technology, connectivity and data is driving multilayered changes in societies, and in the way we work and communicate. In humanitarian settings, this is affecting not only the expectations and needs of affected populations and other stakeholders, but also the way humanitarian programmes and services can be delivered.

While offering new opportunities for the humanitarian sector to implement and scale its response, the digital transformation is also creating new and/or amplified risks for conflict-affected populations. Yet, the digital transformation of societies and businesses is no longer something you can really opt out from. It is happening, and we need to learn to live and work with it.

The humanitarian sector has a double obligation to look into the relevance of digital solutions and find the right type of ethical engagement with the tech industry. The first reason for this is because technology and data have the potential to improve humanitarian responses and therefore help alleviate the suffering of affected populations. The second reason is because these technologies, depending on their uses and misuses, can create risks for populations and/or damage the credibility of humanitarian organizations. This potential for improved humanitarian response through the use of digital technologies needs to be

20 ICRC, above note 11.

analyzed, understood, explored and deployed responsibly, following the "do no harm" principle and relevant rules of data protection.

In the same vein that the humanitarian sector can learn from the tech industry in designing a more efficient response, the tech industry can learn from the humanitarian sector in thinking beyond its labs, in understanding the real-world consequences that "technocolonialism"[21] can have on the lives and safety of affected populations, and in jointly finding ways to mitigate those consequences.

Massimo Marelli: I think Delphine has just about covered what I would have to say on this front. I echo her statement that we can no longer "opt out" from digital transformation processes, and we see this with the dominance of the private sector's tech industry. On our end, in the humanitarian sector we have an obligation to engage with these actors and to ensure that we adapt our work in such a way that we continue to put the safety and trust of affected people first in our operational work.

Are social media giants responsible for having policies that protect affected populations from widespread misinformation or disinformation campaigns?

Massimo Marelli: I think Delphine is best placed to speak to this.

Delphine van Solinge: Thanks, Massimo. Well, I think it is a shared responsibility. MDH does not happen in a vacuum; it is rooted in history, societal behaviours, chronic tensions and politics, to name just a few factors.

The platforms are not the cause of the spread of harmful information *per se*; they certainly have a role in amplifying, magnifying and increasing the speed of how information is shared, but they do not initiate the post. This should not be interpreted as a way for social media companies to "wash their hands" and do nothing about the dangerous content that is mushrooming on their platforms. Since they provide this lucrative service to people, where algorithm, artificial intelligence and data analytics play a major role in making content more or less visible and accessible based on one's profile, interest and/or social behaviour, they have a responsibility to put in place the necessary mitigating policies and technical measures to limit the spread of information in ways that can be harmful for individuals.

During the COVID crisis, many social media platforms, such as Facebook and Twitter, have taken steps to limit, for example, the amount of misinformation circulating around curative treatments which could clearly have lethal consequences

21 "Technocolonialism", as defined by Mirca Madianou, refers to "how the convergence of digital developments with humanitarian structures and market forces reinvigorates and reshapes colonial relationships of dependency". See Mirca Madianou, "Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises", *Social Media and Society*, Vol. 5, No. 3, 26 July 2019, available at: https://journals.sagepub.com/doi/full/10.1177/2056305119863146.

for people – that is, incorrect and potentially harmful medical advice.[22] While many people would argue that such information is obviously and totally false, in the midst of a health crisis, the fear of death can make people think irrationally. In situations of armed conflict and even dormant violence, primal fears and survival instinct can be easily tapped into. Posting and circulating alarming or decontextualized information in these circumstances can foster polarization, increased stress and fear, unreasonable behaviours and, ultimately, violence.

This is why it is important to look at the different actors, roles and responsibilities beyond the platforms. Politicians, authorities and civil society have a responsibility and a role to play, in their different capacities and functions, to help limit the creation and spread of MDH on social media. However, we also need to be realistic: misinformation and disinformation are deeply rooted into politics, power and social behaviour, and as such, they will continue to be used in different ways, shapes and forms. What can be done, however, is that we can increase people's resilience to MDH by promoting digital literacy, critical thinking and – let's be idealistic a notch – humanitarian values.

What are the benefits and disadvantages of affected populations having increased access to digital connectivity?

Delphine van Solinge: I'll let Massimo take this one.

Massimo Marelli: Thanks, Delphine. First, let me start by saying that we have to see connectivity as a fact of life and not simply as a challenge or opportunity. As more and more social and material life moves online, and more and more of the world is connected, humanitarian organizations have no choice but to carve out a presence in the online spaces that affected populations congregate in, leverage new resilience mechanisms that are enabled by connectivity, and factor connectivity into their programming. This can be relatively simple, for example by providing "connectivity as aid"; but it can also be more complex, for example where new digital services are concerned. Beyond data protection compliance – and remaining fully responsive to the needs of those who are not connected – we see the main challenge as one of responsible innovation. We have to invest in building safe and protected digital humanitarian spaces where we can be confident that we are actually doing no harm as well as working within the confines of digital environments as we find them today. This is far from easy. It means educating partners, States and technology providers as to why these spaces

22 For example, Google is removing false or misleading information about COVID-19 from its various platforms and in advertisements; see Sundar Pichai, "COVID-19: How We're Continuing to Help", *Inside Google*, 15 March 2020, available at: https://blog.google/inside-google/company-announcements/covid-19-how-were-continuing-to-help/. Twitter now verifies tweets and Twitter accounts for the credibility of information they offer, and has put in place informative #KnowTheFacts search prompts; see "Coronavirus: Staying Safe and Informed on Twitter", *Twitter Blog*, 3 April 2020, available at: https://blog.twitter.com/en_us/topics/company/2020/covid-19.html.

are needed and working with them to develop the infrastructure and applications that neutral, independent and trusted humanitarian action requires.

What are some of the ethical implications of, for example, using digital technologies, such as facial recognition software, to identify missing persons? How do these ethical considerations shape the ICRC's work?

Massimo Marelli: Ethics and data protection frequently overlap. Data protection laws, like most laws, are in the end a reflection of the responses that societies give to ethical questions and the rules they decide to give to themselves to stay loyal to those responses. I think that in the context of the technologies you've mentioned, the key questions are: can these technologies bring real benefits to the ICRC and affected populations, on the one hand, and can we utilize them responsibly and accountably, keeping the affected populations at the centre, on the other?

International Review of the Red Cross (2020), **102** (913), 43–73.
Digital technologies and war
doi:10.1017/S1816383120000491

IRRC

"Doing no harm" in the digital age: What the digitalization of cash means for humanitarian action

Jo Burton*
Jo Burton is a Cash Transfers and Markets Specialist with over twelve years of field experience in conflict-affected countries. Having first worked with Cash and Voucher Assistance (CVA) in 2004, she has specialized in CVA in conflict and post-conflict contexts. She is currently the ICRC's Institutional Lead for CVA, ensuring that it is an integral and multidisciplinary part of the ICRC's response to the needs of people affected by armed conflicts and other situations of violence. Email: joburton@icrc.org.

Abstract
Cash transfers have changed the way the humanitarian sector delivers assistance, and at the same time, digitalization is changing the way our world works in fundamental ways. The digitalization of cash means that the simple click of a button can put money in the hands of hundreds of thousands, if not millions, of people within minutes. Digital payments have been a game changer, opening the door to faster and more efficient delivery of life-saving assistance. Although physical currency will not disappear with the rise of digital payments, it is essential to balance the benefits of these digital processes with the risks. As humanitarians, we need to articulate what "do no harm" means in the digital age, applying this equally to the way we use digital payments to support people affected by armed conflicts and other situations of violence.

* This article reflects the author's views alone and not necessarily those of the International Committee of the Red Cross.

Keywords: cash and voucher assistance, cash transfers, digitalization, digital payments, do no digital harm, digital risks, data protection, data responsibility.

Money makes the world go around, or so many would say. But why is cash – an item so familiar in our daily lives – such a talking point in the humanitarian sphere?

It is because the use of "cash", a common shorthand for Cash and Voucher Assistance (CVA),[1] has been one of the biggest changes in humanitarian action in the last decade. CVA use doubled between 2016 and 2019 – in 2019, $5.6 billion of international humanitarian assistance, or 17.9% of the total budget for such assistance, was delivered through cash or vouchers[2] – and it is transforming the way humanitarian assistance is being delivered.

This focus on delivering CVA – the provision of cash[3] and/or vouchers to individuals, households or communities to enable them to access the goods and services that they need,[4] rather than distributing goods – has changed the way we think about and deliver humanitarian assistance, putting an emphasis on the choices of affected people and changing the power dynamics within the humanitarian sector. At the same time, as the use of CVA has increased, we have seen a rise in digital transformation processes across the humanitarian sector, which has been further amplified by the onset of the COVID-19 pandemic. Digitalization is changing how our world works in fundamental ways, from how people communicate to how we form networks, how we travel, and how we make decisions and ensure our voices are heard. In short, digitalization is changing the way people interact with their social, political and built environment, not only in the digital space, but also in the physical space.

When it comes to CVA, digitalization has also changed the way we deliver vital assistance, significantly increasing both the speed of delivery and the volume of people we can reach with that assistance.

This article aims to unpack how CVA and digitalization processes have come together, exploring the "do no digital harm" concept specifically in relation to the use of digital payments to support people affected by armed conflicts and other situations of violence. Starting with a broad overview of some of the digital

1 CVA is the most current terminology used in the humanitarian sector, with the previously used synonyms being Cash Transfer Programming, Cash-Based Assistance and Cash-Based Interventions. See Cash Learning Partnership (CaLP), "Glossary of Terms", available at: www.calpnetwork.org/library-and-resources/glossary-of-terms/ (all internet references were accessed in December 2020).

2 CaLP, *The State of the World's Cash 2020: Cash and Voucher Assistance in Humanitarian Aid*, July 2020, p. 9, available at: www.calpnetwork.org/publication/the-state-of-the-worlds-cash-2020-full-report/.

3 "Cash" in this definition of CVA refers to both physical and digital payments.

4 The International Committee of the Red Cross's (ICRC) definition – in line with CaLP and other CVA providers – focuses on direct transfers to individuals, families and communities in need. The ICRC does not count larger transfers of money to partners like National Red Cross and Red Crescent Societies and large businesses, or salary top-ups to staff in the relevant authorities with whom the organization works. CVA also excludes remittances and micro-finance, although micro-finance institutions may be used for the actual delivery of CVA. ICRC, *Cash Transfer Programming in Armed Conflict: The ICRC's Experience*, Geneva, November 2018, p. 14, available at: www.icrc.org/en/publication/cash-transfer-programming-armed-conflict-icrcs-experience.

trends influencing humanitarian work, it will first define digital payments, exploring the rise of their use in the humanitarian sector. It will then explore how the "do no harm" concept must evolve in an increasingly digitalized world, accounting for the risks and potential pitfalls associated with digital payments, in particular with regard to data protection; here it will focus on the role of the International Committee of the Red Cross (ICRC) as an organization that has been a leading actor in this field over the last decade. The analysis will also highlight how the outlined digital risks associated with CVA can be mitigated. It will conclude with a call to action for humanitarians to be more mindful in our decision-making around digital payments and our interactions in the digital world, always putting affected people at the centre of our work.

Why are humanitarians talking about the digitalization of cash?

To understand the benefits and potential pitfalls associated with the digitalization of cash, it is necessary to unpack three major concepts: digitalization, digital payments, and do no harm.

Digitalization and digital payments

Digitalization covers a wide spectrum of different areas, many of which are explored in the articles collected in this issue of the *Review*. Major trends impacting humanitarian action and the wider world include, but are not limited to, ubiquitous connectivity; big data; the impact of technological disruptors on traditional business models; the spread of misinformation, disinformation and hate speech; cyber warfare; surveillance; and artificial intelligence. Ubiquitous connectivity means that people are better connected than ever; in 2019, more than 53% of the world's population were able to access the Internet[5] and 67% had subscribed to some form of mobile communications services.[6]

This digital connectivity – including the increased use of digital payments – both requires and generates huge quantities of data. The digitalization of this data, including financial data, is enabling increased surveillance, which today can be carried out by governments, lawfully or not, and by corporations that take advantage of the services they provide in order to gather, process or infer information. Additionally, as spying software becomes cheaper and easier to obtain, there are real-world risks beyond a basic invasion of privacy. Evidence shows that despite the duty of States to uphold and protect the rights of citizens, "surveillance of individuals – including government critics, journalists, and human rights advocates – is largely unimpeded, leading to detention, torture, and

5 International Telecommunication Union, *Facts and Figures 2019: Measuring Digital Development*, Geneva, 2019, available at: https://itu.foleon.com/itu/measuring-digital-development/home/.

6 GSMA, *The Mobile Economy 2020*, London, 2020, p. 3, available at: www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Global.pdf.

extrajudicial killings".[7] Beyond the data itself, our digital world requires both physical and digital infrastructure. Cyber attacks – in the sense of hostile operations conducted through data streams against computers, computer systems, connected devices or networks – have been increasing year on year, and this has only been amplified during the COVID-19 pandemic.[8] Cyber attacks can target any infrastructure relying on connected networks, with real-world consequences; this can include financial systems, as we will explore later.

How we use money has also changed as a result of digitalization. If I were to ask you how much cash you are carrying in your pocket right now, what would your answer be? One dollar? Ten dollars? Nothing at all? Increasingly, people rely on digital payments – by card, by mobile phone or using internet banking – to pay for the goods and services they need, whether that be paying for groceries, a meal out, rent, or health insurance. Many times, I've seen a look of consternation on a person's face on that (rarer and rarer) occasion where a shopkeeper utters the words, "We don't take cards." It does not follow, however, that cash will cease to exist. Numerous reports by central banks and financial experts, while recognizing the rise of digital payments, note that – to misquote Mark Twain – "reports of the death of cash have been greatly exaggerated".[9]

Though the terminology is likely to change as technology evolves, it is important to understand what is meant by the digitalization of cash, and digital payments, in order to have a common understanding of what we are speaking of, and to be able to analyze the impact of these phenomena on humanitarian action. In essence, we are talking about "digital payments" or "electronic transfers", where money is moved electronically – for example, a bank transfer or mobile money transaction. The Electronic Cash Transfer Learning Action Network (ELAN), which has been a pioneer in the humanitarian sector on the use of digital payments, defines them as a "digital transfer of money or vouchers from the implementing agency to a program participant. E-transfers provide access to cash, goods and/or services through mobile devices, electronic vouchers, or cards (e.g., prepaid, ATM, credit or debit cards)."[10]

7 Siena Anstis, Ron Deibert, Miles Kenyon and John Scott-Railton, "The Dangerous Effects of Unregulated Commercial Spyware", *Citizen Lab*, 24 June 2019, available at: https://citizenlab.ca/2019/06/the-dangerous-effects-of-unregulated-commercial-spyware/.

8 Maggie Miller, "FBI Sees Spike in Cyber Crime Reports during Coronavirus Pandemic", *The Hill*, 16 April 2020, available at: https://thehill.com/policy/cybersecurity/493198-fbi-sees-spike-in-cyber-crime-reports-during-coronavirus-pandemic.

9 A 2014 European Central Bank study which focused on cash and non-cash (including digital) payments across seven European countries found that cash is still widely in use, even where many other payment mechanisms exist. Importantly, the study found that use of cash is strongly correlated with demographics and point-of-sale characteristics. While this study looks at industrialized countries, it is important to consider how these results might be extrapolated into the types of contexts that humanitarians work in. See John Bagnall, David Bounie, Kim P. Huynh, Anneke Kosse, Tobias Schmidt, Scott Schuh and Helmut Stix, *Consumer Cash Usage: A Cross-Country Comparison with Payment Diary Survey Data*, Working Paper Series No. 1685, European Central Bank, June 2014.

10 ELAN, "Vocabulary and Usage", available at: www.calpnetwork.org/wp-content/uploads/2020/01/elan-vocab-and-usage-expanded-jan-2017.pdf.

Digital payments should not be confused with digital currencies such as Bitcoin, one of the most popular forms of cryptocurrency. Cryptocurrencies are digital forms of currency created by a public network, rather than any government, so they are not legal tender.[11] There are countries – like Canada – that are considering introducing digital legal tender, thus making paper and metallic currency obsolete.[12] However, a recent study of humanitarian CVA notes that despite the increased interest in distributed ledger technology and digital currency, "real-life application at scale is a long way off".[13]

For the purposes of this article, I will use the term "digital payments" to refer to end-to-end digital transactions. This means that both the payer and payee use an electronic medium of transfer (such as a bank transfer or mobile money transfer) and that the payment instrument (how the payment instructions are carried) is itself digital and not paper-based (i.e., not cash, cheques or money orders).

Of course, digital payments do not mean that no physical currency is involved. For example, a recipient might receive funds into their personal account and then withdraw some or all of the money in cash using an ATM card; or, with mobile money, the funds could be moved to the recipient's digital wallet, and then the person can make digital payments (e.g., an electronic payment in a shop) and can also "cash out" (e.g., withdraw cash from any point of sale).

In general, long gone are the days of humanitarians driving around with a Land Cruiser full of bank notes in order to be able to make payments to a community so that they can cover their needs, whether that entails buying food, clothes and households items, or paying for transport, shelter or health care. While it is true that there are still places where there are no means to transfer money electronically, humanitarians are increasingly disbursing funds digitally – through smart cards or mobile phones, or to people's bank accounts. It is the digitalization of the delivery of cash, and not just the cash itself, that is the game changer. The simple click of a button can put money in the hands of hundreds of thousands – if not millions – of people within minutes, and this is changing decades-old processes, opening the door to faster and more efficient delivery of life-saving assistance.

An evolution of the "do no harm" imperative: The example case of the ICRC

Hand in hand with the increasing use of digital payments comes the need to mitigate potential adverse spillover effects from this digitalization process. While recognizing that there is no such thing as a risk-free world, as a part of our duty of care to affected people, the humanitarian sector needs to think about the impact of its

11 Legal tender is anything recognized by law as a means to settle a public or private debt or meet a financial obligation, including tax payments, contracts, and legal fines or damages.
12 Better Than Cash Alliance (BTCA), "Payments Measurement Toolkit", available at: www.betterthancash.org/tools-research/toolkits/payments-measurement/focusing-your-measurement/introduction.
13 CaLP, above note 2, p. 116.

work, harnessing opportunities but also anticipating and minimizing possible negative impacts. This is where the imperative of "do no harm"[14] plays a crucial role. "Do no harm" requires humanitarian actors to endeavour not to cause further damage and suffering because of their actions.

First coined by Mary B. Anderson, this imperative is fleshed out in the first Protection Principle of the Humanitarian Charter and Minimum Standards in Humanitarian Response.[15] For its part, the ICRC emphasizes the "do no harm" imperative in its Protection Policy,[16] which highlights how it ensures that its actions do not have adverse impacts on, or create new risks for, individuals or populations. Beyond the ICRC, numerous humanitarian organizations have embedded "do no harm" into their work and policies over the decades, including when it comes to digital transformation processes, such as the digitalization of cash. Oxfam was one of the first agencies to develop an explicit policy on data responsibility[17] and has produced toolkits and training for staff to help operationalize data responsibility throughout its programmes.[18] More recently, the United Nations (UN) Office for the Coordination of Humanitarian Affairs (OCHA) Centre for Humanitarian Data's work on data responsibility applies a "do no harm" lens in creating tools and guidance to help staff navigate the technical and ethical aspects of working with humanitarian data, including but not limited to data generated by digital payments.[19]

Examining the digitalization of cash through the framework of "do no harm" – otherwise referred to as "do no *digital* harm" – is crucial as it allows humanitarian organizations to act ethically as we proceed in integrating digital transformations.

"Do no digital harm" is a critical imperative particularly in relation to the way humanitarian organizations and their partners manage data, implement activities and connect with affected people in the digital space. This applies equally to the way we use digital payments to support crisis-affected people. In the case of the ICRC, this means that not only does the organization use cash and vouchers (both physical and digital) in its responses, but it also has a clear understanding of how the digitalization of cash could impact humanitarian action, particularly in situations of armed conflict. The ICRC knows that "it is not enough to understand only the physical environment of armed conflict. It is

14 Mary B. Anderson, *Do No Harm: How Aid Can Support Peace or War*, Lynne Rienner, Boulder, CO, 1999. For a recent foundational text, see Hugo Slim, *Humanitarian Ethics: A Guide to the Morality of Aid in War and Disaster*, Oxford University Press, Oxford, 2015.

15 Sphere Project, *The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response*, 2018, Protection Principle 1, available at: https://handbook.spherestandards.org/en/sphere/#ch004_002_002.

16 ICRC, "ICRC Protection Policy", *International Review of the Red Cross*, Vol. 90, No. 871, 2008, p. 753, available at: www.icrc.org/en/doc/assets/files/other/irrc-871-icrc-protection-policy.pdf.

17 Oxfam, "Responsible Program Data Policy", February 2015, available at: https://policy-practice.oxfam.org.uk/publications/oxfam-responsible-program-data-policy-575950.

18 Oxfam, "Responsible Data Management", 2017, available at: https://policy-practice.oxfam.org.uk/our-approach/toolkits-and-guidelines/responsible-data-management.

19 OCHA, Centre for Humanitarian Data, "Data Responsibility", available at: https://centre.humdata.org/data-responsibility/.

essential to overlay this with readings of the virtual or digital environment."[20] To harness the opportunities that digitalization brings – for cash, but also beyond – the ICRC's implementation of the "do no harm" imperative is centred on ensuring that we keep the people we serve at the centre of our analysis.

CVA supporting people-centred responses

The advantages of CVA are well known: it increases affected people's dignity, power, autonomy and choice in how they manage their survival and recovery.[21] When it comes to putting affected people at the centre of our work, CVA is a critical tool because it puts financial resources in the hands of crisis-affected people so that they can recover, whenever and however they choose. In so doing, it enables affected people to make their own decisions – decisions which may well be different from those a humanitarian organization would make on their behalf. CVA has the potential to be transformative because it starts to change the balance of power between humanitarian agencies and crisis-affected people, as well as the multiple other stakeholders involved in humanitarian responses: donors, governments and civil society. When agencies are well prepared, CVA can be delivered quickly and at scale. Additionally, CVA can provide greater operational flexibility and achieve wider social and economic multiplier effects beyond its specific purpose, compared with in-kind assistance (giving goods directly). This is due to the indirect effects of cash transfers whereby increased expenditure by recipients contributes to income growth for non-recipients, expansion of markets for local goods, and increased demand for services.

Nevertheless, the use of CVA is subject to risks similar to those involved in providing in-kind assistance – market interference, accountability tracking, social tensions, protection issues – and these need to be managed carefully. In this respect, the digitalization of CVA can enhance both the benefits and risks for affected people. This is the focus of this analysis going forward: to highlight how the digitalization of cash is impacting humanitarian action, what the potential risks of the digitalization of cash are, and how we in the humanitarian sector can work to mitigate those risks.

How the digitalization of cash is transforming humanitarian action

However we look at it, digital payments are here to stay – both in the wider society and in humanitarian action. According to an analysis from MasterCard, "the ways

20 ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict, London, 11–12 December 2018*, Geneva, 2019, p. 1, available at: www.icrc.org/en/publication/4403-symposium-report-digital-risks-armed-conflicts.

21 Overseas Development Institute (ODI), *Doing Cash Differently: How Cash Transfers can Transform Humanitarian Aid. Report of the High Level Panel on Humanitarian Cash Transfers*, London, September 2015.

we pay for things has been changing more in the past 15 years than in the previous 150, and nearly every innovation we have seen has taken share away from cash."[22] Although when looked at from a global perspective, cash remains the most commonly used method of payment, there has been a significant increase in the growth of digital payments, with several countries moving rapidly towards becoming "cashless".

Why digital payments?

This has been mirrored in a shift away from cash payments and towards digital payments in the humanitarian sector. For example, in the International Red Cross and Red Crescent Movement, more than half of all payments to affected people are today made digitally.[23] This is in part because of the general global shift towards digitalization, but also because digital payments, transfers and remittances are considered by many actors, such as the World Bank, as contributing to the G20 goals of broad-based economic growth, financial inclusion and women's economic empowerment.[24] The Better than Cash Alliance (BTCA)[25] examined five main drivers behind the rise in digital payments, and concluded that digital payments bring:

- Cost savings through increased efficiency and speed
- Transparency and security by increasing accountability and tracking, reducing corruption and theft as a result
- Financial inclusion by advancing access to a range of financial services, including savings accounts and insurance products
- Women's economic participation by giving women more control over their financial lives and improving economic opportunities
- Inclusive growth through building the institutions that form the bedrock of an economy and the cumulative effect of cost savings, increased transparency, financial inclusion, and greater women's economic participation.[26]

Additional factors have played a role in the shift towards digital payments, including safety and security concerns, as handling large amounts of cash can be very visible and can put recipients or humanitarian staff delivering the cash at risk of theft or looting – much as can be the case with in-kind assistance. For the recipient, holding money in an account rather than keeping cash under their mattress is generally safer. Another important factor is the recipient's preferences – where

22 Hugh Thomas, "Measuring Progress towards a Cashless Society", MasterCard Advisors, available at: https://newsroom.mastercard.com/wp-content/uploads/2014/08/MasterCardAdvisors-CashlessSociety-July-20146.pdf.

23 See the Cash Hub interactive cash maps, available at: www.cash-hub.org/resources/cash-maps.

24 Leora Klapper and Dorothe Singer, *The Opportunities of Digitizing Payments*, World Bank, Washington, DC, 28 August 2015, p. 91.

25 The BTCA is a partnership of seventy-five governments, companies and international organizations intended to accelerate the "transition from cash to digital payments in order to reduce poverty and drive inclusive growth". See the BTCA website, available at: www.betterthancash.org.

26 BTCA, "Why Digital Payments?", available at: www.betterthancash.org/why-digital-payments.

people already use digital payments in their daily lives, it makes sense to use the same mechanisms that they are familiar and comfortable with.

At the time of writing, almost a year into the COVID-19 pandemic, a new driver can potentially be added: the perception that digital payments decrease the risk of virus transmission. Although it has been concluded in previous studies looking at influenza that "handling banknotes and coins is not practically avoidable and will confer no discernible increased risk compared with handling almost any other communal object used in daily life",[27] there has been a general preference for digital payments during the COVID-19 pandemic. Digital payments, such as mobile money transactions or using debit or credit cards, require less handling than physical currency. Additionally, the hardware involved (cards, mobile phones, point-of-sale devices etc.) can be regularly cleaned with a simple disinfectant. A guidance note on CVA and health during COVID-19 issued by the World Health Organization (WHO) and the Global Health Cluster emphasized the preference for digital payments, which reduce the need for people to gather at distribution points and allow for regular disinfection of surfaces such as ATM keypads, noting that "where this is possible, contact-less electronic or mobile payments should be the preferred option to reduce the risk of transmission".[28]

How humanitarians have embraced digital payments

Although different humanitarian organizations have used cash assistance over the decades, it was the formation of the Cash Learning Partnership (CaLP) in 2005, to promote and improve CVA across the humanitarian sector, which first structured the conversation around CVA, and with it, digital payments. By the early 2010s, and with increasing evidence of the effectiveness of CVA behind us, there was a convergence, and several collaborative initiatives emerged. The BTCA, which launched in 2012 in response to public and private sector demand, continues to provide strategic advocacy, research and guidance on how to transition to electronic payments. 2016 saw the launch of the ELAN, and during a four-year project this network produced numerous valuable resources for practitioners, hosted learning events, and contributed to growth and foresight in the humanitarian sector's work on digital payments. This period also saw a significant investment by the world's largest humanitarian network, the International Red Cross and Red Crescent Movement, in building its capacity to

27 European Centre for Disease Prevention and Control, *Technical Report of the Scientific Panel on Influenza in Reply to Eight Questions concerning Avian Flu*, Stockholm, 5 June 2006, p. 26, available at: www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/0606_TER_Eight_Questions_Concerning_Avian_Flu.pdf.

28 WHO and the Global Health Cluster, *Guidance Note on the Role of Cash and Voucher Assistance to Reduce Financial Barriers in the Response to the COVID-19 Pandemic, in Countries Targeted by the Global Humanitarian Response Plan COVID-19*, Geneva, April 2020, p. 8, available at: www.who.int/health-cluster/about/work/task-teams/Guidance-note-CVA-COVID.pdf?ua=1.

deliver CVA,[29] with the majority of its payments being made digitally. Such approaches require a level of interoperability between the various humanitarian agencies carrying out this work, which means in terms of risk-benefit analysis, the entire humanitarian sector is affected by the use of digital payments in its operations.

As outlined above, digital payments are becoming more and more prominent in the humanitarian sector. However, this raises the question: what are the risks and benefits of this facet of digital transformation? In the following section, I will address this question by outlining how the ICRC analyzes when to use digital payments.

How the ICRC uses digital payments in armed conflict and other situations of violence

Like many others within the humanitarian sector, the ICRC has embraced the use of both physical and digital payments, based on a thorough analysis of the risks and benefits for people affected by armed conflict and other situations of violence. As will be highlighted in the following sections, not only has the ICRC been using CVA for over 100 years, but the organization also has a critical focus on digital risk and data protection.[30] Thus, the ICRC's work can serve as an example of how to examine digital risks in CVA.

A dip into the ICRC's archives shows that it was already using cash transfers during the First World War, when its International Prisoners of War Agency was in charge of processing money orders and forwarding registered letters, including money, sent to interned civilians and prisoners of war by their families.[31] Today, the ICRC uses CVA in the majority of its eighty-plus operations, and its experience demonstrates that it is possible to use CVA in armed conflicts and other situations of violence.[32] From 2012 to 2020, the ICRC has seen a 600% increase in the number of people reached with CVA across its programmes, ranging from small cash-in-hand transfers to help families pay for transport costs to visit detained relatives, to larger grants for rebuilding homes or starting income-generation activities. During this time there has been a huge increase in digital payments, with most large-scale transfers – those to large numbers of people – being made digitally.

The choice between (physical) cash payments or digital payments is made very pragmatically by ICRC teams, based on what resources and services are available in a given country. In South Sudan, for example, where there are extremely limited means of digital payments, small-scale payments – for example, to help a separated person pay for transport to be reunited with their family – are

29 International Red Cross and Red Crescent Movement, *Cash Transfer Programming: Guidelines for Mainstreaming and Preparedness*, Geneva, 2015, available at: https://cash-hub.org/wp-content/uploads/sites/3/2020/10/Cash-Transfer-Programming-Guidelines-for-Mainstreaming-and-Preparedness.pdf.

30 See "Q&A: Humanitarian Operations, the Spread of Harmful Information and Data Protection", in this issue of the *Review*.

31 ICRC, *L'agence internationale des prisoners de guerre, Genève, 1914–1918*, Geneva, 1919, p. 105.

32 ICRC, above note 4, p. 7.

made in cash directly to the affected person. Compare this to Somalia, where the ICRC makes most of its payments digitally as the population have long been using mobile money. The ICRC asks itself what payment instruments people are familiar with, which ones they have access to, which will be fast, efficient and safe for affected people and for ICRC staff alike. This sometimes means using multiple payment methods in one country. For example, in the Democratic Republic of the Congo, the ICRC uses three: cash payments directly to people, digital payments via mobile money, and digital payments via personal accounts held with cooperatives.

Of course, it is not a choice between digital payments or no payments at all; cash remains a valid payment instrument in many cases. In Ukraine, to accompany its work supporting families with the search for missing persons, strengthening the authorities' capabilities to conduct the search according to international standards, and providing psychosocial, legal and administrative support to families, the ICRC also provided monthly payments to enable the families to better meet their basic needs. These payments were made in cash directly to the families by ICRC staff, in part because this gave the staff a practical reason to visit the families regularly. ICRC staff noted that they would not feel comfortable visiting the families every month if they had no new information about the missing person, whereas making the cash payments provided a reason to visit. The team considered that with digital payments, this vital contact would have been much less frequent and the relationships – and trust – with the families may have taken longer to build.

What are the risks of digital payments in humanitarian action?

Understanding the possible risks and potential pitfalls of digital payments in humanitarian settings and taking steps to mitigate them is a core part of the critical broader work of redefining what the "do no harm" principle means in our increasingly digitalized world. In fact, digital payments have already mitigated some of the traditionally perceived risks of CVA: evaluations of programmes using digital payments have shown that these programmes have reduced theft, reduced risks for staff in transporting money, and proved popular with recipients due to the privacy they afford.[33] However, it is important to start with the understanding that it is impossible to eliminate all risk; the risks in using digital payments can never be entirely avoided, only mitigated. Risk is a part of daily life – we take a risk just crossing the road in the morning. Humanitarian organizations must weigh the benefits to the recipient (as well as the organization) of the speed and efficiency of digital payments, versus the potential risks explored below. Therefore, risk assessment remains vital: allowing the identification and analysis of risks, understanding which can be mitigated and which cannot, putting relevant

33 Laura Gordon, *Risk and Humanitarian Cash Transfer Programming: Background Note for the High Level Panel on Humanitarian Cash Transfers*, ODI, London, May 2015, available at: www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9727.pdf.

measures in place, and documenting these decisions, thus also improving accountability and promoting learning.

Digital payments are not universally accessible.

We must start this examination of "do no digital harm" with a focus on the affected people themselves. The fact is, digital payments are not accessible to everyone in the same way that cash is, and the increase in digital payments may deepen the "digital divide" (defined as inequality in access to technology and its associated benefits). Any person can use cash if they can get their hands on it, and providing the goods and services they need to pay for are available. For the recipient, cash does not require any formal identification or any physical infrastructure. Cash is not linked to the person, and as such does not discriminate or identify.[34] It simply requires the coins or notes that are the means of exchange, and a basic level of numeracy.[35] However, to use digital payments, the recipient will require a level of digital and financial literacy. It is estimated that only one in three adults globally shows an understanding of basic financial concepts, and that there are lower standards of financial literacy amongst women and the poor.[36] The recipient will also require a personal account with a financial service provider and will be subject to Know Your Customer (KYC)[37] regulations set globally by the Financial Action Task Force (FATF), transposed into national law by States and applied by every commercial service provider. While the FATF requires "identifying the customer and verifying that customer's identity using reliable, independent source documents, data or information",[38] most States transpose this to mean a legally accepted form of identification such as a passport, birth certificate or national identity card. According to the World Bank, an estimated 1 billion people – that is, 13% of the world's population – do not have an official proof of identity.[39] Some 91% of these live in low-income or lower middle-income

34 AGIS Consulting, *Cash Essentials: Beyond Payments*, Paris, 2015, available at: https://cashessentials.org/?ref=xranks.

35 It has been argued that cash is the preferred means for criminal activities seeking to avoid detection. However, digitalization is changing this, with reports that "electronic money laundering, also known as Transaction Laundering, is the most common, but least enforced, method of money laundering". Ron Teicher, "Transaction Laundering – Money Laundering Goes Electronic in the 21st Century", *Finextra*, 4 June 2018, available at: www.finextra.com/blogposting/15423/transaction-laundering—money-laundering-goes-electronic-in-the-21st-century.

36 Leora Klapper, Annamaria Lusardi and Peter van Oudhuesden, *Financial Literacy around the world: Insights from Standard & Poor's Ratings Services Global Financial Literacy Study*, 2015, available at: https://gflec.org/wp-content/uploads/2015/11/3313-Finlit_Report_FINAL-5.11.16.pdf?x28148.

37 KYC is a process enabling businesses to check the identity of their customers in order to comply with regulations and legislation on money laundering and corruption, and includes collecting information from the customer such as name, identity document number, phone number and address. PwC, *Know Your Customer: Quick Reference Guide*, January 2016, available at: www.pwc.lu/en/anti-money-laundering/docs/pwc-kyc-qrg-final-interactive-2016.pdf.

38 FATF, "FATF Recommendation 5: Customer Due Diligence and Record-Keeping", available at: www.un.org/sc/ctc/wp-content/uploads/2016/03/fatf-rec05.pdf.

39 World Bank, "ID4D Data: Global Identification Challenge by the Numbers", available at: https://id4d.worldbank.org/global-dataset.

countries, and among these, one in two are women in low-income countries.[40] There are many reasons why people do not have an official proof of identity, including low levels of literacy, the often high costs of official documents (which are a barrier for the poorest), legal requirements that vary for different sections of the population in a given country, absence or paucity of national identity management systems or lack of resources committed to ensuring that all citizens are registered, or simply that it is not common practice in a community to register with the authorities. These issues are often exacerbated by armed conflict and other situations of violence, where State services may be disrupted and shifting frontlines inhibit movement of people and change geographical boundaries, sometimes leaving people living in territories where their identity documents are no longer recognized as valid.

Of course, if communities have no access to digital payments, humanitarians can choose to provide cash payments, vouchers, in-kind assistance or services. The problem arises when most of a population do have access, leading agencies to choose digital payments and thus resulting in the possible – albeit unintentional – exclusion of those groups without access. Humanitarians must always have the flexibility to offer a variety of solutions, in order to ensure that no one is left behind.

Digital payments are often promoted based on the concept that they drive financial inclusion; however, financial inclusion does not necessarily translate to financial well-being. Clear links have been established between increased financial inclusion, poverty reduction and economic security, but it is not as simple as ensuring that people have access to financial services. A 2017 study conducted by the ICRC and the British Red Cross in Nigeria and Kenya found that, in these contexts, "people's main problem remains poverty and not financial inclusion".[41] The provision of digital payment options will not, in and of itself, automatically lead to financial inclusion.

The increased focus on linking humanitarian cash assistance to social protection adds governments – the main providers of social protection systems – into the mix, bringing complicated protection and coordination concerns. This focus is only increasing due to the COVID-19 pandemic and the predicted global economic depression, which will stretch existing governmental social protection and humanitarian aid systems to the limits of their capacities and beyond. However, making these links is not a straightforward matter; the findings of a recent practitioner survey[42] show that the main perceived challenges of linking CVA with social protection systems remain a lack of coordination between actors, the fact that social protection systems are not designed to be shock-

40 Vyjyanti T. Desai, Anna Diofasi and Jing Lu, "The Global Identification Challenge: Who Are the 1 Billion People without Proof of Identity?", *Voices: World Bank Blogs*, 25 April 2018, available at: https://blogs.worldbank.org/voices/global-identification-challenge-who-are-1-billion-people-without-proof-identity.

41 Paul Harvey, Kokoévi Sossouvi and Annie Hurlstone, *Humanitarian Cash and Financial Inclusion: Findings from Red Cross Movement Projects in Kenya and Nigeria*, British Red Cross and ICRC, London, February 2018, p. 6.

42 CaLP, above note 2, p. 144.

responsive, lack of experience of humanitarians in social protection, and that governments are not perceived as impartial in addressing the needs of the most vulnerable. As the UN has noted:

> Many laws that formally restricted access to social protection and public services to certain population groups have been repealed. Nevertheless, discrimination continues to reinforce some of the barriers they face, including a lack of information on entitlements or the political voice or representation necessary to claim such entitlements.[43]

These exclusions can be exacerbated in times of crisis, and particularly in conflict. In contexts where the rule of law is weak and corruption is endemic, certain people and groups will benefit – and some will suffer – from the changes to the economic model brought about by conflict. Humanitarians have a role to ensure that marginalized groups and those that face protection risks are supported, whether that be through existing systems or through humanitarian protection and assistance activities.

There is also a risk of conflating digital and physical proximity. Digital payments may allow vital assistance to be provided to communities remotely, but this does not mean that humanitarians do not need to be present in communities; field teams still need to conduct assessments in order to get a true picture of needs and then follow these up with monitoring and evaluation to gauge the impact of interventions on the affected communities. Digital proximity – for example, through digital payments – "does not replace the need for physical access to vulnerable communities nor can it replace wider efforts to ensure they enjoy protection under relevant laws".[44]

Data responsibility in humanitarian action

To ensure inclusion in digital payments, humanitarian organizations must collect and process an enormous amount of data, including the personal data of affected people who wish to access those payments. Whether using a "closed loop" system managed entirely by the humanitarian agency, or working through local financial service providers, this data must be securely collected, managed, stored and shared in line with good data management practices.

CaLP's 2020 *State of the World's Cash* report highlights that "digital risk and data management is a 'newly emerged risk'" and notes that "whilst some large operational CVA actors (by their own admission) are now on top of responsible data management, many CVA practitioners still find this a paralysing topic".[45]

43 UN, *Promoting Inclusion through Social Protection: Report on the World Social Situation 2018*, New York, 2018, p. 18, available at: www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2018/07/1-1.pdf.
44 ICRC, above note 4, p. 9.
45 CaLP, above note 2, p. 52.

A series of events in 2019,[46] some focusing specifically on CVA and others on humanitarian action more broadly, explored digital risks and the concepts of "do no digital harm" and "digital dignity" as part of the digital transformation of humanitarian action. Several themes emerge when examining digital risks in humanitarian action, and more specifically the risks associated with digital payments, but fundamentally they are all rooted in the core topic of data responsibility, "a set of principles, processes and tools that support the safe, ethical and effective management of data in humanitarian response. This includes data privacy, protection, and security, as well as other practical measures to mitigate risk and prevent harm."[47] Data responsibility requires humanitarians to collect, manage, store and share data conscientiously.

The issue of data responsibility is frequently raised when aiming for collaboration between humanitarian agencies. There have been several initiatives related to improving collaboration for delivery of CVA, including the Collaborative Cash Delivery Network[48] and the UN common cash system.[49] Such collaborative approaches have been broadly welcomed in the spirit of fostering greater collaboration, in line with Grand Bargain[50] commitments to increase the effectiveness, efficiency and accountability of CVA operations and provide better support to those affected by crises.[51] However, they have experienced various challenges. Questions have been raised around intellectual property rights in co-created systems, the complexity of data interoperability, and issues of power and resource control between different agencies. Lack of data interoperability and data sharing between agencies that are using collaborative platforms is a significant impediment to programme quality, contributing to delays in programming and meaning that use of information cannot be maximized. Data sharing requires trust in each other's systems and data management and protection practices. In February 2019, the World Food Programme (WFP) and data analytics and

46 The ICRC convened a symposium on "Digital Risks in Armed Conflicts" in December 2018 in London: see ICRC, above note 20. CaLP convened a data responsibility workshop in April 2019 in Geneva: see CaLP, "Data Responsibility: Let's Not Wait for Another Wake-Up Call", 8 May 2019, available at: www.calpnetwork.org/blog/data-responsibility-lets-not-wait-for-another-wake-up-call/. Findings from the CaLP workshop were taken into a first meeting on "Data Responsibility in Humanitarian Action" convened by the OCHA Centre for Humanitarian Data, in collaboration with Wilton Park, in May 2019: see Wilton Park, "Data Responsibility in Humanitarian Action: From Principle to Practice", available at: www.wiltonpark.org.uk/event/wp1688/. A second Wilton Park event in October 2019 focused more broadly on "Digital Dignity in Armed Conflict": see Wilton Park, *Digital Dignity in Armed Conflict: A Roadmap for Principled Humanitarian Action in the Age of Digital Transformation*, October 2019, available at: www.wiltonpark.org.uk/wp-content/uploads/WP1698-Report.pdf.

47 Wilton Park, *Data Responsibility in Humanitarian Action: From Principle to Practice*, June 2019, available at: www.wiltonpark.org.uk/wp-content/uploads/WP1688-Report-1.pdf.

48 See Collaborative Cash Delivery Network, "Our Story", available at: www.collaborativecash.org/the-network.

49 See Inter-Agency Standing Committee (IASC), "Statement from the Principals of OCHA, UNHCR, WFP, and UNICEF on Cash Assistance", 5 December 2018, available at: https://interagencystandingcommittee.org/other/content/statement-principals-ocha-unhcr-wfp-and-unicef-cash-assistance-5-december-2018.

50 See the official Grand Bargain website, available at: https://interagencystandingcommittee.org/grand-bargain.

51 IASC, "Increase the Use and Coordination of Cash-Based Programming", available at: https://interagencystandingcommittee.org/increase-the-use-and-coordination-of-cash-based-programming.

intelligence-gathering firm Palantir Technologies signed a $45 million partnership, attracting criticism

> from human rights and data transparency advocates, who argued that Palantir has facilitated rights abuses through its previous work with organisations including the Central Intelligence Agency (CIA), Immigration and Customs Enforcement (ICE) and Cambridge Analytica. … They argued that, in the name of increased efficiency and cost savings, the highly sensitive data of the 92 million people served annually by the WFP was being put at risk.[52]

The partnership raised questions about the implications for the UN common cash system and how data collected by any organization participating will be protected, along with how this could impact the affected people that the system seeks to serve.

The rise of digital identities and the use of biometrics

Technology-assisted innovation in digital identity has increased in recent years and is often used in programmes providing CVA. But what is digital identity?

> A digital identity is a collection of electronically captured and stored identity attributes that uniquely describe a person …. A person's digital identity may be composed of a variety of attributes, including biographic data (e.g., name, age, gender, address) and biometric data (e.g., fingerprints, iris scans, hand prints) as well as other attributes …. [These data] can be used to identify a person by answering the question "who are you?"[53]

Combined with other credentials, they can also answer the question, "Are you who you claim to be?"

A digital identity is not automatically an official or legal proof of identity; only governments can provide legal identification to citizens, although many are choosing to do so digitally. In Estonia, seen as the world leader in digital integration, it is mandatory for every Estonian citizen above the age of 15, and every European citizen residing in Estonia, to obtain the Estonian digital identity card.[54] India has the world's single largest biometric-based digital identification system, called Aadhaar.[55]

The digital identities created by humanitarian organizations, enabling people to have access to their programmes, are therefore not official legal identities, and as such have limited use beyond their specified purpose, unless negotiated with the relevant authorities (such as in the case of refugee ID cards

52 Barnaby Willitts-King, John Bryant and Kerrie Holloway, *The Humanitarian "Digital Divide"*, Humanitarian Policy Group Working Paper, ODI, London, November 2019, p. 15.

53 World Bank Group, GSMA and Secure Identity Alliance, *Digital Identity: Towards Shared Principles for Public and Private Sector Cooperation*, July 2016, available at: www.gsma.com/mobilefordevelopment/wp-content/uploads/2016/07/Towards-Shared-Principles-for-Public-and-Private-Sector-Cooperation.pdf.

54 World Bank Group, *Privacy by Design: Current Practices in Estonia, India, and Austria*, Washington, DC, 2018, available at: https://id4d.worldbank.org/sites/id4d.worldbank.org/files/PrivacyByDesign_112918web.pdf.

55 *Ibid.*

issued by the Office of the UN High Commissioner for Refugees (UNHCR), which provide access to services beyond those provided by the UNHCR, including financial services in some cases[56]). The creation of digital identities also raises the question of ownership of those identities. The 2019 Wilton Park conference on "Digital Dignity in Armed Conflict" concluded that "to promote digital dignity, individuals who receive aid should be perceived as data agents who have agency over their digital identity and digital anonymity".[57] There could be significant risks in creating interoperable systems and identity frameworks for vulnerable groups, who may have very good reasons for wishing to remain anonymous, and who could face discrimination or worse by virtue of their being identifiable.

Biometrics are one such form of digital identity. Biometrics were primarily developed and deployed for the purposes of border and migration control, and were then widely deployed in updated national identity systems. It is the drive for legal identity meeting due diligence requirements that makes biometrics so attractive, because they cover both needs. However, this blurring of the boundaries between immigration control and counterterrorism or security makes biometrics collected by private actors and humanitarian organizations particularly interesting for States.[58]

The use of biometric identification systems by humanitarian organizations to support digital identity has significantly increased, due to the perception that they bring efficiency and accountability to operations, in particular with regard to reducing fraud.[59] However, this rapid uptake of biometrics has caused much debate. The early use of biometrics by the UNHCR was seen as a success story, but a UN internal audit in 2016 found that in four out of five country operations reviewed, the information being given to refugees about the biometric programme was insufficient for them to be properly informed. [60] As an example, in response to the Rohingya refugee crisis, biometrics enabled the UNHCR and its partners to cope with a crisis of enormous scale and speed, but some have argued that this has created risks for the Rohingya refugees:

> The biometric data which humanitarian agencies and the government have been collecting is not only being used to distribute aid to the Rohingya people; it is also being used to control their movements. … The fear for the

56 UNHCR, "Documentation", available at: www.unhcr.org/registration-guidance/chapter5/documentation/.
57 Wilton Park, *Digital Dignity in Armed Conflict*, above note 46, p. 4.
58 UN Security Council Resolution 2396 requires all States to use "biometric data, which could include fingerprints, photographs, facial recognition, and other relevant identifying biometric data, in order to responsibly and properly identify terrorists, including foreign terrorist fighters". See UNSC Res. 2396, 21 December 2017, available at: http://unscr.com/en/resolutions/doc/2396; Fionnuala Ní Aoláin, "The UN Security Council, Global Watch Lists, Biometrics, and the Threat to the Rule of Law", *Just Security*, 17 January 2018, available at: www.justsecurity.org/51075/security-council-global-watch-lists-biometrics/.
59 The Engine Room and Oxfam, *Biometrics in the Humanitarian Sector*, March 2018, p. 8, available at: www.theengineroom.org/biometric-tech-review-report/.
60 Elsie Thomas, "Tagged, Tracked and in Danger: How the Rohingya Got Caught in the UN's Risky Biometric Database", *Wired*, 12 March 2018, available at: www.wired.co.uk/article/united-nations-refugees-biometric-database-rohingya-myanmar-bangladesh.

> Rohingya is that this biometrically-enabled control system could be used to send them back to Myanmar.[61]

Taking a measured approach, in 2015 Oxfam imposed a moratorium on the use of biometrics in its work,[62] later stating that "given the number of unknowns … we felt it was best not to become an early adopter".[63] The ICRC has long used biometrics, but only to support the implementation of its mandate where particular objectives cannot be realized without using them, for example in forensics and the restoration of family links: "In this case the ICRC processes the biometric data as a matter of 'public interest' (in the implementation of the ICRC's mandate)."[64] When it comes to using biometrics for beneficiary management and assistance distribution, the ICRC's stance differs:

> Because the purpose here is primarily linked to efficiency, and insofar as aid can be (and long has been) distributed without the need for biometrics, the ICRC would have to establish that the "legitimate interest" it has in establishing any biometric identity management system does not outweigh the potential impact on the rights and freedoms of the individuals concerned.[65]

While the use of biometrics is increasing in daily life – fingerprint scans are promoted as an easy way to secure your smartphone, for example – it can still be possible to "opt out". However, increasingly in humanitarian programming, this possibility is somewhat theoretical, as proving your identity is linked to assistance delivery. Biometrics in and of themselves are not inherently risky, but as we will explore below, the processing of any personal data is sensitive and subject to various risks and concerns.

Data protection: Protecting life, integrity and dignity

The risks highlighted so far have the issue of identity at their core: how do people prove who they are, who owns that identity, and what is done with the identity data that people provide? Directly related to the issue of identity is that of data protection. Anyone with an email account will remember the introduction of the European Union's General Data Protection Regulation, which became enforceable in May 2018, and with it the flood of emails asking us to give our consent to our data being used. Who among us ever reads the reams of pages of terms and conditions before shopping online or downloading an application on our smartphones? People give away their data freely all the time without necessarily considering the implications, the desire for convenience and connectivity often overriding the potential risks, even when their trust in

61 *Ibid.*
62 The Engine Room and Oxfam, above note 59.
63 E. Thomas, above note 60.
64 Ben Hayes and Massimo Marelli, "Facilitating Innovation, Ensuring Protection: The ICRC Biometrics Policy", *Humanitarian Law and Policy Blog*, 18 October 2019, available at: https://blogs.icrc.org/law-and-policy/2019/10/18/innovation-protection-icrc-biometrics-policy/.
65 *Ibid.*

companies or governments is low. For example, despite initial concerns after the Facebook/Cambridge Analytica privacy scandal,[66] and with very little actual improvement to privacy for its users, Facebook's user numbers continue to climb, with a 12% rise year on year.[67]

Protecting individuals' personal data is an essential part of protecting their life, integrity and dignity, which makes the protection of personal data of fundamental importance for humanitarian organizations. Humanitarians frequently collect personal data in order to carry out their mandate, whether that be to trace missing people, to help reunite separated families or to provide life-saving assistance. When it comes to digital payments, data protection is a critical issue because in order to make or receive payments through a financial service provider, recipients must identify themselves, and humanitarian organizations must therefore collect and share people's personal data. As new technologies allow for exponentially faster and easier processing of this data, there are increasing concerns about privacy. While the right to privacy has been recognized globally as a human right since 1948,[68] the right to have one's personal data protected is more recent, with the first regional data protection treaty entering into force almost forty years later.[69] Today, a majority of States, the UN and international organizations with a humanitarian mandate recognize core data protection principles, even if the content of policies and scope of legislation varies across the world.[70] However, the very circumstances within which humanitarian organizations operate – crisis situations – create unique challenges regarding data protection.

For humanitarians, obtaining the consent of recipients for the processing of their personal data is one of the first challenges. In the context of data protection, consent means the freely given, specific and informed indication of a data subject's (i.e., the recipient's) wishes by which they signify agreement to personal data relating to themselves being processed. This means that "the individual must be put in a position to fully appreciate the risks and benefits of data Processing, otherwise Consent may not be considered valid".[71] Using consent as a legal basis for data processing differs from the way the humanitarian sector has traditionally

66 Julia Carrie Wong, "The Cambridge Analytica Scandal Changed the World – but It Didn't Change Facebook", *The Guardian*, 18 March 2019, available at: www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook.

67 Dan Noyes, "The Top 20 Valuable Facebook Statistics", Zephoria, October 2020, available at: https://zephoria.com/top-15-valuable-facebook-statistics/.

68 See Universal Declaration of Human Rights, UN Doc. 217 A (III), 10 December 1948, Art. 12; International Covenant on Civil and Political Rights, 999 UNTS 171, 16 December 1966 (entered into force 23 March 1976), Art. 17.

69 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No. 108, 28 January 1981 (entered into force 1 October 1985), available at: www.coe.int/en/web/conventions/full-list/-/conventions/treaty/108.

70 United Nations Conference on Trade and Development, "Data Protection and Privacy Legislation Worldwide", available at: https://unctad.org/page/data-protection-and-privacy-legislation-worldwide.

71 Christopher Kuner and Massimo Marelli (eds), *Handbook on Data Protection in Humanitarian Action*, 2nd ed., ICRC and Brussels Privacy Hub, Geneva, June 2020, p. 51, available at: www.icrc.org/en/data-protection-humanitarian-action-handbook.

used consent as a basis for its humanitarian responses. Data processing by humanitarian organizations – including the ICRC – is often "based on vital interest or on important grounds of public interest, for example in the performance of a mandate established under national or international law".[72] This means that consent does not have to be valid for data processing to go ahead, provided that the processing is being carried out on the legal basis of "public interest".

However, we must recognize that the very concept of freely given consent in a crisis is a misnomer, particularly if consenting to the processing of personal data is a precondition for receiving assistance. If you had lost everything and someone was offering to help you, you would likely accept that help with very few questions asked and would pay very little heed to complex legal questions about what might happen to your data and what the potential future risks may or may not be. Ensuring informed consent is also a challenge when data protection policies include complex legal concepts which many of us would struggle to understand, regardless of levels of literacy and education. One of the key recommendations from the 2019 Wilton Park conference was that:

> The notion of informed consent should be reformulated as meaningful consent, addressing coercive consent practices. Alternatives to the current framework of informed consent should provide an opt-out for people who choose not to provide information. This should not prevent access to services. An opt out will shift the power dynamic and go some way to addressing the relationship imbalance.[73]

Of course, data protection is important across the full spectrum of humanitarian work, but it has a particular resonance when it comes to digital payments. The use of digital payments requires humanitarians to outsource part of their supply chain, which can be positive in terms of efficiency and accessing specific expertise, but also requires giving up a level of control in terms of the management of data. The very nature of digital payments requires sharing of data with a third party – the financial service provider who will facilitate the digital payment. Once transferred, the data is therefore no longer under the humanitarian organization's control.

Key concerns relating to data protection arise when using digital payments through financial service providers, who are bound by national legislation implementing FATF recommendations.[74] Concerns include the use of data by authorities for law enforcement purposes, including surveillance and profiling of individuals, and the use of personal data for commercial purposes, such as service providers offering targeted services or advertisements, or profiling customers for

72 *Ibid.*, p. 60.
73 Wilton Park, *Digital Dignity in Armed Conflict*, above note 46.
74 It is important to note that FATF recommendations do require data retention, access for law enforcement, the establishment of financial intelligence units, the submission of suspicious financial transaction reports, etc. See FATF, *The FATF Recommendations*, February 2012 (updated October 2020), available at: www.fatf-gafi.org/publications/fatfrecommendations/documents/fatf-recommendations.html.

creditworthiness. Data may also be used to cross-check people against master lists of customer debts, potentially leading to the financial institution directly deducting sums owed from the humanitarian assistance a person is set to receive. There are times when data protection concerns mean that digital payments are not appropriate, even if they are available. This is particularly common in conflict environments and other sensitive settings, where the choice of financial service providers is often limited and tends towards control by one particular party to the conflict.[75]

In many cases, people already have a relationship with a financial service provider, such as a bank or a mobile money operator; these individuals have already accepted and fulfilled the necessary requirements (including KYC) to access the service, and as such accept the benefits and risks that come with this. While this does not absolve humanitarians of their responsibility to undertake the necessary due diligence, there is a difference between, on the one hand, leveraging an existing financial relationship that a recipient has, and, on the other, asking someone to commence a new financial relationship for the sole purpose of receiving assistance through digital payments.

While the risks to individuals are critically important, there is a second layer of risk. We cannot ignore the fact that the personal data collected by humanitarians may be accessed by authorities, via financial service providers, and used for law enforcement purposes, which can have potential risks for wider humanitarian action. As the ICRC has observed:

> These risks are not limited to individuals. Humanitarian organizations can also face such risks. If data generated by a humanitarian organization are then used for non-humanitarian purposes, whether law enforcement or commercial, the neutrality and independence of humanitarian action could be affected. A humanitarian organization may then be perceived as supporting one party to a conflict by providing data that could lead to security risks for the organization and/or loss of access to the population of concern.[76]

Both outcomes could eventually lead to a reduction in essential humanitarian services for the affected population.

Beyond the legitimate and legal use of data, the data collected, stored, shared and analyzed by humanitarian agencies can be valuable to parties to the conflict. Data which identify people, their locations, their networks and their connections can be used with malicious intent by individuals, groups or organizations. It is not only the data that we commonly understand as personal data (e.g. name, phone number) but also the often overlooked metadata – data that provides information about other data – which is a mine of information, for those who know how to use it. General Hayden, former director of the US National Security Agency and Central Intelligence Agency, articulated this most

75 ICRC, above note 4, p. 50.
76 *Ibid.*, p. 51.

clearly when he said "[w]e kill people based on metadata",[77] highlighting how metadata is used to make life-and-death decisions. In an age of increased surveillance, the protection risks for people of concern are clear: "[e]ven though software typically can't kill people directly in the way that bullets can, the end result is often the same".[78] Metadata has always existed – for example, the information contained on the outside of a letter which says nothing of its content but still provides information on the sender and recipient. However, with digital information flows and transactions – whether financial or of any other type – the quantities of metadata generated are huge. For example, the volume of metadata generated by mobile money transactions is immense. The issue is not simply the volume of this metadata but rather what can be inferred by it, including what social groups people belong to (if a particular group was targeted for assistance), where they may have moved after the crisis, and their "network of family or friends, based on transfers received or made that didn't involve the humanitarian organisation. Information can then be inferred about these individuals in turn, even though they were not directly involved" in the programme.[79]

The way in which metadata can be used for good while simultaneously eroding the privacy of the individual has been very clearly highlighted by the rapid development of contact tracing apps during the COVID-19 pandemic, many of which require users to upload personal information and to agree to their location being tracked and shared using different types of geolocation technology. As of July 2020, close to fifty countries were developing or had rolled out contact tracing apps to help fight the spread of the disease.[80] While some countries, such as South Korea, have credited the use of contact apps with "flattening the curve" of the disease, helping to bring the outbreak under control, there are real privacy concerns being raised. An analysis of COVID-related apps by Privacy International exposed some of the risks of adopting these technologies, not least that "apps are notorious for their lack of security and privacy safeguards, exploiting people's data and devices".[81] While billions of people around the world have accepted restrictions on their day-to-day life in the service of bringing the pandemic under control, there will be limits; now people are "being asked to trust governments with their proposed apps – of which there are many. These are the very same governments who have been keen to exploit data in the past."[82]

77 David Cole, "'We Kill People Based on Metadata'", *New York Review of Books*, 10 May 2014, available at: www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/. And see Johns Hopkins University, "The Price of Privacy: Re-Evaluating the NSA", Johns Hopkins Foreign Affairs Symposium, 2014, available at: www.youtube.com/watch?time_continue=1022&v=kV2HDM86XgI.

78 Ron Deibert, Citizen Lab, quoted in Stephanie Kirchgaessner, "'Cat and Mouse Game': How Citizen Lab Shone a Spotlight on Israeli Spyware Firm", *The Guardian*, 12 May 2020, available at: www.theguardian.com/world/2020/may/12/cat-and-mouse-game-how-citizen-lab-shone-a-spotlight-on-israeli-spyware-firm-nso.

79 ICRC and Privacy International, *The Humanitarian Metadata Problem: "Doing No Harm" in the Digital Era*, Geneva, October 2018, p. 74.

80 Niall McCarthy, "Which Countries Are Deploying Coronavirus Tracing Apps?", *Forbes*, 22 July 2020, available at: https://tinyurl.com/j8wse55q.

81 Privacy International, "There's an App for That: Coronavirus Apps", 20 April 2020, available at: https://privacyinternational.org/long-read/3675/theres-app-coronavirus-apps.

Reducing privacy standards in an emergency could be the start of a slippery slope; once standards have dropped, it can be difficult to raise them again.

Cyber security and the increase in cyber attacks

Digital payments also require an infrastructure, both digital and physical. Payments infrastructure requires computers and servers, which must physically be based somewhere and which require energy to run and resources to maintain. Infrastructure can break down, and it can also be damaged, particularly in situations of armed conflict.[83] If the electricity fails or the "computers are down", the economy may come to a standstill. In 2018, a massive outage of the Visa payment system left millions in Europe unable to make payments.[84] The same happened to MasterCard a few weeks later, disrupting digital payments globally.[85] A statement from Visa said the issue was not associated with any unauthorized access or cyber attack, but this is a real risk when it comes to critical infrastructure, and alongside health, water, energy and transport infrastructure, financial infrastructure is indeed critical. In 2019, financial services firms reported huge year-on-year increases in the number of cyber attacks, breaches and data thefts, with 25% of all malware attacks being made against financial services organizations, more than any of the other twenty-seven industries monitored.[86] The COVID-19 pandemic is likely to push more and more financial institutions to become fully digital, and with this increasing reliance on e-commerce and contactless payments, there should be continued investment in resilient payment systems.[87] However, humanitarian organizations are also at risk: "Humanitarian organisations collect, store, share and analyse data that is attractive to parties to armed conflict. … As a result, humanitarian organisations are exposed to a growing wave of digital attacks and cyber espionage, and have become highly prized targets."[88] In the summer of 2019, the UN was the target of a complex cyber attack: "'The attack resulted in a compromise of core infrastructure components', said UN spokesperson Stéphane Dujarric, who classified it as 'serious'. … The 'core infrastructure' affected included systems for user and password management, system controls, and security firewalls."[89] When it comes

82 *Ibid.*

83 On the protections that international humanitarian law affords against the effects of cyber operations during armed conflicts, see "Twenty Years On: International Humanitarian Law and the Protection of Civilians against the Effects of Cyber Operations during Armed Conflicts", in this issue of the *Review*.

84 Patrick Collins, "Visa Card Payments System Returns to Full Capacity after Crash", *The Guardian*, 2 June 2018.

85 Martin Arnold, "MasterCard Customers Suffer Outages around the World", *Financial Times*, 12 July 2018.

86 Hadar Rosenberg, *Banking and Financial Services: Cyber Threat Landscape Report*, IntSights, April 2019, p. 3.

87 World Economic Forum, *Impact of COVID-19 on the Global Financial System: Recommendations for Policy-Makers Based on Industry Practitioner Perspectives*, Geneva, April 2020, available at: www3.weforum.org/docs/WEF_Impact_of_COVID_19_on_the_Global_Financial_System_2020.pdf.

88 ICRC, above note 20, p. 12.

to digital payments, data collected to make payments to recipients can be at risk of "hacks" while held with the humanitarian organization, and again when held by the financial service provider.

Risk cannot be entirely eliminated

As noted above, the risks in using digital payments can never be entirely avoided, only mitigated. Critical risks including exacerbating the digital divide, increased cyber attacks and data breaches, unknown implications of the creation and use of digital identities, and wider concerns around data protection which must be weighed against the benefits of digital payments. Decisions must be taken mindfully and based on each country or population group, as the severity of impact can change for different groups in different situations. Affected people must have the option to opt out of providing personal data (biometric or otherwise) without it prejudicing their access to essential assistance. Risk analysis – a weighing of the cost–benefit parameters – should be done in consultation with affected people themselves. While the risk may be the same for different groups, the impact of that risk on people's daily lives can vary wildly, and we should listen to their perspectives when taking decisions. A Ground Truth Solutions study which found that "users want cash transfers delivered through mechanisms that are flexible, trustworthy and reliable"[90] dovetails with other evidence – both researched and anecdotal – that people prefer payment mechanisms with which they are familiar, whether those are digital or not. Expecting people in crisis situations to absorb new risk or trust unfamiliar systems is a big ask and must not be undertaken unless it is in their best interest and with their agreement. It is not only today's risks that are a concern; technology is developing much faster than we can keep pace with, and it is obvious that humanitarians cannot accurately imagine all possible future risks. Despite this, we must continue to examine the risks and benefits of digital solutions, including digital payments, while respecting the principle of "do no digital harm".

How to mitigate the risks of digital payments

While we have focused on some of the main risks of digital payments, it is worth remembering all the evidenced benefits that have led the humanitarian sector to embrace their use, particularly as the risks and potential pitfalls of digital payments described above can all be mitigated to a certain extent. The question that must be asked every time humanitarians choose digital payments is whether

89 Ben Parker, "The Cyber Attack the UN Tried to Keep under Wraps", *The New Humanitarian*, 29 January 2020, available at: www.thenewhumanitarian.org/investigation/2020/01/29/united-nations-cyber-attack.

90 Ground Truth Solutions, *Improving User Journeys for Humanitarian Cash Transfers*, December 2018, available at: https://groundtruthsolutions.org/wp-content/uploads/2018/12/User_Journeys_Summary-Report_2018.pdf.

those mitigation measures are sufficient to balance the advantages with the risks. We have noted that there is no such thing as risk-free action, and this is especially true in conflict settings. In weighing the risks and benefits of any action, we touch on one of central dilemmas of humanitarian action in crisis: the risk of acting versus the risk of doing nothing.

Closing the digital divide

If digital payments are not accessible to all, there are several measures that humanitarians can take, depending on the cause. Sometimes the issue is simply that people have never used digital payments before. In this case, basic financial literacy classes can be provided to ensure that people are comfortable with the payment instrument being used. For example, Mercy Corps "embraces a broad definition of financial inclusion, seeking to improve access [and] ensure quality and actual usage of financial products and services", and one of its key strategies is increased client-level financial capability.[91] Organizations such as the Center for Financial Inclusion[92] and the Consultative Group to Assist the Poor[93] were founded with the very objective of promoting financial inclusion and advocating for inclusive, responsible finance. The ICRC, where possible, tries to use financial service providers that people already hold personal accounts with, so that they are familiar and comfortable with the services, have already met the KYC requirements and are not being asked to trust unfamiliar systems. If the issue is that people cannot meet the KYC requirements, this can be harder to resolve. The ICRC can refer people to legal services to help them obtain missing documentation such as identity papers, and even cover the costs of these documents if needed. Humanitarian agencies such as the Danish Refugee Council, Norwegian Refugee Council, UNHCR and a host of local actors include legal advice and the provision of, or advocacy for, key documentation as part of their services. If people cannot access digital payments services, humanitarians must be ready to offer alternatives such as cash payments or in-kind assistance. It can be challenging to offer different options; for example, if digital payments are not a solution for many people, making cash payments to large numbers of people can be logistically complicated, and setting up a new pipeline to procure and deliver in-kind assistance instead will take time. Humanitarian organizations must ensure that they have the infrastructure and procedures in place for in-kind assistance, cash, vouchers or digital payments where appropriate, enabling them to switch if needed.

While many humanitarian organizations are using technologies to improve digital proximity to populations, such as harnessing different means of (two-way) digital communications, monitoring population movements remotely, or making digital payments so that people can pay for essential goods and services, they

91 Mercy Corps, *Financial Inclusion: Approach and Principles*, 2019, available at: https://tinyurl.com/1f9ejhfq.
92 See the Center for Financial Inclusion website, available at: www.centerforfinancialinclusion.org.
93 See the Consultative Group to Assist the Poor website, available at: http://cgap.org/.

must continue to balance both physical and digital proximity. The ICRC will continue to negotiate for unimpeded access to people affected by armed conflict. Being accountable to people affected by armed conflict is a key element of the organization's identity and the essence of its operational model, which is based on proximity to affected people.

Digital identities and personal data: protecting data "by design"

When it comes to digital identity, humanitarian organizations should take a measured approach to their engagement. If humanitarian organizations use biometrics, thus creating digital identities, steps must be taken to ensure that those digital identities are used only for the specified purposes and are securely managed. Affected people must have the information they need to make an informed decision and to have agency over their digital identity and digital anonymity, and must have the ability to opt out of having a digital identity created on their behalf, without it prejudicing their access to assistance. A middle-ground option may be provided by the development of different forms of "self-sovereign identities", such as blockchain-based identities where the data subject has a digital identity but has ultimate control over who can access it.

In discussing how to mitigate risks related to data protection, it is imperative to analyze the situation of affected people. Some people may be at greater risk of negative consequences due to their individual characteristics, such as their legal status, their socio-economic situation, their race, or their religious or political affiliations. This can be exacerbated in situations of armed conflict, which often fall along ethnic or socio-political fault lines. It is also important to understand how the financial service provider will both store and use the data that it receives from humanitarian organizations to facilitate the digital payments. This is influenced both by the financial service provider's own company practice and by national legislation. For example, the ICRC has checklists that highlight data protection "red flags" and issues of concern which it would wish to discuss or negotiate with the provider. However, the humanitarian sector must be realistic; while it should be possible to negotiate that the service provider only uses the data provided to facilitate the digital payment and not for marketing or analyzing creditworthiness, obligations such as mandatory reporting under national legislation cannot be negotiated. Significant investment is being made by humanitarian organizations in tools and guidance to support analyzing data protection concerns with digital payments. A new short course hosted by CaLP focuses on e-transfers and operationalizing beneficiary data protection.[94] The ELAN Data Starter Kit[95] is designed to help humanitarians plan and improve data management practices, while the Mercy Corps *E-Transfer Implementation*

94 CaLP, "E-Transfers and Operationalizing Beneficiary Data Protection", available at: www.calpnetwork.org/blog/new-calp-online-training-course-e-transfers-and-operationalizing-beneficiary-data-protection/.

95 ELAN, *A Data Starter Kit for Humanitarian Field Staff*, available at: www.mercycorps.org/sites/default/files/2019-11/DataStarterKitforFieldStaffELAN.pdf

Guide[96] provides a step-by-step walkthrough on implementing e-transfers and includes support on analyzing the regulatory environment, KYC requirements and including data protection in its contracting. Many organizations, including UN agencies and NGOs, have their own internal data protection policies and guidance.

In short, every humanitarian organization must complete its due diligence. For the ICRC, this means firstly determining the likelihood of the risk – how likely is the data to be compromised? – and secondly determining the impact of the risk on the individual – how severe would the consequences be for the individual, as well as for the perception and acceptance of the ICRC by the population and the parties to the conflict? In practical terms, this entails ICRC teams completing a Data Protection Impact Assessment, which will help to identify the privacy risks to individuals, identify data protection compliance liabilities for the ICRC, protect the ICRC's reputation, and ensure that the organization does not compromise on the neutrality of its humanitarian action.[97] Similarly, CaLP has shared guidance on data protection in CVA, including model privacy impact assessments,[98] and this work is being built on by multiple humanitarian organizations.

It is critical that any risk assessment – whether for data protection or any other area of risk – includes consulting the affected population themselves. In Afghanistan, the ICRC considered using mobile money transfers to make digital payments. People living in Taliban-controlled communities were well aware of the possibility of geolocalization through mobile phones, and while they said that they trusted the ICRC to protect their data – and often gave the organization their phone numbers so that its teams could be in contact with them – they were concerned about how the mobile money companies would use their data, or which groups their data might be passed to. In direct contrast to this, in Ukraine, where ICRC teams discussed digital payments and technologies, the local communities – somewhat fatalistically – told the organization that the authorities already had all their data (through passport records, mobile phone and banking records, and even through Facebook), and as such, they didn't see any additional risk in the ICRC collecting and sharing their data. These two examples show how different people's perspectives of the likelihood and impact of risk can be, and why it is essential to take affected people's views into account when analyzing risk.

From a pure data protection perspective,

> one possible option in programmes using cash and voucher assistance is for the Humanitarian Organization to transfer, when feasible, a unique identifier (from which the receiving entity cannot identify the final beneficiary) and the amount

96 Mercy Corps, *E-Transfer Implementation Guide*, 2018, available at: www.mercycorps.org/sites/default/files/2020-01/EtransferGuide2018%2C%20Final.pdf.
97 C. Kuner and M. Marelli (eds), above note 71, Chap. 5.
98 CaLP, *Protection Beneficiary Privacy: Principles and Operational Standards for the Secure Use of Personal Data in Cash and e-Transfer Programmes*, 2013, available at: www.calpnetwork.org/wp-content/uploads/2020/01/calp-beneficiary-privacy-web.pdf.

of cash to be distributed to the commercial service provider (e.g. bank or mobile network operator), so as to limit the risks to the individuals concerned.[99]

This can be particularly useful for people who have no formal identity document, or where the risk of their data being shared with the financial service provider is deemed too high. Of course, this approach does not allow for financial inclusion, as the person does not have access to a personal account.[100]

Assessing critical infrastructure

In terms of financial infrastructure, it is much easier to assess its availability than its security. Humanitarian organizations can assess several issues, including the provider's liquidity, its geographical coverage, including the number of points of sale (locations and capacity of those places where recipients can make onward digital payments or "cash out"), and the availability of customer support services. With the growing risk of cyber attacks on financial infrastructure, it becomes more complex and difficult to analyze the security of that infrastructure. Beyond analyzing past incidents in the country and which providers they have affected – likely only from publicly available sources, as financial service providers do not disclose all hacks – there is a limit to what a humanitarian organization can do to be sure of the stability and safety of financial infrastructure. Humanitarians can, however, ensure that they implement appropriate technical and operational security standards in their own operations, and processes should be put in place for the protection of people's personal data from loss, theft, damage or destruction; this includes back-up systems and effective means to respond to security breaches and prevent unauthorized access, disclosure or loss of data that humanitarians store. The *Handbook on Data Protection in Humanitarian Action* also recommends that humanitarian organizations "protect 'by design' the Personal Data they obtain from beneficiaries …. Encryption or compartmentalization of information can be viable solutions to meet this need."[101]

Data responsibility: From theory to practice

All of these mitigation measures fall under the broad category of data responsibility. The OCHA Centre for Humanitarian Data has attempted to address gaps in guidance through its series on data responsibility, providing guidance on a wide range of issues including data incident management.[102] However, this remains a critical area for investment.

99 C. Kuner and M. Marelli (eds), above note 71, p. 154.
100 ICRC, above note 4, p. 26.
101 C. Kuner and M. Marelli (eds), above note 71, p. 155.
102 OCHA, Centre for Humanitarian Data, "Guidance Note: Data Incident Management", August 2019, available at: https://centre.humdata.org/wp-content/uploads/2019/08/guidanceNote2_dataincidentmanagement.pdf.

Of course, humanitarians can't predict the future; we cannot know all the myriad ways in which the data we collect today could be used for different purposes tomorrow, employing technologies that may not even exist today. However, all humanitarian agencies need to ensure data minimization – i.e., that only the personal data necessary for the identification of individuals should be collected and processed for the explicit humanitarian purpose of making the digital payments. Any "excess" information should not be collected, or if collected, should be deleted. We must ensure that the technologies we use to create digital ID, to store, process and share data and to communicate with affected people are secure. While it is vital to uphold clear standards on data protection, the humanitarian sector must also accept that it is impossible to fully protect data from every possible breach or misuse. More fundamentally, we should start from the fact that access to humanitarian assistance should not depend upon people being forced to disclose their personal information.[103] With robust data responsibility policies and practices, and thoughtful decision-making, we can do our due diligence, taking measures to analyze the risks that the method of assistance – and in the case of CVA, the payment instrument – exposes people to, and mitigate them.

Conclusion

The digitalization of cash is happening, with or without the humanitarian sector. Digital payments in humanitarian responses have been proven to have numerous benefits, primarily for recipients but also for humanitarian agencies delivering services, and these benefits are not being questioned *per se*. Rather, we need to ensure that these benefits outweigh the risks, and that as humanitarians we do our due diligence, on behalf of the people we serve, while recognizing that we do not live in a risk-free world. This is not a one-time deal; the analysis of risks and benefits should be done every time we choose whether to use digital payments, and should be done in collaboration with affected people, in order to take on board their views and their own analysis of risk.

Despite the rise in digital payments, it does not follow that cash will cease to exist. Cash (physical currency) is not linked to the person, and as such does not discriminate or identify; cash remains here to stay, at least in the coming years. The humanitarian sector must continue to work with both cash payments and digital payments – as well as offering in-kind assistance and services – based on what is most appropriate to meet the needs of populations affected by armed conflict and other situations of violence. We must always offer alternatives.

103 It is important when providing humanitarian services to collect the minimum amount of data possible, or even none at all. This process of data minimization is separate from the question of identifying the appropriate legal basis for data collection and processing, a principle which is supported by the ICRC's *Handbook on Data Protection in Humanitarian Action* and the "do no harm" principle and is one of the core principles of data protection.

Humanitarians are often accused of embracing change too slowly, particularly with regard to new technologies. Technology is moving faster than any of us – as private individuals or within our organizations – can keep pace with. Humanitarians are not financial service providers, and we are not technologists; we need to find avenues for working with the private sector to harness opportunities, but in ways that are in the best interests of crisis-affected people. At the ICRC's 2018 symposium on "Digital Risks in Situations of Armed Conflict", Professor Nathaniel Raymond, whose research interests have focused on the human rights and human security implications of information communication technologies for vulnerable populations, particularly in the context of armed conflict, challenged the audience by stating: "We are undermining the 'values of Geneva' through a relatively blind embrace of the potential 'promises of Silicon Valley'."[104] Banknotes and coins are a public utility, and companies make no profit from their use, only from the goods and services purchased with them. Hence the drive for "cashlessness", as digital payments generate revenue for financial service providers – and with it, greater surveillance and greater risks of misuse of data. Seen in this context, the approach taken by organizations like the ICRC and Oxfam, of taking time to analyze the opportunities and risks of solutions, should be understood as a measured and appropriate level of caution rather than a reluctance to embrace change. Humanitarians should "experiment in labs, and not on people".[105]

The ICRC will continue to be defined and led by our principled approach to humanitarian action, putting people at the centre of our work. To harness the opportunities that digitalization brings – for digital payments, but also beyond – we will continue to keep the people we serve at the centre of our analysis.

I would finish with a call to action for the humanitarian sector to do two key things.

Firstly, humanitarians must define what "do no digital harm" looks like in reality. "This requires that we understand the risks, protection issues, ethical concerns and challenges before building digital solutions and having a valid and important humanitarian purpose for developing a certain digital capability or using certain data."[106] Digitalization is changing the way our world works in fundamental ways, and this "digital disruption" has both positive and negative impacts on humanitarian action. Organizations must consider not only the way they provide humanitarian protection and assistance in a digital world, but how they transform themselves as digital agencies. We must translate our principled approach to humanitarian action into this brave new digital world.

Secondly, and most importantly, as we have reflected throughout this paper, we must keep people affected by conflict and other situations of violence at the centre of our action. Humanitarians must base both the strategic and day-to-day decisions that we make on conversations with the affected people that we

104 ICRC, above note 20, p. 5.
105 *Ibid.*, p. 2.
106 *Ibid.*, p. 2.

are here to serve, always keeping their interests at the forefront. This does not simply involve providing feedback mechanisms; it requires us to actively listen to, and reflect upon, what people affected by conflict and other situations of violence have to say, and to adapt our approaches to their preferences, needs and capacities. We must ensure a person-centred approach to the analysis of the risks and benefits of any action – including the use of digital payments.

Fundamentally, humanitarians must be more mindful in their decision-making around digital payments and their interactions in the digital world, always putting affected people at the centre of their decision-making.

International Review of the Red Cross (2020), **102** (913), 75–94.
Digital technologies and war
doi:10.1017/S1816383121000084

IRRC

Humanitarian aid in the age of COVID-19: A review of big data crisis analytics and the General Data Protection Regulation

Theodora Gazi and Alexandros Gazis
Theodora Gazi is a lawyer and a PhD candidate in refugee law at the School of Law, University of Athens. She is a Data Protection Specialist for the Danish Refugee Council (DRC Greece) and has been working in humanitarian aid since 2017.

Alexandros Gazis is a PhD candidate in computer science at the School of Engineering, Democritus University of Thrace, where he works as a Teaching Assistant and a Lab Demonstrator. He is also a Software Engineer for Eurobank SA, specializing in core banking systems.

Abstract
The COVID-19 pandemic has served as a wake-up call for humanitarian aid actors to reconsider data collection methods, as old ways of doing business become increasingly obsolete. Although access to information on the affected population is critical now more than ever to support the pandemic response, the limitation of aid workers' presence in the field imposes hard constraints on relief projects. In this article, we consider how aid actors can use "big data" as a crisis response tool to support humanitarian projects, in cases when the General Data Protection Regulation is applicable. We also provide a framework for examining open-source platforms, and discuss the advantages and privacy challenges of big data.

Keywords: big data, humanitarian aid, COVID-19, GDPR, data collection, crisis response.

Introduction

"Big data" has emerged as one of the most used buzzwords in the digital world, promising unique insight into understanding the aftermath of a disaster and the areas of need. In a displacement context, both the lack of information and the overflow of data may be paralyzing. For traditional humanitarian organizations, the use of big data is still uncharted territory. The question that arises is how aid actors can make use of large amounts of data, the majority of which is unstructured. On the one hand, data analytics introduce new opportunities for aid actors to support affected populations. On the other hand, big data could have serious implications for vulnerable individuals and communities if applied without safeguards. Importantly, practitioners should ensure compliance with data protection rules and best practices prior to resorting to innovative data collection methods, as this goes hand in hand with the humanitarian principles of non-discrimination and "do no harm" in the digital environment.

The goal of this study is to address the relationship between data protection and big data. As a result, we do not delve deeper into the intrinsic complexities of either of these issues. To explore the application of big data in humanitarian aid projects, we organize this article into two sections. First, we discuss the different views on what constitutes big data and on its potential use by aid actors to tackle the issues presented by COVID-19, focusing our analysis on two open-source software case studies. Then, we lay out key data protection rules in the EU and present the particularities of applying the General Data Protection Regulation (GDPR) to the processing of data from vulnerable populations. While the GDPR is applicable only to a portion of aid actors, we believe that its careful consideration is important. Indeed, it constitutes a "last-generation" data protection law that is shaping global regulatory trends on how to protect personal data in an increasingly digital world, along with other global benchmarks such as the ICRC's *Handbook on Data Protection in Humanitarian Action*.[1] Our purpose is to summarize the literature on big data, offer insight into its contribution to humanitarian projects and highlight its potential use by aid actors during the pandemic.

Defining big data and its use during the COVID-19 pandemic

"Big data" is an umbrella term that originated in the mid-1990s and became popular from 2011 onwards.[2] Its definition varies depending on the sector, and will likely

1 Christopher Kuner and Massimo Marelli (eds), *Handbook on Data Protection in Humanitarian Action*, 2nd ed., ICRC, Geneva, May 2020, p. 93.

2 Amir Gandomi and Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics", *International Journal of Information Management*, Vol. 35, No. 2, 2015, p. 138, available at: https://doi.org/10.1016/j.ijinfomgt.2014.10.007 (all internet references were accessed in January 2021).

evolve further, since what is defined as big data today may not be classified as such in a few years.[3] According to the independent European working party on the protection of privacy and personal data,[4] big data refers to "the gigantic amounts of digital data controlled by companies, authorities and other large organisations which are subjected to extensive analysis based on the use of algorithms. Big Data may be used to identify general trends and correlations".

In the data science industry, big data is defined by the "three Vs":[5] volume (large amounts of data), variety (data derived from different forms, including databases, images, documents and records) and velocity (the content of the data is constantly changing through complementary data from multiple sources). This list can be further enriched[6] to accommodate the intrinsic characteristics of aid projects by including veracity (credibility of the data for informed decision-making), values (respect of privacy and ethical use of crisis data), validity (mitigating biases and pitfalls), volunteers (motivation and coordination of volunteers) and visualization (presentation of big data in a coherent manner to support informed decisions). Throughout our work, we have adopted this enriched definition for aid projects in order to demonstrate the main data processing principles.

Moreover, big data refers to combining and analyzing information from diverse sources.[7] Depending on its source, data can be both structured (i.e., organized in fixed fields, such as spreadsheets and data sets) and unstructured (e.g., photos or words in documents and reports). In a crisis context, we identify the following sources for big data analysis:[8]

1. Data exhaust: information provided by individuals as by-products during the provision of humanitarian assistance, e.g. operational information, metadata records and web cookies. This refers to data which were not actively collected

3 See National Institute of Science and Technology, *NIST Big Data Interoperability Framework*, Vol. 1: *Definition*, US Department of Commerce, 6 September 2015, pp. 4–5, available at: http://dx.doi.org/10.6028/NIST.SP.1500-1; Council of Europe, *Guidelines on the Protection of Individuals with Regard to the Processing of Personal Data in a World of Big Data*, Strasbourg, 23 January 2017, p. 2.

4 Article 29 Data Protection Working Party, *Opinion 03/2013 on Purpose Limitation*, 2 April 2013, p. 45, available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

5 Jules J. Berman, *Principles and Practice of Big Data*, 2nd ed., Elsevier, London, 2018, p. 2.

6 Junaid Qadir, Anwaar Ali, Raihan ur Rasool, Andrej Zwitter, Arjuna Sathiaseelan and Jon Crowcroft, "Crisis Analytics: Big Data-Driven Crisis Response", *Journal of International Humanitarian Action*, Vol. 1, Article No. 12, 2016, p. 14, available at: https://doi.org/10.1186/s41018-016-0013-9.

7 Indicatively, see Alexandros Gazis and Eleftheria Katsiri, "Web Frameworks Metrics and Benchmarks for Data Handling and Visualization", in Yann Disser and Vassilios Verykios (eds), *Algorithmic Aspects of Cloud Computing*, Lecture Notes in Computer Science, Vol. 11409, Springer, Cham, 2018, available at: https://doi.org/10.1007/978-3-030-19759-9_9; Alexandros Gazis and Eleftheria Katsiri, "A Wireless Sensor Network for Underground Passages: Remote Sensing and Wildlife Monitoring", *Engineering Reports*, Vol. 6, No. 2, 2020, available at: https://doi.org/10.1002/eng2.12170.

8 UN Global Pulse, *Big Data for Development: Challenges and Opportunities*, May 2012, p. 16, available at: www.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay2012.pdf; Bapu Vaitla, *The Landscape of Big Data for Development: Key Actors and Major Research Themes*, Data2x, May 2014, available at: https://data2x.org/wp-content/uploads/2019/09/LandscapeOfBigDataForDevelopment_10_28-1.pdf.

but rather "left behind" from other digital interactions. These data are used as a sensor of human behaviour.

2. Crowdsourcing: information actively produced or submitted by individuals for big data analysis, via online surveys, SMS, hotlines etc. This method has been described as "the act of taking a job traditionally performed by a designated agent and outsourcing it to an undefined, generally large group of people in the form of an open call".[9] This information is valuable for verification and feedback.
3. Open data: publicly available data sets, web content from blogs and news media etc. Web content is used as a sensor of human intent and perceptions.
4. Sensing technology: satellite imagery of landscapes, mobile traffic and urban development. This information monitors changes in human activity.

The use of big data analysis peaked during the COVID-19 pandemic, which progressed from a worldwide public health emergency to a social and economic crisis. Scholars have claimed that at the time of writing (late 2020), all countries are using big data analytics to visualize COVID indicators in real time (such as case data, epidemic distribution and situation trends), inform the public about the epidemic situation and support scientific decision-making.[10]

Big data is especially relevant for aid actors in the context of disaster management, e.g. during migration crises, epidemics, natural disasters or armed conflicts.[11] During the COVID-19 pandemic, aid agencies switched to remote methodologies for data collection, such as phone surveys, remote key informant interviews and secondary data analysis.[12] Remote data collection relies heavily on the use of telecommunications and digital tools, such as phone calls, online surveys, SMS and messaging apps (such as WhatsApp and Signal). Big data analysis can also support aid actors in epidemic surveillance and response. However, the application of big data analysis to medical data is not widespread, due to the sensitive nature of medical records and the lack of a common technical infrastructure that can facilitate such analysis. The use of big data for

9 Jeff Howe, *Crowdsourcing: How the Power of the Crowd Is Driving the Future of Business*, Random House, New York, 2008.

10 See Qiong Jia, Yue Guo, Guanlin Wang and Stuart J. Barnes, "Big Data Analytics in the Fight against Major Public Health Incidents (Including COVID-19): A Conceptual Framework", *International Journal of Environmental Research and Public Health*, Vol.17, No. 17, 2020, available at: https://doi.org/10.3390/ijerph17176161. Their argument is strengthened by the data published through dashboards created by Johns Hopkins University (available at: https://coronavirus.jhu.edu/map.html) and the World Health Organisation (available at: https://covid19.who.int/) regarding active and past positive cases.

11 J. Qadir *et al.*, above note 6.

12 See International Organization for Migration, "Adapting to Change: IOM Faces COVID-19 Pandemic by Strengthening Outreach Tools", 6 February 2020, available at: www.iom.int/news/adapting-change-iom-faces-covid-19-pandemic-strengthening-outreach-tools; Save the Children, "Tipsheet: Remote and Digital Collection & COVID-19", 29 March 2020, available at: www.ready-initiative.org/wp-content/uploads/2020/06/COVID-19-and-MEAL-Remote-Data-Collection_v1.0-Save-the-Children1.pdf; Office of the UN High Commissioner for Refugees (UNHCR), Global Data Service, Innovation Service and Global Tri-Cluster Group, "Data Collection in Times of Physical Distancing", 13 August 2020, available at: www.unhcr.org/blogs/data-collection-in-times-of-physical-distancing/.

epidemic surveillance mainly involves the processing of crowdsourced data from volunteers who report protection needs.

Big data platforms include both commercial and free open-source products – i.e., software whose source code is open and publicly available for organizations to access, adjust or further enhance for any purpose.[13] Crisis management tools[14] may either be built from scratch or be revamped to fulfil existing needs. To better understand the advantages and disadvantages of big data analysis, we draw on a number of previous projects. First, we will review two recent projects led by a government agency and the private sector, which each developed an algorithm to predict migration trends. Then, we will focus on the Ushahidi and Sahana projects, which we believe are the most suitable open-source platforms to support humanitarian operations, and discuss their use for COVID-19 monitoring, depending on the size of the operation and the intended use.

Prediction of migration trends

Predictions of migration flows enable actors to better plan their resources in order to respond in a timely manner to humanitarian needs. The Swedish Migration Agency, the government body responsible for evaluating applications for asylum and citizenship in Sweden, has initiated a relevant big data project.[15] The Agency is using big data analysis to predict migration trends via annual comparisons of stored data. In this way, it gains insight into the expected needs and can plan for up to six months ahead in order to deploy resources to alleviate bottlenecks.[16] For instance, in October 2015, the Agency accurately predicted the number of refugees expected to arrive in Sweden by the end of the year.[17] However, while it predicted a high influx for 2016,[18] the number of submitted asylum applications declined significantly in that year.[19] The lower number of asylum-seekers in

13 Karim Lakhani and Eric Hippel, "How Open Source Software Works: 'Free' User-to-User Assistance", in Cornelius Herstatt and Jan G. Sander (eds), *Produktentwicklung mit virtuellen Communities*, Springer Gabler, 2004, available at: https://doi.org/10.1007/978-3-322-84540-5_13.

14 These include open/interactive mapping platforms (e.g. OpenStreetMap, Ushahidi, Sahana); health-care modules (e.g. OpenMRS); applications for volunteer management (e.g. Collabbit), budget and finance monitoring (e.g. Mifos), relief response (e.g. Relief Response Database), communication (e.g. FrontLineSMS) and food supply (e.g. LibreFoodPantry); and geographical information system tools (e.g. QGIS).

15 "Migrant Crisis: Sweden Doubles Asylum Seeker Forecast", *BBC News*, 22 October 2015, available at: www.bbc.com/news/world-europe-34603796.

16 Organisation for Economic Cooperation and Development and European Asylum Support Office, *Can We Anticipate Future Migration Flows?*, Migration Policy Debates No. 16, May 2018, p. 6, available at: www.oecd.org/els/mig/migration-policy-debate-16.pdf.

17 More specifically, the Swedish Migration Agency announced that 140,000–190,000 refugees were expected to arrive in Sweden by the end of the year, including 29,000–40,000 unaccompanied children. Indeed, 163,000 applications for international protection were registered in 2015, out of which 35,400 were from unaccompanied children. See Swedish Migration Agency and European Migration Network Sweden, *EMN Annual Report on Migration and Asylum 2016: Sweden*, 2017, available at: https://ec.europa.eu/home-affairs/sites/homeaffairs/files/27a_sweden_apr2016_part2_final_en.pdf.

18 An influx of 100,000–170,000 people was predicted for 2016, including up to 33,000 unaccompanied children.

19 In 2016, 28,939 asylum-seekers were registered, out of which 2,199 were unaccompanied children.

Sweden during 2016 was linked to the EU–Turkey Statement[20] signed in March 2016 to stop crossings to the Greek islands[21] and the border closure of the "Balkan route" to Europe. This has led scholars to argue that long-term decision-making based on migration forecasts is prone to error from unforeseen future events, while short-term predictions are far more useful.[22]

Another example of using big data for predictions of migration is the Danish Refugee Council's (DRC) partnership with IBM to develop a foresight model[23] (called Mixed Migration Foresight) in 2018. The project[24] focused on migration patterns from Ethiopia to six other countries. Anonymous data of thousands of migrants interviewed by the DRC revealed the main reasons for migration: lack of rights and/or access to social services, economic necessity, or conflict. Subsequently, these factors were mapped as quantitative indicators. Then, statistics about Ethiopia were processed, including its labour economy, education system, demographics and governance. Using these indicators, forecasts were produced for mixed migration flows to other countries. On average, the model was 75% accurate for 2018 figures.[25]

According to the forecasting software, the COVID-19 pandemic would lead to the displacement of more than 1 million people during 2020 across the Sahel region of Africa.[26] This prediction indeed captured the high increase of displacement that occurred in the area. However, already in November 2020, 1.5 million people had been displaced in the Central Sahel region, due to "unprecedented levels of armed violence and rights violations".[27] Moreover, based on the DRC's analysis, 6 million people residing in Mali, Niger and Burkina Faso were pushed into extreme poverty due to the pandemic. Again, this example shows that, while predictions based on big data may not fully factor the

20 See the full statement in European Council, "EU–Turkey Statement", 18 March 2016, available at: www.consilium.europa.eu/en/press/press-releases/2016/03/18/eu-turkey-statement/.

21 According to the European Commission, the EU–Turkey Statement was a "game changer". Irregular arrivals of migrants to the EU dropped by 97% from 2016 onwards compared to 2015. See European Commission, "EU–Turkey Statement: One Year On", 17 March 2017, available at: https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-migration/background-information/eu_turkey_statement_17032017_en.pdf.

22 George Disney, Arkadiusz Wiśniowski, Jonathan J. Forster, Peter W. F. Smith and Jakub Bijak, *Evaluation of Existing Migration Forecasting Methods and Models*, Economic and Social Research Council, Centre for Population Change, Southampton, 10 October 2015, available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/467405/Migration_Forecasting_report.pdf.

23 Rahul Nair, Bo Madsen, Helena Lassen, Serge Baduk, Srividya Nagarajan, Lars Mogensen, Rana Novack, Rebecca Curzon, Jurij Paraszczak and Sanne Urbak, "A Machine Learning Approach to Scenario Analysis and Forecasting of Mixed Migration", *IBM Journal of Research and Development*, Vol. 64, No. 1/2, 23 October 2019, available at: https://doi.org/10.1147/JRD.2019.2948824.

24 Rahul Nair, "Machine Learning in Action for the Humanitarian Sector", *IBM Research Blog*, 21 January 2019, available at: www.ibm.com/blogs/research/2019/01/machine-learning-humanitarian-sector/.

25 Karen Faarbæk de Andrade Lima, "Ethiopia Prototype", *DRC INSITE*, 21 January 2020.

26 Kate Hodal, "Covid to Displace More than a Million across the Sahel, New Tool Predicts", *The Guardian*, 11 August 2020, available at: www.theguardian.com/global-development/2020/aug/11/covid-to-displace-more-than-a-million-across-the-sahel-new-tool-predicts.

27 DRC, "Central Sahel is Rapidly Becoming One of the World's Worst Humanitarian Crises", 11 November 2020, available at: https://drc.ngo/about-us/for-the-media/press-releases/2020/11/central-sahel-crisis/.

highly politicized migration context, they can recognize migration trends and imminent humanitarian crises.

Consequently, both examples showcase that big data analysis is indeed useful as a prediction tool for recognizing migration patterns and informing decisions on expected needs. However, this analysis becomes "old news" quite soon, since external factors, such as climate change,[28] political decisions and the pandemic, can severely impact migration flows. Recognizing these limitations, however, big data can still serve as an indicator for preparedness, advocacy and programme planning.

The Ushahidi project

The Humanitarian Free and Open-Source Software community developed Ushahidi (meaning "testimony" in Swahili) in 2008 using PHP programming language.[29] Ushahidi is considered a "micro-framework" application, meaning that it adopts a minimalist approach, providing organizations with basic functions to fulfil three specific tasks: data collection, data management and visualization. Its main outputs are the visualization of data on a map, after applying data mining techniques.[30] Despite its fairly simple design, Ushahidi is included in the big data ecosystem for crisis management[31] because it is capable of analyzing both small and large data sets from diverse sources (per the defining "three Vs" of big data: volume, variety and velocity).

Initially, the application analyzed crowdsourced data solely via SMS messages[32] that reported incidents. Text messages were chosen as the most reliable method for data collection, given the limited network coverage at the time in the affected areas.[33]

For instance, during the 2010 Haiti earthquake, Ushahidi was used as a crowdsourcing platform to produce a crisis map based on information shared by volunteers who generated around 50,000 incident reports.[34] At the time, the US Federal Emergency Management Authority proclaimed the Ushahidi map as the "the most comprehensive and up-to-date source of information on Haiti for the

28 See International Organization for Migration, *Climate Change and Migration: Improving Methodologies to Estimate Flows*, IOM Research Series, No. 33, 2008, https://publications.iom.int/system/files/pdf/mrs-33.pdf.

29 Okolloh Ory, "Ushahidi, or 'Testimony': Web 2.0 Tools for Crowdsourcing Crisis Information", *Participatory Learning and Action*, Vol. 59, No. 1, 2009, available at: www.researchgate.net/publication/233563796_Ushahidi_or_%27testimony%27_Web_20_tools_for_crowdsourcing_crisis_information.

30 Stephen Kovats, *The Future of Open Systems Solutions, Now*, UNESCO World Summit on the Information Society, Berlin, 6 May 2013, available at: www.academia.edu/8746057/The_Future_of_Open_Systems_Solutions_Now.

31 J. Qadir *et al.*, above note 6.

32 The use of SMS for crowdsourcing had already been successfully adopted in previous NGO projects by UNICEF: Rapid SMS in 2010 (code available at: https://github.com/rapidsms/rapidsms) and U-Report in 2011 (code available at: https://github.com/unicefuganda/ureport).

33 See the code for using Ushahidi for crowdsourcing by SMS at the project's SMSSync GitHub public repository, available at: https://github.com/ushahidi/SMSSync).

34 Femke Mulder, Julie Ferguson, Peter Groenewegen, Kees Boersma and Jeroen Wolbers, "Questioning Big Data: Crowdsourcing Crisis Data Towards an Inclusive Humanitarian Response", *Big Data & Society*, 10 August 2016, available at: https://doi.org/10.1177/2053951716662054.

humanitarian community".[35] A four-digit telephone number was published and Haitians were encouraged to share urgent needs via text messages or emails, to be made public after they were translated. Three distinct groups contributed to this process: the digital humanitarians who ran the platform, Haitians affected by the earthquake, and global volunteer translators. "Implied" consent was used as a legal basis to make the incident reports public, based on the broad information-sharing about the project's purpose via radio and TV messaging.[36] However, we should note that this approach is problematic according to global data protection standards, since tolerance of a practice should not equal its acceptance, and the vulnerability of data subjects should also be taken into consideration.[37] We will further explore consent as a legal basis in the section below on "Applying Data Protection in Big Data", in light of the GDPR, which introduced strict requirements for the validity of consent. We should clarify that if the GDPR had been in force during that time, it would have applied if the digital humanitarians who ran the platform were EU-based actors.

Moreover, Ushahidi has been used retroactively by researchers to ameliorate aid response. For instance, researchers analyzed the geographic mobile phone records of nearly 15 million individuals between June 2008 and June 2009 in order to measure human mobility in low-income settings in Kenya and understand the spread of malaria and infectious diseases.[38] The Kenyan phone company Safaricom provided de-identified information to researchers, who then modelled users' travel patterns.[39] Researchers estimated the probability of residents and visitors being infected in each area by cross-checking their journeys with the malaria prevalence map provided by the government. This case would raise privacy concerns if the mobile phone data were publicly available, due to the re-identification risks based on persons' unique activity patterns. For this reason, when de-identified personal data are used for analysis purposes, anonymization procedures typically alter the original data slightly (causing a loss of data utility) in order to protect individuals' identities.[40] As we will analyze in the section below on "Applying Data Protection in Big Data", however, true anonymization of personal data is not always possible.

Furthermore, to reduce its cost for participants, the Ushahidi application integrated additional features for data collection and text processing in the

35 Jessica Heinzelman and Carol Waters, *Crowdsourcing Crisis Information in Disaster-Affected Haiti*, United States Institute of Peace, Washington, DC, 29 September 2019.

36 "Crisis Mapping Haiti: Some Final Reflections", *Ushahidi Blog*, 14 April 2020, available at: www.ushahidi.com/blog/2010/04/14/crisis-mapping-haiti-some-final-reflections.

37 See C. Kuner and M. Marelli (eds), above note 1, pp. 61–63.

38 For the full study, see Amy Wesolowski, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow and Caroline O. Buckee, "Quantifying the Impact of Human Mobility on Malaria", *Science*, Vol. 338, No. 6104, 12 October 2012, available at: https://doi.org/10.1126/science.1223467.

39 Harvard School of Public Health, "Using Cell Phone Data to Curb the Spread of Malaria", press release, 11 October 2012, available at: www.hsph.harvard.edu/news/press-releases/cell-phone-data-malaria/.

40 Ling Yin, Qian Wang, Shih-Lung Shaw, Zhixiang Fang, Jinxing Hu, Ye Tao and Wei Wang, "Re-identification Risk versus Data Utility for Aggregated Mobility Research Using Mobile Phone Location Data", *PLOS One*, Vol. 10, No. 10, 2015, available at: https://doi.org/10.1371/journal.pone.0140589.

following years. Nowadays, more data streams may be processed, including emails, web forms and tweets based on hashtags. Since 2017, the application has also adopted artificial intelligence processing to automate data gathering via the use of chatbots.[41] More specifically, automation bots can interact with users via the Facebook Messenger application. Following a short "dialogue" between the user and the bot, immediate suggestions are offered based on algorithms or the request is catalogued for further processing.[42]

During the COVID-19 pandemic, the Ushahidi platform has also been used to map the availability of public services, volunteer initiatives and requests for help. For instance, the Italian organization ANPAS[43] visualized services offered by volunteers across Italy in order to respond to recurrent needs for food, medicine and necessary goods.[44] Similarly, the FrenaLaCurva project[45] allowed Spanish language-speakers to share needs and available resources in Spain and the Canary Islands.[46] The Redaktor project[47] focused on empowering institutions and journalists across the globe, in addition to promoting community-oriented dissemination of information by mapping their needs for support. These examples demonstrate that big data can be and has been used in various ways to support the provision of help and various services to those affected by COVID and related restrictions.

Ushahidi is fairly easy to set up and serves as a crowdsourcing platform which may be accessed from multiple devices in remote areas, even if network connectivity is low. Its main disadvantage is its dependence on unstructured data (i.e., words in different languages and metadata), which frequently results in missing or inaccurate information.[48] Additionally, aid actors should take into consideration that big data analysis may be inherently biased, since it can exclude marginalized and under-represented groups, such as children, illiterate persons,

41 "'Hi This Is the Ushahidi Facebook Messenger Chatbot' – Meeting People Where They Already Are", *Ushahidi Blog*, 25 August 2017, available at: www.ushahidi.com/blog/2017/08/25/hi-this-is-the-ushahidi-facebook-messenger-chatbot-meeting-people-where-they-already-are-1.

42 Joanna Misiura and Andrej Verity, *Chatbots in the Humanitarian Field: Concepts, Uses and Shortfalls*, Digital Humanitarian Network, May 2019, available at: www.academia.edu/40918719/Chatbots_in_the_humanitarian_field_concepts_uses_and_shortfalls.

43 See the interactive map developed by Ushahidi for ANPAS, available at: https://anpas.ushahidi.io.

44 June Mwangi, "ANPAS: Supporting Vulnerable Communities in Italy during Covid-19 Lockdowns", *Ushahidi Blog*, 20 March 2020, available at: www.ushahidi.com/blog/2020/03/20/anpas-supporting-vulnerable-communities-in-italy-during-covid-19-lockdowns.

45 See the interactive map developed by Ushahidi for FrenaLaCurva, available at: https://es.mapa.frenalacurva.net/.

46 Angela Oduor Lungati, "Frena La Curva: Connecting Spanish Speakers with Critical Resources Around Them", *Ushahidi Blog*, 25 March 2020, available at: www.ushahidi.com/blog/2020/03/25/frena-la-curva-connecting-spanish-speakers-with-critical-resources-around-them.

47 See the interactive map developed by Ushahidi for the Redaktor project, available at: https://redaktor.ushahidi.io/.

48 Unstructured data do not have a predefined structure, so big data analysis requires additional processing power (i.e., CPU and RAM usage), a higher execution time or the purchase of costly computer systems to mine the data. Indicatively, see A. Gazis and E. Katsiri, "Web Frameworks Metrics", above note 7; Kiran Adnan and Rehan Akbar, "An Analytical Study of Information Extraction from Unstructured and Multidimensional Big Data", *Journal of Big Data*, Vol. 6, Article No. 91, 17 October 2019, available at: https://doi.org/10.1186/s40537-019-0254-8.

the elderly, indigenous communities and people with disabilities.[49] Furthermore, it does not always provide aid actors with sufficient information on the incidents reported, e.g. location, description and number of affected individuals.[50]

Moreover, the applicable data protection law needs to be taken into consideration when aid organizations invite users to post public reports through the platform. For instance, while Ushahidi has updated its policies and practices to comply with the GDPR,[51] actors which are either EU-based or which target individuals residing in the EU (irrespective of the organization's place of establishment) still need to acquire users' consent as defined by GDPR's strict criteria and inform them accordingly about any processing activities. This is because compliance with data protection is not just about the use of appropriate software tools; it extends to all aspects of data life-cycle management and to respecting data subjects' rights.

Sahana project

In 2009, the Humanitarian Free and Open-Source Software community developed the Sahana project (meaning "relief" in Sinhalese). Sahana consists of two applications, Agasti[52] and Eden.[53] In contrast to Ushahidi, this project includes framework applications providing organizations with versatile options during big data analysis, instead of only core functions. We will focus our analysis on Eden, since its numerous modules serve multiple purposes during humanitarian projects, from support services to programmatic and field needs. Eden, which stands for Emergency Development Environment, is a more sophisticated application than Ushahidi, using the Python programming language. By processing structured data (mainly in CSV format[54]), it supports organizations in managing people, assets and inventory.

The modules integrated in Eden may be utilized[55] for both supporting (e.g. for inventory and human resources management) and programming purposes (e.g.

49 Shweta Bansal, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani and Cécile Viboud, "Big Data for Infectious Disease Surveillance and Modeling", *Journal of Infectious Diseases*, Vol. 214, No. 4, 2016, available at: https://doi.org/10.1093/infdis/jiw400.

50 Patrick Meier, *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*, Routledge, New York, 2015, available at: https://doi.org/10.1201/b18023.

51 Charlie Harding, "Ushahidi has Updated Its Privacy Policy and Is GDPR Compliant", *Ushahidi Blog*, 24 May 2018, available at: www.ushahidi.com/blog/2018/05/24/ushahidi-has-updated-its-privacy-policy-and-is-gdpr-compliant.

52 The Agasti application uses PHP programming language and has two sub-applications: Mayon, for emergency personnel and resource management, and Vesuvius, for disaster preparedness and response. See the code for both projects, available at: https://launchpad.net/sahana-agasti/+series.

53 Mifan Careem, Chamindra De Silva, Ravindra De Silva, Louiqa Raschid and Sanjiva Weerawarana, "Sahana: Overview of a Disaster Management System", in Institute of Electrical and Electronic Engineers, *Proceedings of the International Conference on Information and Automation*, 15–17 December 2016, available at: https://doi.org/10.1109/ICINFA.2006.374152.

54 Khanh Ngo Duc, Tuong-Thuy Vu and Yifang Ban, "Ushahidi and Sahana Eden Open-Source Platforms to Assist Disaster Relief: Geospatial Components and Capabilities", in Alias Abdul Rahman, Pawel Boguslawski, François Anton, Mohamad Nor Said and Kamaludin Mohd Omar (eds), *Geoinformation for Informed Decisions*, Lecture Notes in Geoinformation and Cartography, Vol. 102, Springer, Cham, 2014, available at: https://doi.org/10.1007/978-3-319-03644-1_12.

55 Sahana Software Foundation, "Sahana Eden: Open Source Disaster Management Software Platform", 13 December 2011, available at: www.slideshare.net/SahanaFOSS/sahana-eden-brochure-10577413.

registry of disaster survivors and messaging system for the reception of and automated response to emails, SMS and social networks). Moreover, Eden can visualize inputs in maps and produce automated scenario templates for managing crisis, based on predetermined resources and past experience (e.g. number of resources and employees needed, time frames). Additionally, Sahana modules are particularly relevant to COVID-19 response. They cover shelter and inventory management, which can be used to track the availability of hospital beds, quarantine centres, childcare facilities (e.g. for medical staff or patients) and medical supplies (e.g. surgical masks and COVID-19 tests). Sahana also allows incident reporting and mapping of requests for food and supplies.

Indeed, Sahana has been utilized to respond to the COVID-19 pandemic, improve data collection and coordinate volunteers globally. In the northwest of England, the county council of Cumbria[56] used Sahana to track vulnerable individuals, coordinate the distribution of protection equipment and supplies to families, and manage volunteers. Additionally, Pakistan's government used the relevant applications for supply chain and mapping to plan its logistic needs and perform case tracking.[57]

Sahana has integrated Ushahidi's functions, so it can process crowdsourced data and visualize them on a map. However, due to its ease of use, Ushahidi better fits the missions of smaller aid actors to coordinate rapid responses to disaster situations. Eden allows organizations to transfer data generated from Ushahidi[58] when they need to scale up their operation, but the opposite is not feasible automatically. In sum, Sahana is suitable for long-term projects and larger organizations, offering a vast range of options for designing, monitoring and executing disaster relief interventions. While both platforms have their benefits in the appropriate operational contexts, both come with privacy concerns depending on the data processed and the outputs produced. These concerns will be examined next.

Applying data protection in big data

The right to privacy

The information processed for big data analysis is not always personal data – i.e., data relating to an identified or identifiable natural person.[59] However, in the field of humanitarian assistance, personal data are typically processed to facilitate

56 Devin Balkind, "Sahana EDEN Used for COVID-19 Responses", *Sahana Foundation Blog*, 23 April 2020, available at: https://sahanafoundation.org/sahana-eden-used-for-covid-19-responses/.

57 Sahana Software Foundation, "Sahana Applicability for COVID-19", 20 March 2020, available at: https://tinyurl.com/rb2fpbgw.

58 Mark Prutsalis, "Developing a Service Industry to Support the Sahana Disaster Management System", *Open Source Business Resource Journal*, December 2010, available at: https://timreview.ca/article/400.

59 This definition of personal data is stated in the General Data Protection Regulation (Regulation on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC), (EU) 2016/679, 27 April 2016 (GDPR), Art. 4(1).

the identification of individuals in need and the recognition of patterns.[60] When aid actors perform data analysis, they usually promote a participatory model through the combination of open and crowdsourced data, especially when outputs are used to inform decision-making.[61] Despite this, the potential for abuse is high, since data analytics could lead to infringement of privacy and discrimination if proper safeguards are not adopted. While big data analysis promises insight, there is a risk of establishing a "dictatorship of data" in which communities are judged not by their actions, but by the data available about them.[62] Thus, issues of privacy must be tackled before applying big data analysis in crisis contexts.

The right to privacy is a fundamental human right recognized in international law by numerous international instruments, such as the United Nations (UN) Declaration of Human Rights, the International Covenant on Civil and Political Rights and the European Convention on Human Rights. Moreover, the UN Special Rapporteur on the Right to Privacy, whose mandate is to monitor, advise and publicly report on human rights violations and issues, plays an important role in highlighting privacy concerns and challenges arising from new technologies.[63] Additionally, an important binding legal instrument on data protection (a notion which originates from the right to privacy[64]) is the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) adopted by the Council of Europe.

In the EU context, the Charter of Fundamental Rights of the EU states that everyone has the right to respect for their private and family life (Article 7), in addition to the protection of personal data concerning themselves (Article 8). Moreover, the GDPR sets out common rules both for EU-based actors who process personal data of individuals located within or outside the EU and for actors who target their services to EU residents, irrespective of the actors' place of establishment. The focus of the remainder of this article is the GDPR, which came into force in May 2018. The reasons why we decided to analyze this legislation are threefold. Firstly, following the 2015 refugee migration crisis, multiple EU-based actors are currently implementing aid projects both in countries outside the EU and within member States, and these require the continuous processing of beneficiary communities' data. Secondly, the existing literature on the application of the GDPR to the humanitarian aid sector is limited. Thirdly, while the GDPR applies to a portion of aid actors, it is a "last-

60 European Data Protection Supervisor (EDPS), *Meeting the Challenges of Big Data: A Call for Transparency, User Control, Data Protection by Design and Accountability*, Opinion 7/2015, 19 November 2015, p. 7, available at: https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf.

61 F. Mulder *et al.*, above note 34.

62 Datatilsynet (Norwegian Data Protection Authority), *Big Data: Privacy Principles under Pressure*, September 2013, p. 7, available at: www.datatilsynet.no/globalassets/global/english/big-data-engelsk-web.pdf.

63 See Joseph A. Cannataci, *Recommendation on the Protection and Use of Health-Related Data*, UN Doc. A/74/277, December 2019; Joseph A. Cannataci, *Preliminary Evaluation of the Privacy Dimensions of the Coronavirus Disease (COVID-19) Pandemic*, UN Doc. A/75/147, 27 July 2020.

64 EDPS, "Data Protection", 2017, available at: https://edps.europa.eu/data-protection/data-protection_en.

generation" EU law which has incorporated international data protection principles and is highly likely to set the standard and affect global regulatory trends.

When is the GDPR applicable to big data analysis?

An important question is that of when big data analysis falls within the scope of the GDPR. The Regulation applies in principle to every kind of operation and activity performed by EU-based public authorities, companies and private organizations that process personal data, regardless of the location of the individuals whose data are processed (within or outside the EU). Additionally, the GDPR applies to non-EU actors when they target their services to individuals residing in the EU.[65]

It is important to note that the GDPR does not apply to anonymous information or to personal data that is rendered anonymous, in that the data subject is no longer identifiable.[66] This includes personal data collected for humanitarian aid, provided that they can be truly anonymized. However, this is not always possible, as one cannot exclude the possibility of re-identification of individuals from other data, even when anonymization techniques have been applied. This is because anonymization is not achieved by just deleting direct identifiers, since the accumulation of different pieces of data increases the probability of re-identification. This is especially true when the target population is small and/or the subjects have a combination of rare and intrinsic characteristics. The UN Special Rapporteur on the Right to Privacy has also highlighted this risk when combining closed and open data sets.[67] Because a person's identity could be revealed by combining anonymous data with publicly available information and other data sets, de-identified data may be considered personal even after anonymization techniques have been employed. Subsequently, when NGOs attempt to anonymize personal data, they should examine whether there is a risk of re-identification. In any case, anonymization is not a one-off activity; anonymization techniques and software are constantly updated with new modules and more complex algorithms to prevent re-identification and to preserve the anonymity of data sets.

As for pseudonymized data, they fall inside the scope of the GDPR[68] because they may still be attributed to an identifiable person by the use of additional information. In a big data context, pseudonymized data may be the preferred approach given that identifiability is sometimes necessary for validating the outputs.[69] Consequently, EU-based organizations remain subject to data

65 GDPR, above note 59, Art. 3.
66 *Ibid.*, Recital 26.
67 Joseph A. Cannataci, *Report of the Special Rapporteur on the Right to Privacy*, UN Doc. A/72/540, 19 October 2017, available at: https://undocs.org/A/72/540.
68 GDPR, above note 59, Recital 26.
69 For further reading, see ICRC, *The Humanitarian Metadata Problem: "Doing No Harm" in the Digital Era*, Geneva, 2018, available at: www.icrc.org/en/download/file/85089/the_humanitarian_metadata_problem_-_icrc_and_privacy_international.pdf.

protection rules when they analyze big data that has been pseudonymized or may be re-identified through reverse engineering.[70]

Applicable legal bases for big data analysis

According to the GDPR, a legal basis must be identified for any data processing activity. The majority of data handled by aid actors are sensitive, especially the information required for COVID-19 monitoring, which includes the processing of health data. Based on Article 9(2) of the GDPR, the applicable legal bases for aid organizations to process sensitive data are: (i) the data subject's explicit consent; (ii) protection of the data subject's vital interests and those of others who are incapable of providing consent; and (iii) public interest in the area of public health.

As mentioned, crowdsourced data – i.e., data retrieved from individuals based on their consent and on a voluntary basis – is a key data source for big data analysis. Based on Recital 32 of the GDPR, consent should be specific, freely given and informed. This means that individuals must have a clear understanding of what they are agreeing to. Consent may be expressed in writing, electronically or orally; however, silence does not imply consent.[71] The definition of "explicit" is not provided by the Regulation; in practice, it means that consent should be confirmed by a clear statement for a specific purpose, separately from other processing activities.[72] Moreover, for consent to be meaningful, data subjects need to have efficient control over their data.[73] Consent is valid until it is withdrawn and for as long as the processing activity remains the same.[74] Interestingly, we notice that while lawfulness of processing is a separate requirement to the rights of data subjects,[75] both of these GDPR requirements are interlinked for consent to be valid.

To be more specific regarding humanitarian assistance, valid consent is not just about ensuring that individuals "tick a box" to indicate their informed decision. Data subjects need to be informed about the use of their data, in a language and format they understand. Moreover, the request for consent must be direct and explicit, and an equivalent process must be available to withdraw consent. Indeed, valid consent presents many difficulties during a crisis context, due to language barriers and the complexity of data processing activities for the provision of

70 C. Kuner and M. Marelli (eds), above note 1, p. 93.

71 European Data Protection Board, *Guidelines 05/2020 on Consent under Regulation 2016/679*, 4 May 2020, available at: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf.

72 Information Commissioner's Office (ICO), *Consultation: GDPR Consent Guidance*, March 2017, available at: https://ico.org.uk/media/about-the-ico/consultations/2013551/draft-gdpr-consent-guidance-for-consultation-201703.pdf.

73 EDPS, above note 60, p. 11.

74 UNESCO, *Report of the International Bioethics Committee of UNESCO on Consent*, 2008, p. 17, available at: https://unesdoc.unesco.org/ark:/48223/pf0000178124.locale=en.

75 Data subjects' rights are analyzed further in Theodora Gazi, "Data to the Rescue: How Humanitarian Aid NGOs Should Collect Information Based on the GDPR", *International Journal of Humanitarian Action*, Vol. 5, Article No. 9, July 2020, available at: https://doi.org/10.1186/s41018-020-00078-0.

humanitarian aid. Since aid organizations target specific communities, information about the intended big data analysis must be provided in the local language, in an understandable manner, regardless of the reader's educational level.[76] Therefore, big data analysis based on crowdsourced data may rely on explicit consent as long as individuals are properly informed about the processing purpose in a user-friendly way, such as a pop-up window or a text message with the relevant information and consent request. Consequently, aid actors' mandates to assist conflict-affected populations do not give them carte blanche to perform data processing.[77]

The debate on whether beneficiaries' consent to the use of their data is valid is not new. Indeed, when data processing is necessary for the provision of life-saving services, consent is not the appropriate legal basis. Recital 46 states that the "vital interests" legal basis may apply when actors are processing data on humanitarian grounds, such as to monitor epidemics and their spread or in situations where there is a natural or man-made disaster causing a humanitarian emergency. Indeed, data protection must never hinder the provision of assistance to vulnerable people at risk, who would be excluded from data collection when incapable of providing consent.[78] The "vital interests" basis applies when personal data must be processed in order to protect an individual's life, while the person is incapable of consenting and "the processing cannot be manifestly based on another legal basis". In these cases, big data analysis which facilitates the rapid assessment of patients' needs and their access to life-saving aid services can be based on vital interests. However, the condition of vital interest is not met when big data analysis is undertaken in non-urgent situations. Thus, processing of personal data focused on research or donor compliance cannot rely on vital interest. When data processing could be performed in a less intrusive manner, the conditions for applying this legal basis are not met.[79]

Lastly, based on Recital 54 of the GDPR, processing of sensitive data may be necessary for reasons of public interest, without acquiring the data subjects' consent. Moreover, according to Article 9, sensitive data may be processed "for reasons of public interest in the area of public health, such as protecting against serious cross-border threats". Based on the above, the "public interest" legal basis can be invoked by aid actors, for instance when they collaborate with the public authorities to support medical aid. Indeed, data processing for reasons of public health is an outcome of the State's duty vis-à-vis its citizens to safeguard and promote their health and safety. Given that public interest is determined by the States themselves, this regulatory leeway allows for States to acquire sensitive

76 Article 29 Data Protection Working Party, *Guidelines on Transparency under Regulation 2016/679*, 11 April 2018, p. 11, available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.
77 Nicole Behnam and Kristy Crabtree, "Big Data, Little Ethics: Confidentiality and Consent", *Forced Migration Review*, No. 61, June 2019, p. 6, available at: www.fmreview.org/sites/fmr/files/FMRdownloads/en/ethics/ethics.pdf.
78 Lisa Cornish, "Is Data Consent in Humanitarian Contexts Too Much to Ask?", 6 August 2018, available at: www.devex.com/news/is-data-consent-in-humanitarian-contexts-too-much-to-ask-93133.
79 C. Kuner and M. Marelli (eds), above note 1.

personal data in the context of a global pandemic. However, this basis enables data processing even when the purpose is not compatible with the data subjects' best interests. The risk of "aiding surveillance" should be highlighted as a significant concern when applying this legal basis, since big data analysis by aid actors could potentially be weaponized to achieve national security objectives.[80] As a result, actors should use this legal basis with caution, taking proportionality into consideration when asked to collect or share data with public authorities. Data-sharing agreements to regulate this data exchange and strict application of the basic data protection principles analyzed in the next section (especially data minimization and purpose limitation) are crucial to avoiding excessive data collection.

Data protection principles and big data analysis

The basic principles of data protection, set out in Article 5 of the GDPR, constitute the backbone of the legal framework when engaging in data analytics. In this section we will analyze these principles in the context of humanitarian assistance and big data analysis.

First, the Regulation requires "lawfulness, fairness and transparency". This means that apart from identifying the relevant legal basis, actors must ensure that data processing is fair and transparent. When performing big data analysis, its purpose constitutes an important factor for assessing the "fairness" and "transparency" principles – i.e., that individuals are informed about the envisioned use of their data in simple, clear language.[81] Fairness is linked to whether data will be handled in an expected way while not causing unjustified adverse effects on data subjects, individually or as a group. Equally, vulnerabilities must be considered when assessing the data subject's likely level of understanding.[82] Lack of transparency could entail that individuals have no idea or control over how their data are used.

Moreover, the GDPR refers to the principles of data minimization, storage limitation and purpose limitation. These principles were already well established in the humanitarian aid sector, long before the GDPR.[83] They specify that aid actors should limit the collection and retention of personal data to the extent that is necessary to accomplish a specific purpose. It is true that data minimization and storage limitation could clash with the key prerequisite for big data – i.e., "volume". Indeed, stockpiling personal data "just in case" they become useful clearly breaches the GDPR. However, a "save everything" approach does not

80 Ben Hayes, "Migration and Data Protection: Doing No Harm in an Age of Mass Displacement, Mass Surveillance and Big Data", *International Review of the Red Cross*, Vol. 99, No. 904, 2018, p. 193, available at: https://doi.org/10.1017/S1816383117000637.

81 ICO, *Big Data, Artificial Intelligence, Machine Learning and Data Protection*, Version 2.2, 2017, p. 20.

82 Article 29 Data Protection Working Party, above note 76, p. 11.

83 See UN Office for the Coordination of Humanitarian Affairs (OCHA), *Building Data Responsibility into Humanitarian Action*, May 2016, available at: www.unocha.org/fr/publication/policy-briefs-studies/building-data-responsibility-humanitarian-action.

necessarily benefit big data analysis. Scholars have argued that storage of data for big data analysis is considered a thing of the past, during the present era of real-time data.[84] Additionally, appropriate data classification and clear policies on data processing improve data quality and the outputs of data science.[85] In any case, data protection "by design" solutions could involve anonymization, where possible, when personal data storage is not justifiable.

As for the purpose limitation principle, big data projects for COVID-19 have a specific aim – namely, to limit the spread of the virus and to protect public health. However, the reuse of personal data collected during humanitarian assistance may place this principle under pressure. Based on Article 5 of the GDPR, personal data "shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes". This principle allows data subjects to make an informed choice about entrusting their data to the aid actor, and to be certain that their data will not be processed for irrelevant purposes, without their consent or knowledge. In some cases, determining whether big data analysis is compatible with the initial purpose might not be straightforward. In any case, the purpose must be expressly stated and legitimate – i.e., there should be an objective link between the purpose of the processing and the data controller's activities.

When big data analysis is used as a tool for policy and decision-making, an important data protection principle that must be respected is that of accuracy. This principle applies when individuals are affected by the outcome of the analysis. Big data typically processes information from diverse sources, without always verifying their relevance or accuracy. This presents several challenges. Firstly, analysis of personal data initially processed in different contexts and for other purposes may not portray the actual situation. Similarly, while working with anonymized data is less intrusive, it increases the risk of inaccuracy. As such, open data may not constitute an appropriate factual basis for decision-making, since this information may not be verified to the same degree as personal data. Predictive analysis can result in discrimination, promotion of stereotypes and social exclusion;[86] this is why big data has been accused of presenting misleading and inaccurate results that fail to consider specific particularities of the community or individuals. In any case, predictive models are inherently biased, regardless of what data they draw on. Data quality can increase the accuracy of predictive models but is not a remedy for their methodological bias.

Therefore, open and anonymous data must be selected with diligence so as to guarantee that the data processed are of the right quality and produce credible outputs. In contexts where actors largely rely on open data, best practices include giving prominence to updated and relevant data sets and enhancing cooperation between aid actors in order to encourage regular information-sharing. Furthermore,

84 Quentin Hardy, "Jeff Hawkins Develops a Brainy Big Data Company", *New York Times*, 28 November 2012, available at: https://bits.blogs.nytimes.com/2012/11/28/jeff-hawkins-develops-a-brainy-big-data-company/.

85 ICO, above note 81, p. 42.

86 EDPS, above note 60, p. 8.

open data need to be validated via beneficiaries' inputs.[87] The combination of open data and big data analysis, through crowdsourcing, enables actors to cross-reference and triangulate the data of specific groups, understand their needs and increase the effectiveness of the operation.

Another key data protection principle is that of confidentiality, meaning that data must be sufficiently protected from unauthorized disclosure.[88] Security measures for big data are linked to the outputs of the analysis, especially when it produces more sensitive data than those included in the initial data sets. Data security is also achieved by applying the data minimization and storage limitation principles, since decreasing the collection of data reduces the risk of data breaches. Additionally, aid actors should use safe data analysis tools and train employees on their proper use. While aid actors' data sets usually uphold security standards, users have been identified as "the weakest link" for data breaches, e.g. due to loss of IT equipment and phishing scams. Encryption and pseudonymization of data sets – i.e., replacing personal identities with codes – are also promoted by the GDPR as a security measure[89] that can prevent the misuse of data.

Finally, a separate requirement introduced in Article 25 involves applying the above principles "by design and by default", along with other legal obligations. Since any data processing activity includes inherent protection risks, aid actors must always assess these risks and adopt appropriate safeguards. In any case, prior to launching big data analysis, aid actors are advised to conduct a data protection impact assessment (DPIA), as described in Article 35 of the GDPR. DPIAs[90] are a key requirement prior to activities that include the processing of sensitive data on a large scale.[91]

87 European Data Portal, *Open Data Best Practices in Europe: Learning from Cyprus, France, and Ireland*, May 2020, available at: www.europeandataportal.eu/sites/default/files/report/20200518_AR16_ODM%20Top%20Performing%20Countries_V1.1_FINAL.pdf.

88 For an in-depth analysis of the principle of confidentiality, see Kurt Schmidlin, Kerri Clough-Gorr and Adrian Spoerri, "Privacy Preserving Probabilistic Record Linkage (P3RL): A Novel Method for Linking Existing Health-Related Data and Maintaining Participant Confidentiality", *BMC Medical Research Methodology*, Vol. 15, Article No. 46, 30 May 2015, available at: https://doi.org/10.1186/s12874-015-0038-6; Loredana Caruccio, Domenico Desiato, Giuseppe Polese and Genoveffa Tortora, "GDPR Compliant Information Confidentiality Preservation in Big Data Processing", *IEEE Access*, Vol. 8, 9 November 2020, available at: https://doi.org/10.1109/ACCESS.2020.3036916.

89 GDPR, above note 59, Art. 32(1)(a).

90 Useful DPIA templates have been developed by the ICRC (see C. Kuner and M. Marelli (eds), above note 1, pp. 300–302) and the French supervisory data protection authority (Commission Nationale Informatique & Libertés, *Privacy Impact Assessment Template*, February 2018, available at: www.cnil.fr/sites/default/files/atoms/files/cnil-pia-2-en-templates.pdf).

91 A DPIA determines the relevant data sources (e.g. open data, pre-existing data sets and crowdsourcing) and the safeguards implemented to achieve fair results and compliance with the GDPR. It also contains a description of the processing activities involved, a risk analysis of the rights of data subjects, and an examination of whether anonymization is applicable; the latter is achieved by running tests to assess the probability of re-identification, especially when the group of data subjects is not sufficiently large. Dariusz Kloza, Niels Van Dijk, Simone Casiraghi, Sergi Vazquez Maymir, Sara Roda, Alessia Tanas and Ioulia Konstantinou, *Towards a Method for Data Protection Impact Assessment: Making Sense of GDPR Requirements*, Vrije Universiteit Brussel, Policy Brief No. 1, 2019, available at: https://cris.vub.be/files/48091346/dpialab_pb2019_1_final.pdf.

Conclusions

The COVID-19 pandemic intensified existing inequities, increasing the financial insecurity of vulnerable people. This meant that the number of households in need of humanitarian support multiplied, while direct access to them became harder. Specifically, the health risks and government measures caused by COVID-19 have severely limited traditional methods for primary data collection, such as conducting household visits, field assessments and focus group discussions.[92] Aid actors can address these major challenges by applying big data analysis to continue their operations and monitor their response to the pandemic. Big data has been defined as a technological phenomenon which relies on the interplay of technology (the use of computation power and algorithmic accuracy to link and compare large volumes of data) and analysis (the recognition of patterns in order to predict behaviours, inform decisions and produce economic or social indicators).[93]

Indeed, the continuation of humanitarian assistance and the monitoring of epidemic responses can be facilitated by technological innovations. As with any technological tool, big data may support disaster management responses, provided that its use does not derail humanitarian efforts or harm beneficiaries' rights. The UN Office for the Coordination of Humanitarian Affairs (OCHA)[94] has underlined that using big data for humanitarian purposes is one of the greatest challenges and opportunities of the network age. Big data is addressed in this context with the possibility of predicting, mapping and monitoring COVID-19 responses.[95]

The belief that big data is a "panacea for all issues" is the main cause of concern expressed by scholars.[96] Big data analysis entails privacy risks and may produce biased results, leading aid actors to misguided decisions and inequity in the provision of humanitarian assistance.[97] Aid actors should be mindful of the shortcomings of both big data and open data. First, both categories often lack demographic information that is crucial for epidemiological research, such as age and sex. Second, this data represents only a limited portion of the population – i.e.,

92 UNHCR, above note 12.

93 Danah Boyd and Kate Crawford, "Critical Questions for Big Data", *Information, Communication and Society*, Vol. 15, No. 5, 2012, pp. 662–663.

94 See OCHA, *Humanitarianism in the Network Age*, OCHA Policy and Study Series, Geneva, 2013; OCHA, above note 83.

95 Alana Corsi, Fabiane Florencio de Souza, Regina Negri Pagani and João Luiz Kovaleski, "Big Data Analytics as a Tool for Fighting Pandemics: A Systematic Review of Literature", *Journal of Ambient Intelligence and Humanized Computing*, 29 October 2020, available at: https://doi.org/10.1007/s12652-020-02617-4; Pravin Kumar and Rajesh Kr Singh, "Application of Industry 4.0 Technologies for Effective Coordination in Humanitarian Supply Chains: A Strategic Approach", *Annals of Operations Research*, 3 January 2021, available at: https://doi.org/10.1007/s10479-020-03898-w.

96 See UN Global Pulse, above note 8, pp. 24–34; Miguel Luengo-Oroz, "10 Big Data Science Challenges Facing Humanitarian Organizations", UNHCR Innovation Service, 22 November 2016, available at: www.unhcr.org/innovation/10-big-data-science-challenges-facing-humanitarian-organizations; Iffat Idris, *Benefits and Risks of Big Data Analytics in Fragile and Conflict Affected States*, 17 May 2019, available at: https://tinyurl.com/5xmftqcy.

97 C. Kuner and M. Marelli (eds), above note 1, p. 93.

excluding marginalized and under-represented groups such as infants, illiterate persons, the elderly, indigenous communities and people with disabilities – while potentially under-representing some developing countries where digital access is not widespread.[98] Third, specifically for the COVID pandemic, short-term funding of big data projects does not acknowledge the long timelines required to measure health impact. During emerging outbreaks, aid agencies may lack accurate data about case counts, making it challenging to adapt decision-making models.[99] Finally, capacity-building of aid workers in information management is a prerequisite for them to develop the necessary know-how in applying data analysis.

As a matter of law, aid actors must adopt a privacy-first approach with any data collection methods implemented. For crowdsourced data, they must provide adequate information to data subjects, so that their consent is meaningful, instead of an illusory choice. When they use personal data collected for different purposes, they must check that further data processing is compatible and whether anonymization can apply. Failure to address these issues may compromise compliance with core data protection principles.

Though there are many challenges and risks involved, aid actors should adopt technological innovations such as big data in order to address the impact of the pandemic. Past big data projects could serve as case studies for identifying best practices and lessons learned. In any case, humanitarians must ensure that they "do no harm" – i.e., that big data does not cause or exacerbate power inequities. The European Data Protection Board has stressed the importance of protecting personal data during the COVID-19 pandemic, but it has also noted that "[d]ata protection rules … do not hinder measures taken in the fight against the coronavirus pandemic".[100] Even in the context of a pandemic, there is no real dilemma between an effective or a GDPR-compliant use of data. The GDPR does introduce exceptions (e.g. vital interest basis) so as not to hinder access to aid services, while ensuring that privacy principles are respected. Robust data protection policies and practices should help aid actors to mitigate the challenges of big data. Finally, any measure to address the COVID-19 pandemic should be consistent with the aid actor's mandate, balancing all relevant rights, including the rights to privacy and health.

98 Shweta Bansal, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani and Cécile Viboud, "Big Data for Infectious Disease Surveillance and Modeling", *Journal of Infectious Diseases*, Vol. 214, No. 4, 14 November 2016, available at: https://doi.org/10.1093/infdis/jiw400.

99 Caroline Buckee, "Improving Epidemic Surveillance and Response: Big Data Is Dead, Long Live Big Data", *The Lancet Digital Health*, Vol. 2, No. 5, 17 March 2020, available at: https://doi.org/10.1016/S2589-7500(20)30059-5.

100 European Data Protection Board, "Statement by the EDPB Chair on the Processing of Personal Data in the Context of the COVID-19 Outbreak", 16 March 2020, available at: https://edpb.europa.eu/news/news/2020/statement-edpb-chair-processing-personal-data-context-covid-19-outbreak_en.

International Review of the Red Cross (2020), **102** (913), 95–115.
Digital technologies and war
doi:10.1017/S1816383121000060

IRRC

The struggle against sexual violence in conflict: Investigating the digital turn

Kristin Bergtora Sandvik and Kjersti Lohne*
Kristin Bergtora Sandvik (SJD, Harvard Law School, 2008) is a Professor of Legal Sociology at the Faculty of Law, University of Oslo, and a Research Professor in Humanitarian Studies at the Peace Research Institute Oslo (PRIO). Her work focuses on refugee resettlement, legal mobilization, humanitarian technology, ethics, innovation and accountability.

Kjersti Lohne is a Postdoctoral Researcher at the Faculty of Law, University of Oslo. Her work focuses on the sociology of international law, international criminal justice, human rights advocacy and transnational networks.

Abstract
Digital technological innovations make new types of responses to conflict-related sexual violence (CRSV) possible, bringing with them both potential promises and pitfalls. Aiming to provide a conceptual starting point for further analysis, this article problematizes the trend towards data extraction in exchange for aid, protection and justice, and argues for the importance of complementing technology-driven approaches to the struggle against CRSV with the inclusion of strategies for user participation and investment in digital literacy as key aspects of the response. To explore how the digital turn shapes the struggle against CRSV, the article offers a three-part analytical framework. First, the article unpacks how digital technologies create corresponding "digital bodies" – comprised of images,

* Research for this paper was funded by the PRIO-hosted project "Do No Harm: Ethical Humanitarian Innovation" and the University of Oslo-hosted project "Vulnerability in the Robot Society", both funded by the Research Council of Norway.

information, biometrics and other data stored in digital space – which represent the bodies of individuals affected by sexual violence, and which interplay with the risks posed upon the physical bodies of those facing CRSV. Second, the article maps out the role of digital technologies in a cycle of intervention, including prevention of, response to, documentation of and accountability for CRSV. Third, recognizing the increasing importance of data governance to the struggle against CRSV, the article considers how divergent humanitarian, human rights and international criminal law approaches to data may create different pathways for CRSV data. This could also give rise to new tensions in how international actors approach CRSV.

Keywords: conflict-related sexual violence, digital bodies, digital literacy, humanitarian data, sexual violence, human rights, international criminal justice.

Introduction

Historically, the cultural stigma of rape and the international community's long-standing disregard of conflict-related sexual violence (CRSV) have meant that little was known about this form of violence. Currently, CRSV actors – from field responders to international judges – have more than sufficient information regarding the occurrence of CRSV: they are overwhelmed by it and need tools for sorting, identifying and analyzing actionable data and viable paths for action.[1] As a result of global connectivity and the affordability and mass distribution of digital devices, new types of responses to CRSV are now possible. Increasingly, digital technological innovations such as humanitarian mega-databases, or "superplatforms",[2] are harvesting and registering data on a mass scale. Yet, this trend towards the extraction of data from vulnerable individuals in exchange for protection, aid and justice also confronts the struggle against CRSV with new dilemmas. Moreover, CRSV actors must find ways to respond effectively to the proliferation of digital technology-based threats.

1 Elisabeth Jean Wood, "Conflict-Related Sexual Violence and the Policy Implications of Recent Research", *International Review of the Red Cross*, Vol. 96, No. 894, 2014; Carlo Koos, "Sexual Violence in Armed Conflicts: Research Progress and Remaining Gaps", *Third World Quarterly*, Vol. 38, No. 9, 2017; Jelke Boesten and Marsha Henry, "Between Fatigue and Silence: The Challenges of Conducting Research on Sexual Violence in Conflict", *Social Politics: International Studies in Gender, State, and Society*, Vol. 25, No. 4, 2018; Anette Bringedal Houge, "Sexualized War Violence: Knowledge Construction and Knowledge Gaps", *Aggression and Violent Behavior*, Vol. 25, 2015.

2 "Superplatforms", or "platforms of platforms", are giant internet companies that operate across multiple sectors, such as Apple, Facebook, and Google. See, for example, David Porteous and Olga Morawczynski, "The Superplatforms are Coming … and They Will Transform Financial Inclusion", *NextBillion*, 21 December 2018, available at: https://nextbillion.net/the-superplatforms-are-coming-and-they-will-transform-financial-inclusion/ (all internet references were accessed in January 2021). We use this term for the emergence of platforms such as the Office of the UN High Commissioner for Refugees' (UNHCR) PRIMES (see: www.unhcr.org/primes.html) and the World Food Programme's SCOPE (see: https://tinyurl.com/y4axb5br).

Despite considerable interest in and abundant literature on CRSV,[3] there is a dearth of critical research and reflection on the ability of digital technologies to assist in the struggle against it. This article suggests cautious optimism regarding the potential role and relevance of digital technology in preventing, mitigating, treating and punishing CRSV. While we advise against seeing digital technology as a panacea, we emphasize the importance of thinking carefully through the potential for improvements and positive change. The contribution of the article is to offer an analytical framework for assessing the role and relevance of digital technologies, broadly defined as digital devices and data, in addressing CRSV. Our analytical framework has three main components:

1. Digital technologies interact with the insecure and unstable context of conflicts in ways that may produce and exacerbate risk and harm. CRSV actors need a better grasp of how this happens and what this means for the struggle against CRSV. To that end, we introduce the notion of "digital bodies" as a cross-cutting analytical concept to better understand how technologization may also engender risk and harm.[4] Digital technologies are a fundamental part of the contemporary experience and re-conceptualization of the body. At the same time, in conflict-ridden and fragile settings, the provision of data about the body is increasingly a precondition for receiving services, legal protection and justice from the international community. To that end, "digital bodies" as an analytical concept enables cross-sectoral conversations about power and responsibility. Understanding the nature of risk for women's physical bodies as well as women's digital bodies, and the interplay between the two – but also the potential for activism and participation – is crucial for understanding the role and relevance of digital technology in grappling with CRSV. The point is not that women's digital and physical bodies are the same, but that compromising or neglecting the security of digital bodies may be *as consequential* as compromising the security and well-being of physical bodies.
2. We classify and identify digital technological trends and problem framings in the CRSV field through the introduction of a "cycle of intervention", where we assess the digital transformation of the struggle against CRSV in the fields of prevention, response, documentation and accountability. The cycle follows a set of transnational continuums: from the conflict setting to international institutions and global data brokers, and from the prevention stage to the justice and accountability phase, taking place long after the violent event.
3. Recognizing the importance of how data collection, storage, sharing, aggregation and use shape all phases of the struggle against CRSV, we reflect on how divergent perspectives on data may shape the pathways of CRSV data. The article compares and contrasts humanitarian, human rights and

3 See also the thematic issue of the *Review* on "Sexual Violence in Armed Conflict", Vol. 96, No. 894, 2015, available at: https://international-review.icrc.org/reviews/irrc-no-894-sexual-violence-armed-conflict.

4 We understand technologization as the incremental development and application of technology-based approaches.

international criminal justice approaches to data in order to illustrate how new dilemmas may arise with respect to the management and sharing of CRSV data. By looking at the different sectors' objectives, time frames, and perceived agency of those targeted for intervention, we reflect on how the objectives, approaches and values of the three sectors may complement each other, but also create new frictions in the context of the digital turn.

For the purposes of this article, we rely broadly on the United Nations (UN) conceptualization of the term "conflict-related sexual violence" as referring to "rape, sexual slavery, forced prostitution, forced pregnancy, forced abortion, enforced sterilization, forced marriage, and any other form of sexual violence of comparable gravity perpetrated against women, men, girls or boys that is directly or indirectly linked to a conflict".[5] While the article generally addresses all categories of victims/survivors listed in this definition, the analytical focus of the article and the empirical and scholarly contributions on which it draws are largely concerned with the experiences of women. As noted above, so far little attention has been given to digital technologies and CRSV. Thus, to develop a comprehensive conceptual framework for further analysis, the article draws on scholarly and grey literature and media reports on how technology is used to address sexual violence also in non-conflict settings.[6] Acknowledging the resulting partial scope of the article, we hope that it may serve as a resource for others working to plug important knowledge gaps.

The article proceeds in three main parts. In laying out the first part of our analytical framework, we offer a conceptualization of digital bodies. This includes a brief account of how digital technologies transform dynamics of conflict, crisis and injustice, but also constitute a potential for change. We also situate the digital body in a potentially powerful *moral economy* emerging from the digital turn, and the political economy generated by the fight against CRSV.[7] The second component of the analytical framework helps us illustrate how digital technologies shape different aspects of the struggle against CRSV. We introduce a "cycle of intervention", mapping out initiatives involving the use of digital technologies to prevent, provide treatment for, investigate and enhance criminal accountability for sexual violence. For the third part of the analytical framework, we provide a typology for understanding the nature and implications of different perspectives on data in the fields of humanitarian aid, human rights practice and international

5 António Guterres, *Conflict-Related Sexual Violence: Report of the United Nations Secretary-General*, UN Doc. S/2019/280, 29 March 2019, available at: https://peacekeeping.un.org/sites/default/files/annual_report_of_the_sg_on_crsv_2018.pdf.

6 To that end, the article builds on and develops insights from Kristin Bergtora Sandvik, *Technologizing the Fight against Sexual Violence: A Critical Scoping*, PRIO Working Paper, Oslo, 2019, available at: https://gps.prio.org/Publications/Publication/?x=1274; Kristin Bergtora Sandvik, "Digital Dead Body Management (DDBM): Time to Think it Through", *Journal of Human Rights Practice*, Vol. 12, No. 2, 2020, available at: https://doi.org/10.1093/jhuman/huaa002.

7 While the notion of moral economy is used to describe those norms and habits embedded in market rationalities, more broadly it is also concerned with what it is that lends legitimacy to the constitution of markets and the economy. See Susanne Karstedt and Stephen Farrall, "The Moral Economy of Everyday Crime: Markets, Consumers and Citizens", *British Journal of Criminology*, Vol. 46, No. 6, 2006.

criminal justice. We conclude by re-emphasizing the importance of digital literacy and participation in the struggle against CRSV. Our conceptualization of digital literacy is a critical one: it includes not only the capacity to use a device and understand the basic purpose of using it, but also having a basic grasp of issues of law, digital risk and rights, and an awareness of what it means to have a digital body – that is, a body made legible as data. Digital literacy thus goes beyond technical competence to include awareness and perceptions about technology, law, rights and risk.

Re-conceptualizing security through digital bodies

Whether in peace or war, digital technology is a fundamental part of the contemporary experience and re-conceptualization of the body. Via technological means, there is an "intensification of the extension, abstraction, and reconstruction"[8] of the body. The first aspect of our analytical framework concerns how the use of digital technologies creates corresponding "digital bodies" – i.e., images, information, biometrics and other data stored in digital space – that represent the physical bodies of individuals affected by sexual violence, but over which they have little say or control. Understanding this double risk – for the physical body as well as the digital body, and the interplay between the two – is crucial for properly gauging the role and relevance of digital technology in the struggle against CRSV.[9] We argue that the digital body should also be a separate point of departure for security considerations.

Early cultural approaches to the study of cyborgs explored how "digital bodies" operated in the discourses of digital culture to refer to those avatars and images that represented and simulated humans on-screen.[10] While these digital bodies were tropes in popular culture in the 1980s and 1990s, technological innovation has increasingly given them a presence in everyday life. Haggerty and Erickson describe how "surveillant assemblages" operate by abstracting human bodies from their territorial settings and separating them into a series of different flows, to be reassembled in different locations as discrete and virtual "data doubles", which can be scrutinized and targeted for intervention.[11] In this way,

8 Chris Shilling, "The Body in Sociology", in Claudia Malacrida and Jacqueline Low (eds), *Sociology of the Body*, Oxford: Oxford University Press, 2008; Carey Jewitt, Sara Price and Anna Xambo Sedo, "Conceptualising and Researching the Body in Digital Contexts: Towards New Methodological Conversations across the Arts and Social Sciences", *Qualitative research*, Vol. 17, No. 1, 2017.

9 See K. B. Sandvik, *Technologizing the Fight against Sexual Violence*, above note 6. See also Kristin Bergtora Sandvik, "Making Wearables in Aid: Digital Bodies, Data and Gifts", *Journal of Humanitarian Affairs*, Vol. 1, No. 3, 2019; Kristin Bergtora Sandvik, "Wearables for Something Good: Aid, Dataveillance and the Production of Children's Digital Bodies", *Information, Communication & Society*, Vol. 23, No. 14, 2020.

10 Donna Haraway, "Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s", *Socialist Review*, No. 80, 1985. A cyborg is a person whose physiological functioning is aided by or dependent upon a mechanical or electronic device; see: www.dictionary.com/browse/cyborg.

11 Kevin D. Haggerty and Richard V. Ericson, "The Surveillant Assemblage", *British Journal of Sociology*, Vol. 51, No. 4, 2000.

translating human identity into information patterns not only provides more information, it also creates new conceptions of identity. The body itself becomes the source of information – the coded body can "talk". As illustrated by scholarly analysis of migration management, this may imply that "a talking individual, who owns the body, is in fact seen as unnecessary" and, more importantly, may be considered insufficient and even suspect as a source of identification, as the coded body is considered more "truthful".[12]

Much of the literature investigating the *gendered* dimensions of the digital body takes as its starting point the social nature of the body, and is concerned with contestations over sex/gender/nature/culture, and the bodywork undertaken to forge links between physical and virtual bodies.[13] As with the broader feminist discourse on anti-essentialism,[14] there is little agreement about what a body is. However, with increasing technologization, attention must be paid to how digital technologies have become instruments facilitating the making of truth claims about the body.[15] In the context of CRSV, this concerns what happened when, to whom, by whom and how. While digital technologies offer novel ways of constructing and communicating about gender, gender roles and gendered violence, they also bring the possibility of new modes of disempowerment and abuse, with implications for real-life sexual violence.[16]

It is important to reiterate that for the CRSV context, the digital body is not a metaphysical concept. In conflict-ridden and fragile settings, it is common practice for underserved, abandoned and structurally discriminated-against communities to be called upon to provide data as a precondition for receiving services, legal protection and justice in the form of legal accountability, truth and reparations.[17] Whereas technology actors and States engage in heavy dataveillance of consumers/citizens,[18] the extra-democratic governance structures of the international community and the vulnerability of individuals and communities in crisis magnify the power exercised over such communities and exacerbate existing

12 Katja Franko Aas, "'The Body Does not Lie': Identity, Risk and Trust in Technoculture", *Crime, Media, Culture*, Vol. 2, No. 2, 2006.

13 Kate O'Riordan, "Revisiting Digital Technologies: Envisioning Biodigital Bodies", *Communications*, Vol. 36, No. 3, 2011.

14 Candace West and Don H. Zimmerman, "Doing Gender", *Gender & Society*, Vol. 1, No. 2, 1987.

15 Theresa M. Senft, "Introduction: Performing the Digital Body – a Ghost Story", Women & Performance: A Journal of Feminist Theory, Vol. 9, No. 1, 1996, available at: https://tinyurl.com/y3nmjhmq; Breanne Fahs and Michelle Gohr, "Superpatriarchy Meets Cyberfeminism: Facebook, Online Gaming, and the New Social Genocide", *MP: An Online Feminist Journal*, Vol. 3, No. 6, 2010.

16 See Ian Sample, "Internet 'Is not Working for Women and Girls', Says Berners-Lee", *The Guardian*, 12 March 2020, available at: www.theguardian.com/global/2020/mar/12/internet-not-working-women-girls-tim-berners-lee.

17 Kristin Bergtora Sandvik, Katja Lindskov Jacobsen and Sean Martin McDonald, "Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation", *International Review of the Red Cross*, Vol. 99, No. 904, 2017; Mirca Madianou, "Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises", *Social Media + Society*, Vol. 5, No. 3, 2019.

18 Dataveillance is the systematic monitoring of people or groups by means of personal data systems in order to regulate or govern their behaviour. Sara Degli Esposti, "When Big Data Meets Dataveillance: The Hidden Side of Analytics", *Surveillance & Society*, Vol. 12, No. 2, 2014, available at: https://doi.org/10.24908/ss.v12i2.5113.

power differences between organizations (with their private sector partners), on the one hand, and communities in crisis, on the other. Increasingly, this form of governance centres on extracting data in exchange for aid, protection and justice.[19] In what follows, we explore the nature of risk for women's physical and digital bodies posed by digital transformations – but also the potential for activism and participation – as a means to start analyzing the role and relevance of digital technologies in the struggle against CRSV.

Digital transformations, harms, solution(ism)s and change

For those seeking to remedy harm in conflict settings, the availability and use of mobile phones, social media platforms, satellites, drones, digital cash and biometric technology have transformed how human suffering is identified, registered, understood and addressed, by whom, and from where.[20] At the same time, the opportunities provided by technological developments exacerbate gendered risks[21] and harms and produce new ones.[22] The sophistication and affordability of off-the-shelf commercial devices enable intimate surveillance and the production of false information and fake imagery – i.e., false digital bodies. For example, "deepfake" pornographic videos of Chandrani Murmu, Indian's youngest parliamentarian, were circulated as part of the widespread "trolling" and online sexual harassment of India's female politicians.[23]

Moreover, digital technologies also shape the way in which individuals experience violence. Global connectivity extends the reach of offenders and deprives victims/survivors of privacy by facilitating access to information, providing private ways to communicate, preserving images of assaults, and making anonymous harassment possible. Over the last ten to fifteen years, it has been increasingly recognized that "technology-related forms of violence against women cause psychological and emotional harm, reinforce prejudice, damage reputation, cause economic loss and pose barriers to participation in public

19 K. B. Sandvik, "Wearables for Something Good", above note 9.

20 See K. B. Sandvik, "Making Wearables in Aid", above note 9; Kristin Bergtora Sandvik, Maria Gabrielsen Jumbert, John Karlsrud and Mareile Kaufmann, "Humanitarian Technology: A Critical Research Agenda", *International Review of the Red Cross*, Vol. 96, No. 893, 2014; K. B. Sandvik, K. L. Jacobsen and S. M. McDonald, above note 17. See also special issue of the *International Journal of Transitional Justice* on "Technology and Transitional Justice", Vol. 13, No. 1, 2019.

21 For an exploration of the gendered aspects of risk – that is, how risk shapes the lives of different genders because of their gender – see Kelly Hannah-Moffat and Pat O'Malley (eds), *Gendered risks*, Routledge-Cavendish, London and New York, 2007.

22 We also note the risk that digital technology reproduces biases and discrimination in society and in data sets. For example, if the working assumption is that CRSV only affects women and girls, then men, boys, and sexual and gender minority individuals will remain invisible. For an analysis of two examples of algorithmic exclusion of men and boys – in the UNHCR's vulnerability assessments and in drone strikes – see Kristin Bergtora Sandvik, "Technology, Dead Male Bodies, and Feminist Recognition: Gendering ICT Harm Theory", *Australian Feminist Law Journal*, Vol. 44, No. 1, 2018.

23 See Eliza Mackintosh and Swati Gupta, "Troll Armies, 'Deepfake' Porn Videos and Violent Threats: How Twitter Became So Toxic for India's Women Politicians", *CNN*, available at: https://edition.cnn.com/2020/01/22/india/india-women-politicians-trolling-amnesty-asequals-intl/index.html.

life".[24] Nevertheless, there is little reporting of or response to the intersections of CRSV and digital technologies, and the way risk and harm are evolving is insufficiently understood, particularly in the international domain.[25] Our use of the concept of digital bodies aims to make visible both the multiple emergent forms of CRSV risk and the need for interventions to address the protection needs of this "double" body – both physical and digital.

Digital technologies are also significantly shaping the potential to mitigate and respond to conflicts and other situations of violence. Technology provides tools for victims/survivors and those who work with them or on their behalf. Digital devices and platforms can give victims/survivors therapeutic space by enabling them to receive assistance, speak out, share their stories, and gain recognition of harm and trauma, thus offering possibilities for social change.[26] At the same time, we underline the importance of giving due consideration to digital literacy and the digital body in designing interventions.

Due to the scarcity of analysis on the use of digital technologies in the struggle against CRSV, we here briefly note the mushrooming of initiatives offering digital technology-based solutions to sexual violence in non-conflict settings: these include digital platforms and blockchain technologies designed to raise awareness, advocate for change and offer possibilities for protection, reporting and the crowdsourcing of justice. Such initiatives have been animated by the #MeToo campaign and its interweaving of feminist consciousness – an awareness of women's inequality and a commitment to remedy it – and Silicon Valley solutionism – recognizing "problems as problems based on just one criterion: whether they are 'solvable' with a nice and clean technological solution".[27] We have seen, for example, the emergence of sophisticated legal technology dealing with sexual violence, such as Callisto, a blockchain-based matching system that allows survivors to share their stories and securely connects victims of the same perpetrator to identify repeat offenders,[28] and LegalFling, a platform for uploading consent to sexual activity.[29]

From a CRSV perspective, these approaches may appear to abstract the issue of sexual violence from its systemic context. Access to digital devices is assumed, as is a focus on individual agency and generally high levels of digital

24 Katerina Fialova and Flavia Fascendini, *Voices from Digital Spaces: Technology-Related Violence against Women*, Association for Progressive Communications, 2011; Archana Barua and Ananya Barua, "Gendering the Digital Body: Women and Computers", *AI & Society*, Vol. 27, No. 4, 2012; Rhonda Shaw, "'Our Bodies, Ourselves', Technology, and Questions of Ethics: Cyberfeminism and the Lived Body", *Australian Feminist Studies*, Vol. 18, No. 40, 2003.

25 But see Dubravka Šimonović, *Report of the Special Rapporteur on Violence against Women, Its Causes and Consequences on Online Violence against Women and Girls from a Human Rights Perspective*, UN Doc. A/HRC/38/47, 18 June 2018, available at: https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/47.

26 See, for example, "Sexual Assault and Technology Misuse", *VAWnet*, available at: https://vawnet.org/events/sexual-assault-and-technology-misuse.

27 Evgeny Morozov, "The Perils of Perfection", *New York Times*, 2 March 2013, available at: www.nytimes.com/2013/03/03/opinion/sunday/the-perils-of-perfection.html; see also Evgeny Morozov, *To Save Everything, Click Here: The Folly of Technological Solutionism*, PublicAffairs, New York, 2013.

28 See the Callisto website, available at: www.projectcallisto.org.

29 See the LegalFling website, available at: https://legalfling.io/.

literacy – and the problem and solution are both "produced" in the global North. However, we suggest that they are also examples of serious attempts to rethink the gender-based and sexual violence equation, offer visibility, build feminist consciousness and strengthen advocacy in new ways.[30] As such, these innovations are a starting point, not an end point, for thinking about the use of digital tools in the struggle against CRSV.

Moral economies and the power of problem framing

Furthermore, we must consider the power of problem framing as two contemporary trends intersect. Across the humanitarian, human rights and international criminal justice sectors, the use of digital technologies is gaining traction as a way to make service delivery, truth, accountability and justice more efficient and cheaper.[31] This faith in technology is supported by a burgeoning "ICT for good" literature making optimistic claims about the capacity of technology to solve political and social issues.[32] In parallel, a political economy of sexual violence has emerged whereby the moral, political, legal and financial attention given to CRSV risks crowding out agency, participation and recognition of the continuum of violence affecting women in conflict, which includes but is not limited to sexual violence.[33] While acknowledging the plethora of different policies and tools developed by international, State and non-State organizations alike to prevent CRSV, there is concern that the contemporary and dominant focus on sexual violence may sideline alternative framings with respect to women's insecurity.[34]

Without undermining the attention that CRSV (finally) elicits, the converging interests among activists, academics and politicians – as concerns both the focus on digital technologies and CRSV – should give pause for thought. Especially, difficult questions should be asked about the potential for a powerful moral economy of the technologized struggle against CRSV: who frames problems and solutions when it comes to CRSV, and what are the direct and indirect burdens of this framing? As well, who benefits from the assumption that digital technologies will facilitate the struggle against CRSV, and why? What does it mean for how we calibrate the vulnerability – and utility – of digital bodies in

30 See Doreen Raheena Sulleyman, "Sexual Assault in Ghana: How Technology Can Help Build Visibility", *GenderIT.org*, 6 February 2019, available at: www.genderit.org/feminist-talk/sexual-assault-ghana-how-technology-can-help-build-visibility.

31 Kristin B. Sandvik and Kjersti Lohne, "The Rise of the Humanitarian Drone: Giving Content to an Emerging Concept", *Millennium*, Vol. 43, No. 1, 2014.

32 Patrick Meier, "New Information Technologies and Their Impact on the Humanitarian Sector", *International Review of the Red Cross*, Vol. 93, No. 884, 2011. For a discussion of this literature, see Kristin Bergtora Sandvik, "Now Is the Time to Deliver: Looking for Humanitarian Innovation's Theory of Change", *Journal of International Humanitarian Action*, Vol. 2, No. 1, 2017, p. 8.

33 Paul Kirby and Laura J. Shepherd, "The Futures Past of the Women, Peace and Security Agenda", *International Affairs*, Vol. 92, No. 2, 2016. See also Maria Eriksson Baaz and Maria Stern, *Sexual Violence as a Weapon of War? Perceptions, Prescriptions, Problems in the Congo and Beyond*, Zed Books, London and New York, 2013.

34 See also Anette Bringedal Houge and Kjersti Lohne, "End Impunity! Reducing Conflict-Related Sexual Violence to a Problem of Law", *Law & Society Review*, Vol. 51, No. 4, 2017.

the CRSV context? To scope the range of possibilities and provide a shared frame of reference for critical discussion, the next section maps a cycle of digital technological intervention to CRSV.

Digital technologies and CRSV: Mapping a cycle of intervention

The second part of our analytical framework provides an analytical approach to classifying and identifying digital technological trends and problem framings in the CRSV field. We map a cycle of digital technological interventions, where gadget distribution and acceptance combine with strategies for large-scale data harvesting and effective data management. The cycle is mapped onto a transnational continuum from the conflict setting to international institutions and global data brokers, and from the preventive phase to efforts to obtain justice and accountability occurring long after the violence.

Prevention

An important insight emerging from the last two decades of analysis of CRSV is that sexual violence is not necessarily random, unplanned and unforeseeable, but is often preceded by rumour and hate speech, sudden or irregular troop movements, the separation of men and women at checkpoints and so forth.[35] Early warning is therefore key for preventing sexual violence, and digital technologies can be an essential element of this response. In particular, data-driven predictive approaches have significant potential in contexts where connectivity, mobile phone ownership and social media use are widespread. This includes prediction based on the analysis of big data sets,[36] the identification of sites of sexual violence and harassment,[37] and the use of automated detection based on machine learning and natural language processing to identify patterns of hate speech and rumour in social media that dehumanize and sexualize particular groups or individuals. Satellite surveillance footage and drone imagery can also help map physical movements and actions and identify indicators of heightened risk of sexual violence.[38]

At the same time, technology-driven early-warning initiatives raise difficult questions about operability, relevance and risk. At a technical level, there will always be issues related to algorithmic bias, access to relevant local data traffic, and

35 E. J. Wood, above note 1. See also António Guterres, *Report of the Secretary-General on Conflict-Related Sexual Violence*, 16 April 2018, available at: www.un.org/sexualviolenceinconflict/wp-content/uploads/reports/sg-reports/SG-REPORT-2017-CRSV-SPREAD.pdf.

36 Dara Kay Cohen and Ragnhild Nordås, "Do States Delegate Shameful Violence to Militias? Patterns of Sexual Violence in Recent Armed Conflicts", *Journal of Conflict Resolution*, Vol. 59, No. 5, 2015.

37 Chelsea Young, "HarassMap: Using Crowdsourced Data to Map Sexual Harassment in Egypt", *Technology Innovation Management Review*, Vol. 4, No. 3, 2014.

38 Joshua Lyons "Documenting Violations of International Humanitarian Law from Space: A Critical Review of Geospatial Analysis of Satellite Imagery during Armed Conflicts in Gaza (2009), Georgia (2008), and Sri Lanka (2009)", *International Review of the Red Cross*, Vol. 97, No. 866, 2012.

adequate local knowledge and translation capacity. Aerial surveillance is useful only if timely and of high quality, and where enough imagery analysis capacity is available. Satellites are expensive, and drones require significant local infrastructure. Moreover, surveillance is not necessarily predictive: the violence may already be ongoing. Neither is it active: knowing about atrocities is not the same as acting on them, and there is no necessary causal link between remote monitoring and protection efforts being implemented on the ground. Similarly, the assumption that there is a causal relationship between early warning and self-rescue is problematic. For example, distributing free phones to women at risk is a fairly common strategy. However, in addition to possibly revealing the whereabouts of their users, phones may get lost or be confiscated, or may not work, and they may also create a dynamic whereby women are (inherently) put under pressure to "produce cases" in order to prove efficacy. In addition, the notion of data having an inherent protective effect is based on the assumption that increased amounts of unique and otherwise unobtainable data over wide geographic areas and/or non-permissive environments result in targeted communities having an early warning, which enables them to make better and quicker decisions that are potentially life-saving.[39]

The impact of this type of early-warning system – whether organized by external actors or community-based – will also depend on trust in technology and in the message itself. Over the last two decades, digital technologies have been used for awareness-raising, consciousness-raising, training and capacity-building. For grassroots and community-based actors, they are used to document and disseminate information about harms and threats, create digital support networks, give early warnings and trigger support from powerful constituents.[40] Yet in this regard, digital technologies occupy an ambiguous position. On the one hand, for activists, technology produces indisputable "facts", testimonies, or evidence of events in formats familiar and acceptable to those in power. Even if the physical body cannot speak with credibility – for example, when we take into account that refugee narratives are often met with distrust[41] – digital devices are seen as credible conveyors of information.

On the other hand, in recent years, countering misinformation (false information not created to do harm), disinformation (false information created to do harm) and malinformation ("true" information used to inflict harm) has become increasingly complicated.[42] In the past, the aim was to harness the power

39 Kristin Bergtora Sandvik and Nathaniel Raymond, "Beyond the Protective Effect: Towards a Theory of Harm for Information Communication Technologies in Mass Atrocity Response", *Genocide Studies and Prevention: An International Journal*, Vol. 11, No. 1, 2017.

40 Molly K. Land and Jay D. Aronson (eds), *New Technologies for Human Rights Law and Practice*, Cambridge University Press, Cambridge, 2018.

41 Kristin Bergtora Sandvik, "The Physicality of Legal Consciousness: Suffering and the Production of Credibility in Refugee Resettlement", in Richard D. Brown and Richard Ashby Wilson (eds), *Humanitarianism and Suffering: The Mobilization of Empathy*, Cambridge University Press, Cambridge, 2008.

42 Council of Europe, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*, 2017, available at: www.coe.int/en/web/freedom-expression/information-disorder.

of social media to do good, but social media is now also seen as a source of harm.[43] The rise of deepfake images/videos and generative adversarial networks[44] poses particular problems for early warning. For example, rumours about rapes and child killings usually trigger acts of revenge. Graphical illustrations of such crimes can prove fatal, regardless of whether or not they are verifiable. False attribution of imagery producing false digital bodies –such as a deepfake app generating nude images of women[45] – exacerbates vulnerability.

All these factors dovetail with more general issues related to cyber security and the use by governments and armed non-State actors of surveillance, kill switches and offensive cyber weapons. We therefore urgently need early-warning approaches that will identify hate speech and rumours, produce evidence that they are untrue and, in culturally meaningful ways, rapidly disseminate counter-messages to diffuse potential violence.[46] At the same time, it seems clear that the future of early warning depends on participatory approaches, such as training and capacity-building designed to enhance user competence. Most importantly, increased attention to digital literacy in early-warning approaches will enable communities to gauge the scope of digital manipulation, undertake value assessments about the kind of sexual violence produced by digital devices and the type of gendered harm it causes, and make realistic risk assessments.

Response

There are many examples of how digital technologies are helping activists to self-protect and to organize community protection. For example, crowdsourcing approaches provide documentation of locations and stories of assault,[47] and apps send the times and GPS coordinates of arrests to families, fellow activists, legal advisers and social media outlets.[48] These technologies can also offer individuals at risk and survivors of sexual violence access to services when there is stigma and generalized insecurity, and can help field responders to achieve a greater degree of internal coordination and coherence and to avoid duplication of

43 Sam Gregory, "Cameras Everywhere Revisited: How Digital Technologies and Social Media Aid and Inhibit Human Rights Documentation and Advocacy", *Journal of Human Rights Practice*, Vol. 11, No. 2, 2019.

44 See Joseph Rocca, "Understanding Generative Adversarial Networks", *Towards Data Science*, 7 January 2019, available at: https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29.

45 See James Vincent, "New AI Deepfake App Creates Nude Images of Women in Seconds", *The Verge*, 27 June 2019, available at: www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography.

46 On the trade-offs between freedom of speech and security, as well as the risks inherent in limiting freedom of speech, see David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc. A/74/486, 9 October 2019, available at: www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf.

47 See Anastasia Powell and Tully O'Neill, "Cyber Justice: How Technology is Supporting Victim-Survivors of Rape", *The Conversation*, 5 April 2016, available at: https://theconversation.com/cyber-justice-how-technology-is-supporting-victim-survivors-of-rape-56022.

48 Sheila Peuchaud, "Social Media Activism and Egyptians' Use of Social Media to Combat Sexual Violence: An HiAP Case Study", *Health Promotion International*, Vol. 29, No. 1, 2014.

services and victim interviewing. Digital technologies can provide gender-sensitive screening tools, and training platforms for field responders. Digital screening tools are being developed, along with multiple tools for navigating services and accessing information safely: so-called digital "safe spaces".[49] And while record-keeping in conflict and fragile contexts is challenging, digitized health records are less likely to be lost or destroyed, as long as the integrity of the database is intact.

Ideally, digital technologies can be used to better safeguard patients' physical security and confidentiality. For therapeutic interventions to be effective, accessible information about the nature and function of particular treatments is crucial, and digital technologies can provide this information. In settings with few resources, there is an acute shortage of mental health professionals; in such contexts, Internet-based interventions, text messaging, and mobile phone- and smartphone-based interventions may enhance active self-care among trauma survivors, extend the geographic reach of health-care providers, and facilitate the use of paraprofessionals and peer mental health support.[50]

While these assessments are realistic, the focus on resilience, self-care and self-responsibilization is potentially problematic as a justification for digital health-care initiatives in fragile settings, because – as with prevention and early warning – their effectiveness hinges on the existence of digital literacy, trust, access and functionality, and on the effective mitigation of digital risk. A "safe space" is a questionable entity in the context of conflict: there is a high risk that one's physical geolocation will be revealed and one's personal data compromised. On the other hand, databases in themselves constitute potent targets for cyber attacks. The sensitive nature of health data means that the repercussions of negligent or inadvertent leaks and hacks can be serious not only for the digital body of the patient but also for her physical security. Furthermore, a digital user roadmap for services is only useful if the services are actually accessible to and meaningful for users.

Documentation

In the past, the cultural stigma attached to rape and the international community's long-standing lack of attention to sexual violence in conflict have meant that little was known about this type of violence. Today, however, things are different. Here we identify two specific challenges to CRSV documentation, after first identifying some of the forms that CRSV data may take.

CRSV data can consist of surveys (victimization rates), medical data (such as patient medical records, medical certificates, and sexual assault medical forensic

49 Andrea L. Wirtz *et al.*, "Development of a Screening Tool to Identify Female Survivors of Gender-Based Violence in a Humanitarian Setting: Qualitative Evidence from Research among Refugees in Ethiopia", *Conflict and Health*, Vol. 7, No. 1, 2013, available at: ww.ncbi.nlm.nih.gov/pmc/articles/PMC3695841/.

50 Josef Ruzek and Carolin M. Yeager, "Internet and Mobile Technologies: Addressing the Mental Health of Trauma Survivors in Less Resourced Communities", *Global Mental Health*, Vol. 4, 2017; A. L. Wirtz *et al.*, above note 49.

examinations, commonly known as "rape kits"[51]), crime reports, perpetrator data, public (media) reports or proxy data (such as pregnancies resulting from rape). Such data can be used to reinterpret testimonies in order to identify formerly overlooked patterns of sexual violence, corroborate or confirm witness accounts and detect gender bias in documentation.[52]

There are still many CRSV situations in which data may be non-existent or of poor quality due to collection problems, bias or "digital shadows" where access to digital devices and connectivity are limited;[53] effective data analysis may be hampered by low levels of data literacy in the practitioner community and so forth.[54] Yet, the problem is often no longer data scarcity. While UN Security Council Resolution 1325 of 2000 noted "the need to consolidate data on the impact of armed conflict on women and girls", currently, CRSV actors – from field responders to international judges – are deluged with information about sexual violence, and need tools for data verification, analysis, and making responses operational and effective.

A different kind of challenge concerns the moral dilemma arising in the documentation of CRSV, from field responses to criminal prosecutions. This relates to the operational requirement for first-person testimony and the resulting practice, whereby survivors must conduct multiple interviews with multiple stakeholders, targeted towards multiple audiences. This practice, which may result in re-victimization, epitomizes one of the great failings of the sexual violence bureaucracy. The fundamental problem is the way survivor agency and ownership of the story comes up against institutional needs for credible evidence and believable narratives capable of generating empathy among politicians. With care, it might be possible to create a framework that would record and assess the veracity of victim testimonies from survivors, and ensure that they are the ones in control of the access codes – i.e., that there is one single gateway to the digital body for bureaucracies and service providers to scrutinize and engage with.[55]

However, this solution also raises a number of issues, such as community acceptance, the standardization of process and quality, difficulties in guaranteeing the relevance of the statement to all stakeholder audiences, variations in end-user competence and digital acceptance, and problems related to unauthorized access,

51 Jaimie Morse, "Documenting Mass Rape: Medical Evidence Collection Techniques as Humanitarian Technology", *Genocide Studies and Prevention: An International Journal*, Vol. 8, No. 3, 2014, p. 72.

52 Tony Roberts and Gauthier Marchais, "Assessing the Role of Social Media and Digital Technology in Violence Reporting", *Contemporary Readings in Law and Social Justice*, Vol. 10, No. 2, 2018.

53 This lack of access to digital technologies is also known as the "digital divide". For discussion on this term, see Lina Gurung, "The Digital Divide: An Inquiry from Feminist Perspectives", *Dhaulagiri Journal of Sociology and Anthropology*, Vol. 12, 2018; Jennifer Radloff, "Digital Security as Feminist Practice", *Feminist Africa*, No. 18, 2013.

54 Mark Latonero and Zachary Gold, Data, *Human Rights and Human Security*, Data & Society Research Institute, 2015.

55 Access codes are a series of letters and numbers that allow access. The idea, of course, is that legal regulations apply once consent to access has been given. The authors are grateful to the participants at the Expert Roundtable on "Using Tech Innovation to Combat Conflict-Related Sexual Violence", held in Geneva on 18–19 February 2019 and hosted by Legal Action Worldwide and the Office of the UN High Commissioner for Human Rights, for this point.

destruction and manipulation.[56] Moreover, policies on the storage, sharing and destruction of such testimonies raise extremely difficult questions about ownership and control.[57] Digital technological approaches to documenting CRSV thus require participation and digital literacy to be made core components – but even when solutions are designed that focus on these objectives, old issues pertaining to the meaningful participation of victims/survivors will persist.

Accountability

Through digital devices, users produce a mass of data that can be used for establishing criminal accountability; this includes text messages, multimedia messages, metered data (numbers dialled, time and date of calls etc.), emails, internet browsing data, image, sound and audio files, and geolocation data.[58] For the prosecution of sexual violence, including CRSV, in a domestic or international court, new technology requires standardized and uniform digital forensic guidelines and methods (proper evidence handling, including preservation, storage and maintenance of the chain of custody) for the documentation, collection and preservation of both digital and physical evidence. Thus, digital forensic approaches to sexual violence require investigators, forensic scientists, medical staff and lawyers, among others, to have new types of expertise and training.[59]

A different type of quandary relates to the institutional challenge of keeping in contact with witnesses, both to keep them apprised of developments and to get witnesses to testify when required. Technical solutions, but also political developments, make this increasingly feasible: in particular, digital identities are increasingly bound up with pushes for legal identity as an aim under Target 16.9 of the UN Sustainable Development Goals, whereby a legal identity is presumed to increase access to basic services, financial inclusion, social integration and regularization.[60] When wrapped into a digital format, legal identities make people instantly trackable. In the CRSV domain, this "trackability" presents some difficult issues: prosecutors' offices and victims' units obviously have an interest in finding ways to keep in touch with potential witnesses, and the difficulty of locating and communicating with survivors is a serious obstacle to adjudicating sexual violence crimes, but hard questions arise about the pros and cons

56 Marie-Helen Maras and Michelle D. Miranda, "Overlooking Forensic Evidence? A Review of the 2014 International Protocol on the Documentation and Investigation of Sexual Violence in Conflict", *Global Security: Health, Science and Policy*, Vol. 2, No. 1, 2017.

57 For a good discussion of data protection in emergency contexts, see Ben Hayes, "Migration and Data Protection: Doing No Harm in an Age of Mass Displacement, Mass Surveillance and 'Big Data'", *International Review of the Red Cross*, Vol. 99, No. 904, 2017.

58 M.-H. Maras and M. D. Miranda, above note 56.

59 *Ibid.*

60 See UNHCR, *Global Virtual Summit on Digital Identity for Refugees, Concluding Workshop: Summary Conclusions and Recommendations*, August 2019, available at: www.unhcr.org/idecosystem/wp-content/uploads/sites/69/2019/12/Conclusions_and_Recommendations.pdf.

(including potential risks) of being a digitally accessible – and digitally identifiable – "stand by" victim of sexual violence.

Divergent perspectives on data: Tensions at the interface of humanitarian action, human rights practice and international justice

In recognition of the importance of data governance for the struggle against CRSV, the third part of our analytical framework interrogates how divergent perspectives on data may shape the pathways of CRSV data. In this section, we examine the implications of data extraction and use for humanitarianism, human rights and international criminal justice. These fields have been selected because they have a stronger legal tenor and are more closely interlinked than adjacent fields dealing with CRSV such as peacebuilding, development or transitional justice. A different scope – and a different set of analytical combinations – is of course possible.

Whereas humanitarianism sees misfortune and suffering, human rights sees injustice, discrimination and inequality. International criminal justice, in turn, sees atrocity, barbarism and impunity. We suggest that for the digital transformation of the struggle against CRSV, understanding the differences in how these three fields see survivor bodies – and the function and uses of digital bodies – is crucial. To sketch out how they relate to divergent perspectives on data, we consider some of the overlaps and tensions across these three different international response sectors[61] and identify key structural differences between them in terms of their mission objectives, their time frames and their perceptions of individual and communal agency for those targeted for intervention. Our approach to the different sectors consists of so-called Weberian "ideal types" – simplifications used as analytical tools.[62] This entails that we are not as concerned with hybridity and nuance within and among the sectors as we are with mapping out trends and tendencies. Neither is the critical and significant role of *domestic* actors considered. Instead, we aim to increase our understanding of how data collected by one particular actor for a particular purpose within a particular time frame passes on to other actors by being either borrowed or appropriated — and what the consequences of such transfers are. As such, we turn our attention to how background factors shape the different perspectives of humanitarianism, human rights and international criminal justice actors on "repurposing" data.

Animated by the call to "do something" about human suffering and global violence, the similar ethical and legal motivations of human rights, humanitarianism and international criminal justice have led to the widespread supposition that the

61 As noted previously, this section develops insights articulated in K. B. Sandvik, above note 6; see also Sara Kendall and Sarah Nouwen, *International Criminal Justice and Humanitarianism*, University of Cambridge Faculty of Law Research Paper 69, 2018.

62 See "Ideal Type", *Encyclopaedia Britannica*, available at: www.britannica.com/topic/ideal-type.

spheres of these fields seamlessly overlap. In the cycle of intervention from early warning to accountability, the different sectors are seen as fulfilling complementary roles, each providing something needed by the others: humanitarians supply services, human rights actors look for truth, and international criminal justice produces legal accountability. Often, and this is the ideal, the cycle of intervention is imagined as running so smoothly that the sectors merge and blur into an uninterrupted course of action, as when human rights actors provide evidence for international criminal prosecutions, or when international criminal justice provides humanitarian relief through victim reparation. However, norms and good causes do not necessarily align just because we want them to.

In what follows, we disrupt this notion of unproblematic continuity from one sector to the next and consider some of the tensions between these sectors in relation to their use of data. It should be emphasized that these tensions concern issues of great practical significance. For example, the idea of human rights and international criminal justice actors using their beneficiary data is contentious for humanitarian actors, as it may threaten their principles of neutrality and impartiality and accordingly cause harm by limiting their humanitarian access. For these reasons, some humanitarian organizations have policies of non-cooperation with the International Criminal Court.[63] In other words, despite their shared ethics of addressing human suffering, the sectors operate with divergent views of who they can legitimately borrow, extract or receive data from. The differences between the humanitarian, human rights and international criminal justice sectors thus lead to different perspectives on how to harvest, manage, use and share data. In what follows, we show how these tensions are embedded in the objectives of the different sectors, and in their relationship and approach to time and agency.

Objectives

The different views and approaches to data are closely related to the different agendas of each sector. The aim of humanitarianism is to address needs and save lives, which entails that it – in theory –has little concern for the connection between justice and peace. Human rights aims to provide accountability, transparency and justice, whereas international criminal justice is interested in establishing and adjudicating individual criminal responsibility and legal punishment; it sees criminal justice as a precondition for peace. While the "politicization" of emergencies – including situations of mass violence – is a problem for humanitarianism because it may interfere with the response to humanitarian needs, human rights actors, from their perspective, may consider

63 Fabrice Weissman, "Humanitarian Aid and the International Criminal Court: Grounds for Divorce", Centre de Réflexion sur l'Action et les Savoirs Humanitaires, Fondation Médecins Sans Frontières, 2009, available at: www.msf-crash.org/en/publications/rights-and-justice/humanitarian-aid-and-international-criminal-court-grounds-divorce.

trade-offs in the interests of producing accurate, credible data and securing formal rights protection. These differences matter in terms of what data can be used for. They also matter for secondary objectives: while persuasion, public condemnation and prosecution are complementary modes of action in response to CRSV, they also involve significant tensions regarding the proper handling of data. This might include contestations over how data is collected and the "proper" norms of its collection; data ownership and questions with respect to the interests driving the consolidation of databases; and also issues of consent (what is good enough, how long does it last and for what uses may it be extended?).

While their mandates differ, as does their support for criminal prosecutions, most humanitarian organizations support prosecution as a response to breaches of international criminal and humanitarian law, including CRSV.[64] However, humanitarian organizations may be reluctant to participate in judicial processes because doing so may jeopardize humanitarian access to vulnerable populations.[65] For this reason, humanitarian actors may choose not to actively participate in international criminal procedures, as the "purpose of humanitarian action is, above all else, to save lives, not to establish criminal responsibility".[66] Human rights organizations have shown much more willingness to contribute to international criminal prosecutions.[67] Human rights NGOs are at the forefront of the fight against impunity for CRSV, as the criminal justice system offers a tool to enforce human rights.[68] Reports by human rights organizations have therefore also been submitted as evidence before international criminal courts and tribunals. This practice has attracted significant criticism,[69] precisely because of the different mandates of human rights and international criminal justice actors and what this means for their respective approaches to data. These differences also determine how and to what extent the digital bodies of CRSV victims/survivors are drafted into the accomplishment of institutional objectives.

Temporalities

There are also important temporal differences between the three sectors. To return to the cycle of intervention, humanitarian actors are often first to the scene, and may possess first-hand information on sexual violence, survivors and perpetrators. They may also have knowledge of the nature and scale of the violence, including whether

64 Anne-Marie La Rosa, "Humanitarian Organizations and International Criminal Tribunals, or Trying to Square the Circle", *International Review of the Red Cross*, Vol. 88, No. 861, 2006.
65 *Ibid.*
66 *Ibid.*
67 Kjersti Lohne, *Advocates of Humanity: Human Rights NGOs in International Criminal Justice*, Oxford University Press, Oxford, 2019. See also Kjersti Lohne, "Penal Humanitarianism beyond the Nation State: An Analysis of International Criminal Justice", *Theoretical Criminology*, Vol. 24, No. 2, 2020.
68 Karen Engle, Zinaida Miller and Denys Mathias Davis, *Anti-Impunity and the Human Rights Agenda*, Cambridge University Press, Cambridge, 2016.
69 In the International Criminal Court's very first prosecution, the case against the Congolese militia leader Thomas Lubanga, the use of so-called secondary evidence was heavily criticized. See Elena Baylis, "Outsourcing Investigations", *UCLA Journal of International Law and Foreign Affairs*, Vol. 14, No. 1, 2009.

what has been done meets the criteria for crimes against humanity, war crimes or genocide, which may make it eligible for international prosecution. Similarly, human rights organizations may call attention to and document instances of CRSV in close to real time. International criminal justice actors, on the other hand, are rarely first to the scene. Indeed, they are dependent on data from other actors – State and non-State – to successfully investigate and prosecute CRSV. These different temporal positions in the cycle of intervention matter for how and for what purpose data is collected, managed and passed on. As such, timelines not only create different types of digital bodies, but also shape the kind of work that digital bodies are put to.

As such, the three sectors work with and imagine different time horizons. Humanitarianism operates in the ongoing state of emergency, addressing the needs and suffering of the present to bring relief in the urgent future.[70] For humanitarians, the time horizon is short; in theory at least, what matters most is today and tomorrow. This makes data management part and parcel of the day-to-day governance of vulnerable populations. While it is possible to trace and establish responsibility for harms committed, human rights and international criminal justice operate on a much longer time frame. They seek to address individual and systemic injustices of the past in order to establish accountability, provide redress and prevent future violations of human rights and international law. The long-term goal of human rights and international criminal justice is the institution of a universal moral and judicial community. This being so, these temporal differences also reflect a difference in audience and, critically, in how these sectors view the agency of those they address.

Agency

A final issue thus concerns agency. While all three fields are cosmopolitan in orientation, meaning that they have a "philosophical, ethical, and scientific world view which aims to transcend national boundaries and a nationalist, state-centered outlook on society and justice",[71] their actions are strongly oriented towards States and donors, and their institutions and practices reflect donor priorities.[72] How will this change when tech actors enter the scene? For example, "[a]s social media increases in importance, so do the social media companies themselves. Facebook, YouTube, Twitter, etc. are all corporations, and they host the videos, photos, and reports that are posted to social media on privately-owned data

70 Although humanitarian actors have been present in some locations for several decades and are increasingly present in protracted conflicts, a state of emergency animates their intervention; see Didier Fassin and Mariella Pandolfi, *Contemporary States of Emergency: The Politics of Military and Humanitarian Interventions*, Zone Books, New York, 2010.

71 Cecilia M. Baillet and Katja Franko Aas (eds), *Cosmopolitan Justice and Its Discontents*, Routledge, London, 2011, p. 1.

72 S. Kendall and S. Nouwen, above note 61.

servers."[73] Moreover, while all three sectors push for professionalization that "favours 'technical' and generalizable knowledge of local communities",[74] these developments may also increase the distance between staff and victims/survivors. Questions must therefore be asked about how digital technologies factor into recurrent criticisms of justice that is distant and remote, rather than place-based, and the extent to which technological solutions facilitate or hinder survivors' participation. This also affects how we think about the freedom to engage. Scholars have increasingly begun to ask critical questions about the freedom not to engage with the data market or not to be represented on commercial databases.[75] In our context, this entails asking how much visibility survivors of CRSV owe the state, the international community or the aid sector, and whether their digital bodies should automatically be enlisted in the fight against impunity for sexual violence when technology facilitates the collection of evidence.

Conclusion

This article has offered a three-part analytical framework for investigating the possibilities and pitfalls of the ongoing digital transformation of the fight against CRSV. In particular, we have proposed the use of "digital bodies" as an analytical concept for facilitating cross-sectoral exchanges on power and responsibility in data governance.

Despite good intentions, technology does not always work as planned or intended. Inadequate problem definition may entail that technological solutions fail to respond to the real-life issues that they are set up to deal with. A widespread reason for flawed problem definition remains the fact that affected populations are often not present in innovation processes – they are neither properly consulted nor invited to participate.[76] We therefore suggest that the international community must pay further attention to the serious ethical and legal issues emerging from technological innovations within the aid sector: technology has the potential to produce new digital harms, whether these occur through (in)visibilizing the suffering of particular groups or individuals, creating undesirable consequences, or introducing new risks.

There is also a need to distinguish clearly between what technology does and does not see when addressing CRSV. We noted above that the dual focus on technology and sexual violence potentially generates a powerful moral economy, but that it is important to consider the extent to which one issue – sexual

73 Emma Irving and Jolana Makraiová, "Capture, Tweet, Repeat: Social Media and Power in International Criminal Justice", in Morten Bergsmo, Mark Klamberg, Kjersti Lohne and Christopher B. Mahony (eds), *Power in International Criminal Justice*, Torkel Opsahi Academic EPublisher, Brussels, 2020.

74 S. Kendall and S. Nouwen, above note 61.

75 Linnet Taylor, "What is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally", *Big Data & Society*, Vol. 4, No. 2, 2017.

76 Mirca Madianou, Liezel Longboan and Jonathan Corpus Ong, "Finding a Voice through Humanitarian Technologies? Communication Technologies and Participation in Disaster Recovery", *International Journal of Communication*, Vol. 9, No. 1, 2015.

violence – can crowd out other issues and framings of CRSV and related insecurities. At the same time, increased reliance on digital evidence may alter the types of crimes and victim subjectivities that get attention (and documentation), and may unintentionally contribute to the (re)silencing of sexual violence and CRSV in the face of more explicit criminal imagery – killings, for example.[77] It is also worth considering the extent to which digital technologies – and social media especially – exacerbate issues related to stigma, shame, and re-traumatization and secondary victimization. In short, digital technologies do not solve political and ethical problems.

Finally, we suggest that a deeply problematic aspect of the digital turn is the trend towards extraction of data in exchange for aid, protection and justice. From an emancipatory – and a feminist – perspective, this necessitates the inclusion of participation and investment in digital literacy as key aspects of a technologized approach to the struggle against CRSV. Digital literacy must be mainstreamed to include all actors, from survivors and field practitioners to stakeholders in humanitarian, human rights and international criminal justice responses. To that end, impact assessments, training and capacity-building around cultures of responsible digitalization are needed. Digital literacy can only be acquired when survivors and community members are empowered to build, trust and act on this capacity, and when donors and policy-makers are committed to investing time and resources. Participation is thus key to successful technologized interventions, and stakeholders should bear this in mind when deciding on their actions and interventions.

77 E. Irving and J. Makraiová, above note 73.

International Review of the Red Cross (2020), **102** (913), 117–143.
Digital technologies and war
doi:10.1017/S1816383121000102

IRRC

Media and compassion after digital war: Why digital media haven't transformed responses to human suffering in contemporary conflict

Andrew Hoskins*
Andrew Hoskins is Interdisciplinary Research Professor in the College of Social Sciences at the University of Glasgow. His books include *Digital Memory Studies: Media Pasts in Transition* (Routledge, 2018, ed.) and *Risk and Hyperconnectivity: Media and Memories of Neoliberalism* (Oxford University Press, 2016, with John Tulloch). He is founding Co-Editor-in-Chief of the *Journal of Digital War*, founding Editor-in-Chief of *the Journal of Memory Studies*, founding Editor of the new journal *Memory, Mind & Media*, and founding Co-Editor of the Routledge book series Media, War and Security. His latest book, *Radical War: Data, Attention*

* I am grateful to David Rieff and to Paul Slovic for their insights and for their generosity in speaking with me. I am also grateful to Shona Illingworth, Anthony Downey, Oliver Boyd-Barrett and the *Review* editorial team and anonymous reviewers for their detailed comments and advice. I am grateful for important feedback on early versions of this article presented at the 2016 VoxUkraine Conference, "The Power of Words: Responsibility of the Media and Challenges in 2017", in Kiev, and at the 2019 staff–student seminar series at the School of Social Policy, Sociology and Social Research, University of Kent.

and Control in the 21st Century (with Matthew Ford), will be published in 2022.

Abstract

There is a persistent belief in the power of media images to transform the events they depict. Yet despite the instant availability of billions of images of human suffering and death in the continuous and connective digital glare of social media, the catastrophes of contemporary wars, such as in Syria and Yemen, unfold relentlessly. There are repeated expressions of surprise by some in the West when the dissemination of images of suffering and wars, particularly in mainstream news media, does not translate into a de-escalation of conflict.

In this article I consider today's loosening of the often presumed relationship between media representation, knowledge and response under the conditions of "digital war". This is the digital disruption of the relationship between warfare and society in which all sides participate in the uploading and sharing of information on, and images and videos of, conflict.

Is it the case that the capacity of images of human injury and death to bring about change, and the expectation that they would stir practical intervention in wars, is and has been exaggerated? Even if we are moved or shocked upon being confronted by such images, does this translate into some form of action, individual or otherwise? In this article I contend that the saturation of information and images of human suffering and death in contemporary warfare has not ushered in a new era of "compassion fatigue". Rather, algorithmically charged outrage is a proxy for effects. It is easy to misconstrue the velocity of linking and liking and sharing as some kind of mass action or mass movement.

Humanitarian catastrophes slowly unfold in an age of continuous and connective digital glare, and yet they are unseen. If the imploded battlefield of digital war affording the most proximate and persistent view of human suffering and death in history cannot ultimately mobilize radically effective forms of public response, it is difficult to imagine what will.

Keywords: digital war, social media, images, human suffering, compassion, humanitarian crises, Syria, Yemen.

: : : : : : :

The inescapable untruth

Changes in our lens on the world – what is seen or not seen, how and for how long, through the transforming media environment of a given era – have shaped our capacity to be moved, to respond in some way, when confronted with human suffering as a result of conflict. There is also a history of experiences and assumptions as to the impact of the specific medium of representation on a given spectator or audience. This includes the photograph (and earlier public lantern

slide lectures) being used to mobilize humanitarian action since the nineteenth century.[1] But this history, I argue here, has been upended in the digital era, with the transformation in the relationship between war, media, perception and public response.

In this article, I set out the nature, consequences and paradoxes of this transformation and ask if some of the traditional and established concepts and ideas are up to the task of rendering it intelligible. This transformation notably includes a radical shift in the emergence of a new mass public that is able to express views and opinions in an immediate time frame – to participate and to provide feedback – and in ways that were unthinkable only two decades ago. But also intertwined with this shift is the development of a new kind of war – digital war – which is fought through and on the same media platforms and devices that have enabled mass public participation. Human suffering in conflicts is mostly unfolding in a continuous and connective digital glare which one might think would deliver a new era of humanitarianism, and yet which seems instead to have diminished the prospects of any kind of usable compassion.

The singular photograph was once acclaimed as a reliable output of a medium used for stirring the collective conscience, for radicalizing the crowd, for compelling action. This is not to say that there was a golden (twentieth-century) age of image effects. Rather, there is more like a false memory of the relationship between media images, knowledge and action, driven primarily through photojournalism. The persistent belief in the power of images does not fit with today's digital world – more specifically, debates about compassion, media images and the related prospects for intervention to stop so-called "distant" human suffering,[2] caused predominantly by humanitarian crises, fail to grasp the nature and scale of the transformation ushered in by digital war.

I define "digital war" as the ways in which multimedia smartphones, messaging apps and social media platforms have disrupted the relationship between warfare and society, creating a global, although uneven, participative

1 See Heide Fehrenbach and Davide Rodogno, "'A Horrific Photo of a Drowned Syrian Child': Humanitarian Photography and NGO Media Strategies in Historical Perspective", *International Review of the Red Cross*, Vol. 97, No. 900, 2015. In her article for this issue of the *Review*, entitled "The Camera and the Red Cross: 'Lamentable Pictures' and Conflict Photography Bring into Focus an International Movement, 1855–1865", Sonya de Laat explores the role of combat photography in three major conflicts of the mid-nineteenth century "in expanding a shared vision of who constitutes humanity, and who is worth caring for".

2 By "distant" I mean a predominantly Occidental and privileged position of safety from, and limited fear of, warfare. This idea of "distant suffering" has been associated with a "white saviour complex" narrative of humanitarian action, whereby the "West" goes in to help "distant" and "non-Western" parts of the world. This is also embedded in a history of the development of assumptions about the role of late twentieth-century media in war. But perceptions and meanings of "distance" are also transformed through digital war, a shift that I set out in this paper. For an influential work on this term, see Luc Boltanski, *Distant Suffering*, trans. Graham Burchell, Cambridge University Press, Cambridge, 1999. See also Lilie Chouliaraki's idea of the "post-humanitarian" style of communication in a "dominant Western culture where the de-emotionalization of the suffering of distant others goes hand in hand with the over-emotionalization of our safe everyday life": Lilie Chouliaraki, "Post-Humanitarianism: Humanitarian Communication Beyond a Politics of Pity", *International Journal of Cultural Studies*, Vol. 13, No. 2, 2010, p. 122.

arena in which it is decreasingly clear who is fighting, who is commentating, and who is experiencing the effects of war. This is taking place through a new siege of attention as militaries, civilians, journalists, NGOs, States and militias all upload and circulate images and video from sites of conflict, often in real time, producing an intensely competitive panoply of participation in and perspectives on war.

An astonishing consequence of digital war, in particular for anyone who ever had faith in the idea that images can really end conflict, is its production of the most proximate and persistent view of human suffering and death in history. The plethora of content on digital and social media platforms (YouTube, Facebook, Instagram and others) produces an overwhelming presence not only of unfolding war, but also of all the horrors of previously forgotten or mislaid wars, as a new archive of human suffering is suddenly made available.

The battlefield has taken on an open-access dimension, crowded with the claims, opinions and outrage of anyone who can post, link, like or share on social media. This has produced a seething mass of information, disinformation, misinformation and speculation about unfolding events, uploaded by a connected multitude. Digital war, then, is information war that weakens the distinction between combatants and civilians, with publics increasingly entangled in the digital sharing of information and images about conflict.

This live-streaming and sharing of battle then builds an incredible record, a searchable archive, of information and images of war on an unparalleled scale. A simple YouTube search for "Syrian war", for instance, brings up a seemingly infinite number of videos, where even the first ten hits range from being uploaded in the last few weeks to as long ago as several years, some with viewing figures in the millions.[3] Search results yield a seemingly unlimited mix of images and video of the mainstream remediated, the official, unofficial and authorized, the conventional and the transgressive, all awaiting their algorithmic return in an emergent, patchwork, living archive and history of war. This living archive is in constant flux – videos are continually being re-edited, renamed and reposted, by the various actors pushing their version or experience of this war. Meanwhile, the supreme gatekeeper of YouTube is accused of removing or deleting videos that do not conform to its terms of service, but which potentially document war crimes.[4] In this war of and on social media archives, there is a "radicalization of memory".[5]

3 Included in the range of this three-week- to seven-year-old spread of the top 10 search results (searching from a UK ISP on 29 January 2021) were a mix of mainstream media coverage and professional, amateur and unverifiable organizations supporting a particular side or group in the war. For example, a video with 2.3 million views entitled "Heavy Clashes during the Battle for Al-Ramouseh Aleppo | Syria War 2014" was uploaded by the "WarClashes" channel on 20 April 2014. The description reads, in part: "Heavy clashes erupted as various brigades attacked Al-Ramouseh district in Aleppo. The fighters managed to capture the district from the Syrian Army after fierce clashes went on for a couple days". See: www.youtube.com/watch?v=Iwh49HgfyME&t=10s (all internet references were accessed in January 2021).

4 See, for instance, the work of the Syrian Archive group, available at: https://syrianarchive.org.

5 Andrew Hoskins, "The Radicalisation of Memory: Monuments and Memorials in a Post-Trust Era", Keynote Talk, "Moving Monuments" Conference, Manchester Centre for Public History and Heritage, Manchester Metropolitan University, 20 April 2018.

For the purposes of my argument here, digital and social media afford a new experience of the in/visibility of war, where acts and images capturing those acts seem to be out there in the open. Yet, this seemingly transparent news of events is illusory. What is seen and not seen is dictated by the most effective wagers of information warfare and their methods of attention-hacking, exploiting the algorithms of content hosts and owners. Through posting and liking content, automated trolls give the impression of genuine popularity, tripping the algorithm of social media and video platforms to push content into the feeds of real people, who in turn repost it and propel it into a much wider media ecology. Thus, the non-human and the human work in consort to create an alternative reality – in essence, an in/visible digital war.[6]

In this article I explore the consequences of the emergence of digital war for the nature of, and assumptions about, the relationship between knowledge about war, compassion and the capacity for action, in the face of representations of distant human suffering. My principal point of reference for a paradigmatic shift in the relationship between war, media and the capacity for comprehension and engagement is the late twentieth-century "broadcast era". This is the era in which ideas about the relationship between representation, knowledge and response cohered – namely, a pervasive and persistent belief in the power of media and particularly television images to transform the events they depict. What follows from this assumption is that when images of human suffering no longer seemed to do the work that was expected of them, this was explicable by an over-familiarity with such images, breeding so-called "compassion fatigue".[7] I argue that this notion of compassion fatigue, as a way of considering today's response or lack of response to images of distant human suffering, does not translate into today's digital media ecology.

The shift I am describing stems from the nature and frequency of media representations of suffering. During the broadcast era, images of humanitarian crises and conflicts were largely shared by a select group of mainstream journalists, editors, and channel and newspaper owners. In effect, this was a mostly contained or closed system of communication, with actors having largely monopolistic roles in the production, publication and broadcast of images and footage of war. Furthermore, for several decades, media studies and other disciplines sought to work out exactly what "the audience" understood from "the text" that had been "produced" for them.

6 Matthew Ford and Andrew Hoskins, *Radical War: Data, Attention and Control in the 21st Century*, forthcoming.

7 Compassion fatigue refers to the idea that on being too frequently confronted with, for example, a news image of a child emaciated through starvation or maimed by urban shelling, the "distant" "spectator" not living through humanitarian crises will not be sufficiently moved or outraged to challenge the policy-makers or donate to the aid organizations that might intervene in order to limit or stop the suffering and deaths of civilians under attack. See Susan D. Moeller, *Compassion Fatigue: How the Media Sell Disease, Famine, War and Death*, Routledge, New York, 1999. For an influential critique of Moeller's work, see David Campbell, "The Myth of Compassion Fatigue", in Liam Kennedy and Caitlin Patrick (eds.), *The Violence of the Image: Photography and International Conflict*, Routledge, London, 2020.

With digital technologies and social media, however, anyone is potentially an information producer and sharer. Thus, as Merrin makes clear, the very nature of how audiences engage with media content has changed:

> [E]ven when watching broadcast content on a digital device the user is not the traditional audience: their status is not defined by their consumption but by the active relationships of communication and control producing and maintaining their activities and the electronic signals and records these create.[8]

Here, then, the traditional idea of "media production" is reversed. At one point in the history of broadcasting, audiences were seen as the end point of a linear, reductive notion, of the flow of communication of news and information about war and its consequences. Today, users are not just on the receiving end of news and images, with limited or no opportunity for feedback or for making their opinions public. Rather, they are participants in an ongoing and connected network, actively partaking in the production and distribution of media content

This may all initially sound democratizing amidst the benevolent language of "social" media, in which users "share" content, that peaked with the heralding of the 2011 Arab Spring. But although the advent of digital war has challenged the mainstream media and other elite actors in their capacity to shape what war looks like, they now compete with new forms of surveillance and control employed by the corporations that own and manage the platforms through which users produce what today is understood as war, and with all those who can exploit people's increasing dependency on such platforms and services for news and information.

Nonetheless, the very idea of user-generated content (UGC), of not only being a participant but feeling like a participant, and of being active in the capacity to express – record, produce, publish, share, like, post, edit, caption, retweet – media content about war and human suffering, affords a sense of personal connection, action and control. Are these forms of expression, then, new outlets for, and indeed amplifiers of, individual and public outrage and compassion on a new and unprecedented scale? Or rather, when images of suffering are acclaimed for "going viral", is it the case that the knowledge of this contagion functions to blunt as much as to inspire compassion? Are certain forms of digital "participation" substitutes for, rather than drivers of, action, when the work of outrage is seen as done?

Another view that I dissect in this article is that instead, digital war has created a new era of "compassion fatigue". A common explanation for an inability to feel is that familiarity with images of suffering inevitably desensitizes us and leaches meaning out of what is before us. Viewers are supposedly slowly dulled and numbed into jaded indifference by the sheer scale and persistence of the suffering of innocents on repeat, looped into social media platforms and news cycles that offer only war without end. The compassion fatigue theory posits that

8 William Merrin, "Fight for the Users! Media Studies in the 21st Century", *Media Education Research Journal*, Vol. 5, No. 2, 2015, p. 61.

if audiences turn away, then so does the news, for it is a business ultimately driven by clicks and subscriptions.

Yet, this popular view of overburdened attention blunting the potential mobilization of a response is misconstrued. The very idea of compassion fatigue as it applies to war and media[9] is underpinned by a misguided, persistent presumption that images have substantive effects *and* that there was a critical mass of those who really cared in the first place. In this paper, I probe the nature of the investment in this idea by policy-makers, journalists and academics. Specifically, I question whether the compassion fatigue hypothesis removes any impetus for relief or humanitarian support to be undertaken in a public or even urgent manner.

For instance, almost a decade of footage, images and feeds[10] of the suffering and deaths of civilians in Syria has seemingly been hiding in plain sight across global mainstream and social media, with no effective prompting of any coherent or effective collective expression of outrage. If perceived inaction over Syria's protracted war is not a classic case of compassion fatigue, then what explains the relationship between media representations of suffering, knowledge and response in the digital era?

My answer to this question is the basis for what follows in this article. I argue that this belief in the relationship between representation, knowledge and response has for years fed into an unrealistic image of media effects[11] and, relatedly, the role and potential of humanitarianism. This gap between belief and reality has significantly widened with recent transformations in the nature of and relationships between war and media, that I am here calling digital war. This includes the way in which digital media and technologies have disrupted war and society so that the very audiences that all actors (militaries, governments, NGOs) have traditionally attempted to reach via media are now part of the same communications fabric.

From the above, there are two key and related observations in terms of the individual confronted with images of human suffering in today's digital media ecology. The first is that digital war is inescapable. While human suffering in other parts of the world afflicted with humanitarian crises has often seemed like a distant concept to the so-called Occident,[12] in our current digital age, this figurative and geographic "distance" is collapsed by the personal proximity of social media.

9 See S. D. Moeller, above note 7, for her influential work in this context.

10 One acclaimed example is the photograph taken in 2019 by the journalist Bashar al-Sheikh. See "Syrian Air Strike Sisters Photo: Behind the Image that Shocked the World", *BBC News*, 1 August 2019, available at: www.bbc.co.uk/news/av/world-middle-east-49186145/syria-air-strike-sisters-photo-behind-the-image-that-shocked-the-world.

11 There is a long history of "direct" media effects research in media studies – this is the idea that exposure to or consumption of mass media shapes individuals' behaviour. David Gauntlett offers a useful critique of this work in arguing that "if, after over 60 years of a considerable amount of research effort, direct effects of media upon behaviour have not been clearly identified, then we should conclude that they are simply *not there to be found*". David Gauntlett, "Ten Things Wrong with the Media 'Effects' model", 2006, available at: https://davidgauntlett.com/wp-content/uploads/2018/04/Ten-Things-Wrong-2006-version.pdf.

12 Edward W. Said, *Orientalism*, Pantheon Books, New York, 1978.

Digital media not only afford a potentially continuous view of human suffering, but also offer unprecedented opportunities for individuals to like, share and comment on a tsunami of disturbing images. In digital war, which has for years delivered untrammelled images of the horrors of civilian suffering (for example, in the wars in Syria and Yemen), it is no longer possible to say "I did not know".

The second observation is that we are living through an astonishing collapse in trust in the mainstream news media. Informational production and distribution has suddenly undergone a very public crisis of legitimacy, with a recognition of the difficulty of dividing truth from opinion and doubts raised over who has the right to lay claim to an audience or to truth.[13] In these circumstances, when looking at or sharing messages, images, humanitarian campaigns and so on, and in assessing their truthfulness, it becomes easier to conclude that "I cannot know". The voracity, provenance and reliability of media content seem so difficult to ever pin down in this environment. This is, then, an incredible challenge for those attempting to convey information about human suffering in order to elicit intervention, some kind of help. This includes humanitarian organizations, charities and other NGOs, which are often ultimately reliant on the unique power of the still image, in whatever medium, to persuade the individual to donate.

These two aspects are paradoxically brought to bear on the moral sensibility of the participant in digital war, who is unable to deny knowledge of the world's atrocities unfolding in multiple and simultaneous news feeds, and yet finds it difficult to trust any one of them. Both of these aspects, being aware and being uncertain, follow from the fact that in the circumstances of digital war, our attention is under siege.

In sum, can compassion for suffering during humanitarian crises amidst the effects of digital war be translated into action? To answer this question requires acknowledgement of the paradigmatic shift in the relationship between media, knowledge and action. This is caught up in the paradox of the impossibility of being able to claim ignorance of human suffering in the face of seemingly unlimited information and images, against the difficulty of having certainty about the provenance of any of the same information and images.

Digital war is thriving in what is popularly known as today's "post-truth" society, where any given news about the world can be quickly and conveniently mired in a sea of alternative opinions, conspiracies, fakes and fact-checking. For instance, the Oxford Dictionary defines "post-truth" as "circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief".[14] Although news has always been made up of a mosaic of facts, opinion, conjecture and speculation, the difference today is that there is a brazen quality to misinformation in that it is hiding in plain sight, having become an immovable part of everyone's information diet.

13 Catherine Happer, Andrew Hoskins and William Merrin, "Weaponizing Reality: An Introduction to Trump's War on the Media", in C. Happer, A. Hoskins and W. Merrin (eds), *Trump's Media War*, Palgrave Macmillan, Basingstoke, 2018, pp. 7–8.

14 See: https://languages.oup.com/word-of-the-year/2016/.

Rather than "post-truth", the term "post-trust"[15] seems more appropriate to characterize a seismic fall of trust in what was once seen as a reliable mainstream news media, in that it called the attention of the "Occident" to humanitarian crises and unknown human suffering in "other" parts of the world. Rather, there is a new kind of in/visibility to war and its consequences through our digital practices. Our ability to snap, post, record, edit, like, link, forward and chat serves as a stand-in for trust, in that we feel active and liberated in engaging and shaping the media which we inhabit. This is "part of an epochal shift in how we gather and share information, in a movement to a 'me-dia' world in which individuals increasingly create their own, personal ecology of technologies, platforms, media, content, information, opinions, experiences and knowledge".[16] One consequence of our digital immersion in an ocean of media content, then, is not some kind of revelation but rather a closing down – and, some argue, a polarization.[17]

Considering these shifts, I set out in what follows some of the trends in, and the challenges of, attention, perception, compassion and war for all message shapers. To this end, I firstly expand on the historical context of the critical shift from the broadcast era to digital media ecologies that underpins my argument of revolutionary rather than evolutionary change. I then offer two case studies as lenses to explore these ideas further: the still unfolding wars in Syria and Yemen. I focus on these examples of human suffering because they exemplify a new cycle of image–expectation–inaction under conditions of digital war. Namely, there is a persistent belief that knowledge of this suffering through the unlimited supply of images and information should or will shape some kind of political and/or humanitarian intervention. Yet the suffering and death continues in the face of the expectation that the images and information about that suffering should or must be having an effect on the behaviour of the Occident. I conclude my exploration of this model by highlighting the paradox that the apparent publicness, "open access" and unlimited supply of images of human suffering in humanitarian crises is actually closing off the momentum for intervention – in essence, a reversal of the very idea of "media effects".

From the broadcast era to digital media ecologies

Saturation coverage in the broadcast era was influentially associated with a belief in a kind of shared, collective experience[18] that was unavoidably part of what it was to consume news media (if only given the lack of alternatives). This afforded a sense

15 Catherine Happer and Andrew Hoskins, "Hacking the Archive: Media, Memory, and History in the Post-Trust Era", in Michael Moss and David Thomas (eds), *Post Truth Archives*, Oxford University Press, Oxford, forthcoming.
16 C. Happer, A. Hoskins and W. Merrin, above note 13, pp. 14–15.
17 Tarek Abdelzaher *et al.*, "The Paradox of Information Access: Growing Isolation in the Age of Sharing", 2020, available at: https://arxiv.org/pdf/2004.01967.pdf.
18 Daniel Dayan and Elihu Katz, *Media Events: The Live Broadcasting of History*, Harvard University Press, Cambridge, MA, 1992.

that the "whole world was watching",[19] which shaped a belief in a powerful connected and collective conscience, an audience that could somehow feel or respond or act in consort. This sense of the power of media effects in the global yet contained media ecology of the late twentieth century has oddly continued to underpin assumptions about the visual force of the watching of catastrophes in the twenty-first century.

There is also a long history of the circulation of images depicting suffering during humanitarian crises, including with the aim to elicit sympathy and ultimately action. It wasn't until the broadcast era of the twentieth century, however, that there was a cemented belief, mainly across the Western hemisphere, in the power of the Western mainstream media (WMM)[20] to make an impact on events.

Late twentieth-century mainstream journalism thrived owing to its dominance in the representational order of the day, in which there was little opportunity to challenge what or how media content was produced and distributed. This was particularly the case with regard to the cultural hold enjoyed by television, a medium which has traditionally been referenced most by those advancing the idea of direct media effects and the existence of compassion fatigue. This elite representational order was sustained through what can now be seen as a scarcity of images of war and their control by mainstream media in terms of what was accessible to most global audiences – but today this order has been broken or at least displaced.[21]

This dominance of the late twentieth-century WMM hit a new peak with the satellite news coverage of the 1991 Gulf War.[22] This included the heralding of the so-called "CNN effect".[23] The news network's extended real-time coverage of the 1991 Gulf War established its reputation for being the principal and often exclusive provider of what was at the time a new kind of compelling global public

19 To borrow from Todd Gitlin's classic text: see Todd Gitlin, *The Whole World is Watching: Mass Media in the Making and Unmaking of the New Left*, University of California Press, Berkeley, CA, and London, 1980.

20 Hereafter, I use "WMM" to refer to the Western news media in order to signify that historically there has been a concentration of news and other media ownership and/or production in rich Western countries (some also use the term "global North"). This includes, for example, the BBC, CNN, News Corp, Reuters and the picture agencies. This focus is pertinent to my key delineation of news cultures between the late twentieth-century "broadcast era" and today's digital media ecology. An important consideration is that the WMM were utterly defining of what "Western war" was (see Martin Shaw, *The Western Way of War: Risk-Transfer War and Its Crisis in Iraq*, Polity Press, Cambridge, 2005). This includes CNN being mutually synonymous with the 1991 Gulf War, but importantly, this relationship between the WMM and warfare also influenced a significant trajectory of thought in the analysis of war and media from the early 1990s (see below). In the digital media ecology, the WMM remains influential but competes with (as well as being incorporated by) the US-based so-called "tech giants" of Facebook, Twitter, Google and Apple, dominating how news and information is produced, seen or not seen.

21 See C. Happer, A. Hoskins and W. Merrin, above note 13, pp. 3–22. See also the refreshingly honest reflections of Jeff Jarvis over the past few years in his *BuzzMachine* blog, available at: https://buzzmachine.com.

22 Andrew Hoskins, *Televising War: From Vietnam to Iraq*, Continuum, London, 2004.

23 Steven Livingston, *Clarifying the CNN Effect: An Examination of Media Effects According to Type of Military Intervention*, Research Paper R-18, Joan Shorenstein Barone Centre on the Press, Politics and Public Policy, John F. Kennedy School of Government, Harvard University, 1997.

visual perspective on events unfolding across the multiple geopolitical fronts of this war.

At the time, there was a sense that TV news held a new real-time power to shape events as they occurred.[24] For example, on 17 January 1991, the day after the first Coalition bombing of Baghdad, the then Israeli deputy foreign minister Benjamin Netanyahu stated:

> What we are facing now is political communication. As we speak it may be that in a bunker in Baghdad, they listen to us. In fact, I'll delete the "may be" – I'm sure they listen to us. They are listening to us in Moscow, in Washington and everywhere else. So that the impact of what is seen and said on television is an integral part of the – of the war effort on both sides… television is no longer a spectator.
>
> You know, in physics – in subatomic physics – there's something called the Heisenburg Principle, which basically says that if you observe a phenomenon you actually change it. Well, now we have the Heisenburg Principle of politics: If you observe a phenomenon with television instantaneously you modify it.[25]

The significance of the perception of the global real-time power of this shift should not be underestimated. The media depiction of the 1991 Gulf War spawned a whole new subfield in the study of the relationship between war and media. This included the defining empirical study, written in the same decade by the highly respected journalist Nik Gowing, on the impact of real-time news reporting of conflicts and humanitarian crises on political influencers.[26] Gowing concluded that only in exceptional circumstances did real-time reporting change the policy of presidents and governments, although it more often affected their rhetoric.[27] Despite this work, a belief in the CNN effect and in media effects more widely has persisted into the digital era. This is evident in my core argument here of the persistent expression of surprise in this century at the lack of intervention in the face of WMM reporting of suffering caused by humanitarian crises.

Before the CNN effect took hold of the Western conscience, a harrowing BBC video by the Kenyan cameraman Mohammed Amin and the UK journalist Michael Buerk was released.[28] Covering the famine in Ethiopia, which killed over

24 See, for example, McKenzie Wark, *Virtual Geography: Living with Global Media Events*, Indiana University Press, Bloomington, IN, 1994

25 *Larry King Live*, CNN, 17 January 1991, cited in Andrew Hoskins and Ben O'Loughlin, *Television and Terror: Conflicting Times and the Crisis of News Discourse*, Palgrave Macmillan, Basingstoke, 2007, pp. 38–39. See also Deirdre Boden and Andrew Hoskins, "Time, Space and Television", paper presented at the 2nd Theory, Culture & Society Conference, "Culture and Identity: City, Nation, World", Berlin, 11 August 1995.

26 Nik Gowing, *Real-Time Television Coverage of Armed Conflicts and Diplomatic Crises: Does it Pressure or Distort Foreign Policy Decisions*?, Working Paper 94-1, Joan Shorenstein Barone Centre on the Press, Politics and Public Policy, John F. Kennedy School of Government, Harvard University, 1994.

27 See A. Hoskins and B. O'Loughlin, above note 25.

28 The original report was first broadcast on BBC1 on 23 October 1984.

a million people, the report reached international audiences in October 1984. It propelled the view that images of suffering during humanitarian crises could mobilize mass public philanthropy in the advanced industrial world, on a scale that seemed to match that of the crisis. The coverage depicted starving people arriving at feeding stations in the north of the country and reached unprecedented audiences for the pre-satellite news era.

This footage was seen by almost a third of the adult population in the UK and was rebroadcasted by hundreds of stations around the world, including the NBC evening news in the United States.[29] The film is credited with kick-starting the Live Aid phenomenon[30] and an outpouring of vast institutional and public sums in aid, including an immediate £1.8 million emergency EEC donation within two days of its first airing on the BBC1 lunchtime, 6pm and 9pm news bulletins.[31] Unlike today, when we are continuously digitally connected, in the 1980s it was unusual for a news story to go viral. This exceptional story bucked the constraints of the media ecology of the time.

The video was also exceptional in its simplicity. Drought was a key part of the televisual and UK government explanation for the famine, but a key cause of the famine was a civil war between Mengistu Haile Mariam's Ethiopian regime and Tigrayan and Eritrean insurgents. Both sides were using food supply chains and humanitarian assistance as weapons. What became known as the "Great Famine" of 1983–85 was exacerbated through conflict, with "clear evidence" that the Derg (the ruling military junta in Ethiopia at the time) had employed food as a weapon of war.[32] But these complexities of war, which required messy and protracted solutions, were overridden by the belief in the images of humanitarian crisis delivering a successful campaign to raise funding for humanitarian efforts. Ultimately, it was the simplicity of the drought as the principal cause of the famine and the suffering, and the apparently rapid solution of providing humanitarian assistance, that was an easy sell to those who donated funds, rather than a focus on the intractable complexities of the civil war.

The broadcasts of the famine in Ethiopia became an incredibly powerful template for understanding the influence of the WMM on public consciousness. They were not only depicting disturbing images, but were also set in a developing country that was previously not receiving much coverage in the WMM – in this case, the coverage ran for over six minutes on BBC news programmes.[33] Notably, this coverage took place at a time when there was a real scarcity of images, particularly of human suffering. News bulletins in the mid-1980s had primetime but nonetheless very limited slots in the television schedule: no matter what

29 Suzanne Franks, *Reporting Disasters: Famine, Aid, Politics and the Media*, C. Hurst & Co., London, 2013, p. 1.
30 A. Hoskins and B. O'Loughlin, above note 25, p. 131.
31 *Ibid.*, p. 131.
32 Edmond J. Keller, "Drought, War, and the Politics of Famine in Ethiopia and Eritrea", *Journal of Modern African Studies*, Vol. 30, No. 4, 1992, p. 623.
33 S. Franks, above note 29.

happened in the world, audiences had a limited and predetermined amount of time during which they could watch the news.[34]

After the Ethiopia coverage and its success in raising funds for humanitarian relief efforts, "foreign relief NGOs grew at an unprecedented rate and their relationship with the media was radically and permanently altered".[35] It is this belief in the relationship between media image, knowledge and the expectation of action that has dulled audience and, relatedly, political responses to wars and catastrophes in "other" parts of the world ever since. There is a belief upon being confronted with images of human suffering of sufficient severity or scale that somehow, somebody is responding or intervening. This is notwithstanding a feeling of helplessness when confronted with a scenario in which an individual thinks that their own response or action will not make any difference.

Rather than there being an accumulative effect of compassion fatigue, there is a belief in the power of the visual image, amplified in the late twentieth-century by news media, audiences, policy-makers and NGOs, which seems to persist in the face of any evidence to the contrary.[36] For some, the result is a kind of collusion between audience and image, so that mass media coverage in itself is enough to reassure the viewer that something will be done. This could also be characterized as a form of shared denial. This is the view of Roger Silverstone, who, drawing on the work of Stan Cohen, states:

> Just as families can deny the presence of an alcoholic member, because it would be too painful to acknowledge, so too can societies deny the presence of problems and traumas that they would otherwise have to confront. Media images enable a collusive illusion that the appearance of the other in crisis on the screen is sufficient for us to believe that we are fully engaged with him or her in that crisis.[37]

We are thus prone to presuming that the more catastrophic and the more widely mediated the image, the greater the likelihood that somebody somewhere is acting upon that knowledge.

This process has intensified in the digital era with the image being carried by new kinds of what I call "networked contagion". This is the potential of any image to be shared and seen almost instantaneously by billions of social media users across multiple platforms, ultimately without any global force (i.e., law or moderation system) to stop it, no matter the nature of its content, real or fake, benign or terrorizing.

At the same time, there is increasing surprise and alarm from those who realize that the images of humanitarian crises that they are confronted with and surrounded by are having little palpable effect. Thus, a new polarized or

34 It was not until the 1991 Gulf War, seven years later, that the 24-hour news cycle would take hold.
35 S. Franks, above note 29, pp. 2–3.
36 See also W. J. T. Mitchell, "There Are No Visual Media", *Journal of Visual Culture*, Vol. 4, No. 2, 2005.
37 Roger Silverstone, *Media and Morality: On the Rise of the Mediapolis*, Polity Press, Cambridge, 2007, p. 131.

paradoxical state emerges. While one camp of people believes that others have already taken action because these images are widely circulated across digital media, there is a second camp that cannot understand why people have not taken action, having been exposed to the same images. Both positions are still driven by the flawed expectation that images should or must have substantive effects.

Following this, a new vision (and a new history) of visual images is needed to break the contemporary stasis of a dislocated (i.e., from a different era) expectation of what seeing – and sharing – the visual image means in the digital era. Indeed, there is much important work (notably inspired by Harun Farocki[38]) which claims that images are "operational", in that they "are produced by machines to be seen by other machines, rather than the corporeal, embodied eye".[39] Somewhere in between there is the algorithmically influenced vision whereby what is seen and not seen of war and its consequences is determined by the images and stories that any individual has clicked on before. The counter to the idea of networked contagion of the freely contagious image on today's social media platforms is the echo chamber or "filter bubble"[40] – this is the human at the centre of their own "me-dia" ecology (see above).

Whether people have agency over what they see or not, the sheer scale and accessibility of media on armed conflict today is such that compassion is difficult to focus. This is not just a matter of fatigue as a result of being overloaded with social media and media images of the catastrophe, but a new, continuous state of a shared, nagging complicity in being indifferent to all of that which is always just a few clicks away.

The digital sub-conscience

A pivotal difference between the coverage of twentieth-century humanitarian crises and those of the twenty-first century is that now the incomprehensibility of the suffering and scale of war is matched by a new incomprehensibility of the scale and connectivity of information, images and commentary on war. For example, the enormity of the Syrian war, with hundreds of thousands of civilians killed and injured and millions displaced, is made synchronous with digital media. The broadcast era appeared to at least afford the Occident a collective vision of the world, a mono-global version of events so typified by the satellite television coverage of the 1991 Gulf War.[41] In contrast, the digital era shatters the ability of any one side to monopolize media. When suffering is live-streamed by all sides, the battlefield *is* the digital media ecology.

38 Harun Farocki, "Phantom Images", *Public*, No. 29, 2004.
39 Trevor Paglen and Anthony Downey, "Algorithmic Anxieties: Trevor Paglen in Conversation with Anthony Downey", *Digital War*, Vol. 1, No. 1, 2020, available at: https://doi.org/10.1057/s42984-020-00001-2.
40 Eli Pariser, *The Filter Bubble*, Penguin Books, London, 2012.
41 A. Hoskins, above note 22.

While audiences usually feel overwhelmed and incapacitated by new communication mediums,[42] this is particularly the case now because of the Internet's rapid evolution and unprecedented scale of expansion. As Bratton succinctly puts it, social media "boasts human history's single most prolific consolidation of images".[43] This includes the consolidation of images depicting war and human suffering.

It is nonetheless striking to consider that, from a pre-digital era perspective, the idea of an instantly available and unlimited supply of images of civilians suffering from displacement, starvation, violence and death because of war and other humanitarian crises would have appeared to be potentially the greatest communications asset at the disposal of humanitarian organizations. Relatedly, in a broadcast era news environment, where in countries such as the UK, there were two television news bulletins on two channels per weekday evening, the very idea of rolling 24-hour news would have appeared to have offered a new panacea for knowing the world out there. And this is also why rolling news coverage of the 1991 Gulf War, as I have detailed above, was so defining. These differences highlight the fallacy of the connection presumed in the relationship between knowledge and action, which underpins ideas about representations of suffering in humanitarian crises and their ability to propel people into action.

There has been an important shift from how the mass audience of the twentieth century relied on mainstream news media to mitigate a widespread ignorance of the plight of "distant others", to today's continuous digital sub-conscience with regard to the world's unfolding atrocities, only ever a few clicks away. The digital sub-conscience is a kind of partial awareness and a pushing to one side of the horrors of civilian suffering amidst the ephemeral but persistent flow of images in content feeds, which ultimately creates the state of not being able to say "I did not know".

In what I have called today's "post-scarcity"[44] culture of access to and availability of images and content capturing humanitarian crises, the soup of media content brings forth new limitations. There is too much content potentially available at a touch, a tap, a flick, a swipe or a spoken command, so that individuals are continually reminded of their ignoring of the world out there. As Luciano Floridi argues, "we are witnessing a substantial erosion of the *right to ignore*".[45] Floridi continues:

42 See Marshall McLuhan, *Understanding Media: The Extensions of Man*, Routledge and Kegan Paul, London, 1964.

43 Benjamin H. Bratton, *The Stack: On Software and* Sovereignty, MIT Press, Cambridge, MA, 2015, p. 127.

44 I use the term "post-scarcity" to indicate the abundance, pervasiveness and accessibility of communication networks, nodes and digital media content in this century compared with the "scarcity" culture of media content in the late twentieth century. See Andrew Hoskins, "7/7 and Connective Memory: Interactional Trajectories of Remembering in Post-Scarcity Culture", *Memory Studies*, Vol. 4, No. 3, 2011; Andrew Hoskins and John Tulloch, *Risk and Hyperconnectivity: Media and Memories of Neoliberalism*, Oxford University Press, Oxford, 2016.

45 Luciano Floridi, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*, Oxford University Press, Oxford, 2014, p. 42.

> The more any bit of information is just an easy click away, the less we shall be forgiven for not checking it. Information and communication technologies (ICTs) are making humanity increasingly responsible, morally speaking, for the way the world is, will be, and should be.[46]

The "we" that Floridi refers to here includes the principal producers of the news media, who are also confounded by their increasing impotence.

A key aspect of digital war, as I have set out, is participation in uploading and sharing of images and videos of humanitarian crises and conflicts. This is how new "architectures of participation"[47] offered by Web 2.0 platforms and connected and mobile media devices enable a wide range of actors (militaries, States, journalists, NGOs, citizens, victims) to have their say. At the same time, the apparently liberating and limitless uploading and sharing of the horrors of conflicts and humanitarian crises across digital worlds results in a false analysis of public responses. Specifically, using social media engagements, such as likes and shares, as barometers for action against some of the negative impact of and suffering caused by conflicts and humanitarian crises is misleading. The ease of participation in digital outrage lends itself to the notion that it has some form of critical mass, or that the digital response is an all-of-society action. Participatory outrage, the seemingly lively *act* of sharing in the horror, revulsion or compassion, through retweeting or liking or commenting on the image of the suffering other, paradoxically blunts the rage – the work of making a stand is seen as done.

In the broadcast era, individual opinions on and responses to events had very limited avenues for direct expression in the mass media of the day. To a large extent, the views of audiences were only ever marginally present in any given media ecology, effectively represented and managed by the gatekeepers of the WMM. Public opinion, as such, was a manifestation of those who owned and edited the print and broadcast media. Today, by contrast, wars are fought in an imploded battlespace in which millions of images and videos are uploaded from the field and are subject to instant global comment, sharing, linking and liking. Social media platforms have facilitated a new regime of quantification from which popularity, horror or outrage can be instantly and continuously read, and whose visibility is built into the architecture and is essential to the very character of a given platform or app.

One photograph which attracted just such attention, along with claims as to the power of its contagion and spread to shape political policy and action, was that of the toddler Alan Kurdi.[48] The body of the 3-year-old from Kobane was found washed up on a Turkish beach in September 2015, as his family fled war-torn Syria.

46 *Ibid.*, pp. 42–43.

47 Tim O'Reilly, "The Architecture of Participation", June 2004, available at: https://perma.cc/M7TH-EVBN?type=image.

48 See, for example, "100 Photos: Alan Kurdi", *Time*, available at: http://100photos.time.com/photos/nilufer-demir-alan-kurdi.

Media coverage of the refugee crisis tended to downplay the impacts of European policy and, in particular, the risk that criminalizing irregular migration forces people into ever more unsafe means of attempting to reach their destination country. This fact was mostly overlooked in the reporting of the story of those attempting to escape war and the refugee crisis being seen to engulf much of Europe.

For the purposes of this article, it is important to note that this image became instantly iconic as it seized the world's front pages, the outrage of social media and at least the momentary rhetoric of Europe's politicians. Public sympathy appeared to translate into public empathy, with a spike in donations to refugee aid charities.[49] The *Independent* ran a story with the headline "Aylan [*sic*] Kurdi Images Were Seen by '20 Million People in 12 Hours'", which cited Claire Wardle, co-author of a report on the impact of the image:

> The photo of Aylan [*sic*] Kurdi galvanised the public in a way that hours of broadcasts and thousands of column inches wasn't able to do. It has created a frame through which subsequent coverage has been positioned and compared.[50]

This claim begs the question: what difference did the "galvanised" public make in diminishing the actual causes of Alan Kurdi's death?

There are echoes here of Nik Gowing's aforementioned conclusions as to real-time television news reporting's effects on language rather than policy decisions and outcomes in the 1990s. The actual report to which the *Independent* story refers is an academic study that explores the nature and impact of social media circulation of images of Alan Kurdi, but which also tracks the story in mainstream media and offers a variety of perspectives on its impact.[51] This "rapid research response" report is motivated to comprehend the incredible mediatized outcry over the image of the dead Syrian boy, but its conclusions as to any systematic or substantive effects on policy decisions, and on the war that gave rise to the European refugee crisis of 2015, are at best ambivalent. For instance, Anne Burns, one of the contributors to the report, states that

> political figures joined in with the wider discussion about the Kurdi photograph, as not to do so appeared politically unacceptable, but … their responses were framed in relation to the public outcry, rather than an accurate reflection of their own shift in attitudes.[52]

49 Jamie Merrill, "Refugee Aid Charities See Surge in Donations after Image of Drowned Syrian Toddler Aylan Kurdi moves the nation", *The Independent*, 3 September 2015, available at: www.independent.co.uk/news/uk/home-news/refugee-aid-charities-see-surge-in-donations-after-image-of-drowned-syrian-toddler-aylan-kurdi-moves-10484953.html.

50 Adam Withnall, "Aylan Kurdi Images Were Seen by '20 Million People in 12 Hours'", *The Independent*, 15 December 2015, available at: www.independent.co.uk/news/world/europe/refugee-crisis-aylan-kurdi-images-were-seen-by-20-million-people-in-12-hours-new-study-suggests-a6774181.html. Note that Alan Kurdi was incorrectly referred to as "Aylan" in some early reporting on the incident.

51 Farida Vis and Olga Goriunova (eds), *The Iconic Image on Social Media: A Rapid Research Response to the Death of Aylan Kurdi*, Visual Social Media Lab, University of Sheffield, 2015, available at: https://tinyurl.com/1r9cqlbm.

52 *Ibid.*, p. 39.

The outrage here, then, is a containment of, rather than a conduit for, effects. Again, the ghost of Gowing's conclusions on the real-time effects of 1990s television news hovers over claims to the power of the digital visual image today. The Kurdi image and its effects are an important case, because they are viewed as an exemplar of the power of visual and social as well as mainstream media to at least appear to make a difference.

The Syrian war and the fractaling of attention

I now turn to consider in more detail the case of the Syrian war as exemplifying the conundrum at the heart of my thesis of compassion after digital war: namely, that the mass availability of images of the suffering and the dead in Syria (over the years from the initial protests in the country in March 2011) should have – through public pressure on policy-makers or in support of action – led to military intervention to end the mass suffering but failed to do so, and that this perceived failure of images to provoke such a change was persistently met with expressions of surprise.

Rather, the very existence of the flux of images gives the impression that much of contemporary war and its consequences appears to happen in the open, and is seemingly so thoroughly and continuously mediated that it will also catch the attention of those with the capacity to intervene. The image–expectation–inaction cycle, then, appears ever inflationary.

What might be a principal WMM measure or "news value" of the catastrophe of this ongoing war, namely the numbers of civilian and combatant deaths, was not exactly hidden. For example, in 2016, the then UN Special Envoy stated that around 400,000 people had lost their lives in the conflict. There has not been a more recent statement from an apolitical actor on the death toll.[53] Furthermore, more than half of the entire Syrian population has been displaced since 2011, with 5.6 million registered refugees.[54] As of October 2020, there are approximately 6.7 million displaced people in Syria.[55] The UN Office for the Coordination of Humanitarian Affairs (OCHA) estimates that (as of December 2020) 11.06 million people are in need of humanitarian assistance,[56] and the Office of the UN High Commissioner for Refugees (UNHCR) estimates that 4.65 million are in acute need.[57]

The WMM reporting of civilian casualties of the Syrian war was often based upon figures provided by the Office of the UN High Commissioner for Human

53 UN, "Note to Correspondents: Transcript of Press Stakeout by United Nations Special Envoy for Syria, Mr. Staffan de Mistura", Geneva, 22 April 2016, available at: www.un.org/sg/en/content/sg/note-correspondents/2016-04-22/note-correspondents-transcript-press-stakeout-united.

54 See the Office of the UN High Commissioner for Refugees (UNHCR) Operational Portal, available at: https://data2.unhcr.org/en/situations/syria.

55 UNHCR, "Syria", fact sheet, October 2020, available at: www.unhcr.org/sy/wp-content/uploads/sites/3/2020/12/Factsheet-Syria-October-2020-003.pdf.

56 OCHA, *Humanitarian Response Plan: Syrian Arab Republic*, December 2020, p. 7, available at: www.humanitarianresponse.info/sites/www.humanitarianresponse.info/files/documents/files/2020_syria_humanitarian_response_plan.pdf.

57 UNHCR, above note 55.

Rights (UN Human Rights). However, a sign that both the sheer scale and the complexities of the war had overwhelmed the sense-making mechanisms of the West was that the UN, when providing its last estimate of the death toll in 2016 (relying in part on 2014 data), "said that it was virtually impossible to verify how many had died" as the conflict intensified.[58] But even when numbers of deaths were clearly quantified and there was ample digital evidence, as with the emergence of a digital archive of images that were smuggled out of Syria cataloguing ill-treatment and execution on a mass scale, there was not a major outcry or a pause in the deepening of the civil war.

Indeed, all of this barely seemed to register on the WMM conscience. For instance, Fred Ritchin, writing in the same month as the release of the Caesar images,[59] asked:

> [S]houldn't the existence of 55,000 photographs by the Syrian military police documenting the deaths of some 11,000 detainees who had been executed, and in many cases tortured, still provoke widespread condemnation, particularly in a conflict that is ongoing where millions are still at risk?[60]

The answer, Ritchin suggested, lies in a fraying of a "social contract" around the production and circulation of documentary photographs which was based upon "the willingness of the viewer to accept the photograph's reality as well as its invitation to a response".[61] Habituation to mass horror by publics and by governments is also part of Ritchin's explanation, and relatedly, part of digital overload, in that "we are now also aware that in the United States alone we regularly take more than 55,000 photographs every 15 seconds".[62]

This idea of the volume or scale of the medium being related to the human capacity to process and comprehend its meaning is influential. There is long-standing work which claims that responses to being confronted with representations of suffering or death are shaped by the scale of the catastrophe. For example, Robert J. Lifton coined the phrase "psychic numbing" in the aftermath of the atomic bombings in Japan, Hiroshima and Nagasaki to describe how survivors (known as *hibakusha*) had to turn off their feelings in order to function, amidst the human wreckage left by the bombs.[63]

58 Megan Specia, "How Syria's Death Toll is Lost in the Fog of War", *New York Times*, 13 April 2018, available at: www.nytimes.com/2018/04/13/world/middleeast/syria-death-toll.html. The statement was made by the then Special Envoy: see UN, above note 53.

59 The Caesar images are a collection of 55,000 photographs alleging to show proof of torture or ill-treatment by Syrian government forces. See Garance le Caisne, "'They Were Torturing to Kill': Inside Syria's Death Machine", *The Guardian*, 1 October 2015, available at: www.theguardian.com/world/2015/oct/01/they-were-torturing-to-kill-inside-syrias-death-machine-caesar.

60 Fred Ritchin, "Syrian Torture Archive: When Photographs of Atrocities Don't Shock", *Time*, 28 January 2014, available at: http://time.com/3426427/syrian-torture-archive-when-photographs-of-atrocities-dont-shock/.

61 *Ibid.*

62 *Ibid.*

63 Robert J. Lifton, *Death in Life: Survivors of Hiroshima*, University of North Carolina Press, Chapel Hill, NC, 1991. See also Paul Slovic, "Psychic Numbing and Genocide", *Judgement and Decision Making*, Vol. 2, No. 2, 2007.

Furthermore, a digital archive of atrocity, no matter how vast, has a weightlessness, an ephemerality, that the physical photograph does not have. The horrors of the Syrian war are well-documented in images. Archives of photographs documenting torture by State forces are readily available to the public, and non-State armed groups, including the so-called Islamic State, have released graphic videos of, among other things, the execution of prisoners and hostages.

The idea of the existence of 55,000 photographs of tortured and executed Syrian detainees, then, does not offer much hope for attaining any kind of singular focus. For example, the psychologist Paul Slovic said of the Caesar images:

> We find in our studies that attention begins to get dispersed at two objects. You can't focus as hard and draw as much information from two things you're looking at, as from one. So, what about 55,000? It breaks the system. Our simple system of senses and feelings can't handle it.[64]

With the flux of images of digital war, is it then possible to arrest their movement and mass, to fix the gaze as though we were returned to a pre-digital scarcity culture, before attention was shattered? To make the 55,000 images meaningful required attending to the resistances of both the ephemerality or intangibility of the digital archive, and its scale. To this end, fifteen UN member States sponsored a graphic exhibition, entitled "Caesar Photos: Inside Syrian Authorities' Prisons", of thirty of the photographs, which was shown at the UN headquarters in New York for ten days in March 2015.[65] Enlarged photos were displayed on easels stacked next to each other in a long line down a room at the UN, along with their printed captions. Most mainstream news reporting gave the numbers and told the story of the abuse and the archive, but it largely sanitized the images by blacking out any graphic bodily injury on the page or screen, or by showing the exhibition hall with the photos blurred in the background or at a distance that placed the gruesome details of the injuries out of clear view.[66] The images were accompanied in the exhibition with a sign that read: "WARNING. The following images are disturbing."

Instead of showing the full horror of the Syrian photographs, many news outlets showed the horror as reflected in the expression of a woman viewing the exhibits, in a photograph by Lucas Jackson for Reuters.[67] The unidentified woman holds a scarf over her nose and mouth, clenching her right fist, as though

64 Paul Slovic, telephone interview with author, 12 April 2018, on file with author.

65 See Raya Jalabi, "Images of Syrian Torture on Display at UN: 'It Is Imperative We Do Not Look Away'", *The Guardian*, 11 March 2015, available at: www.theguardian.com/world/2015/mar/11/images-syrian-torture-shock-new-yorkers-united-nations.

66 See, for example, Stav Ziv, "A Plea for Action: Gruesome Photos Smuggled From Syria on Display at U. N.", *Newsweek*, 13 March 2015, available at: www.newsweek.com/plea-action-gruesome-photos-smuggled-syria-display-un-313766; R. Jalabi, above note 65.

67 See Ian Black, "Syrian Regime Document Trove Shows Evidence of 'Industrial Scale' Killing of Detainees", *The Guardian*, 21 January 2014, available at: www.theguardian.com/world/2014/jan/20/evidence-industrial-scale-killing-syria-war-crimes.

trying to contain in her grasp the shock of what befalls her eyes.[68] This image draws our attention away from the story and onto one object, translating the unimaginable scale of suffering into a single human shape. In Slovic's terms (see above), the photograph transforms the digital abstraction of 55,000 images of 11,000 brutally executed people into something more intelligible; these are the "fallen images" which have been resurrected in the one. Has the digital splintering of attention thus been arrested, as though the viewer has been returned to the gaze more typical of an earlier media ecology? That is to say, would this be a less distracted gaze which was afforded time to hold and to dwell?

Jackson's photograph and its use here offer insight into Susan Sontag's view of the relationship between the potential for an image or images' mobilization of opposition to war and "the length of time one is obliged to look, to feel".[69] The moment at which the viewer is confronted with the image, the moment of looking, the moment of shock, is captured and thus extended. At the same time, the photograph offers an ideal, standing in for a moral multitude, all looking on and seeing or expecting their own revulsion in the same moment. There is comfort in the belief in this ideal of a shared recognition of what inhumanity looks like. In this way, the new mediator of shock both shows and hides. The photograph of the shocked response disturbs anew and yet offers relief from our having to see that which provokes the response. It does arrest the attention implosion of the digitally captured scale of the 55,000 images, but its mainstream appeal (it was widely used across the WMM) is achieved through a sanitizing displacement of confrontation with the images themselves.

There was, however, a notable exception to the WMM's limited publication, if not sanitization, of these images. In January 2014, the *New York Times* ran an article on the Syrian archive with four photographs at the very top of the page, depicting parts of emaciated bodies with arrows pointing to wounds including ulcerations and "tramline injuries" (linear bruises said to be consistent with being beaten with a rod).[70] Perhaps this forensic and partial depiction of the bodies is less disturbing to the reader than if the photographs on display at the UN, and others in the archive, had been reproduced in full.

My point is that, despite all of this, the Syrian war rolled on.

Viewers and readers continue to be confronted with a constant glut of information on the unimaginable scale of the human suffering in Syria. Those charged with raising awareness of the consequences of war – governmental, humanitarian and mainstream media organizations – all attempt to distil the numbers and the sources into an intelligible and ultimately digestible form. Yet,

68 *Ibid.*
69 Susan Sontag, *Regarding the Pain of Others*, Farrar, Straus and Giroux, New York, 2003, p. 122.
70 See Ben Hubbard and David D. Kirkpatrick, "Photo Archive Is Said to Show Widespread Torture in Syria", *New York Times*, 21 January 2014, available at: www.nytimes.com/2014/01/22/world/middleeast/photo-archive-is-said-to-show-widespread-torture-in-syria.html.

the very knowledge of the limitless online supply of images and claims of seemingly limitless catastrophe shapes a new kind of digital stasis of response.[71]

This stasis is particularly marked in the premediation of suffering, the persistent unheard warnings of impending disaster that are a characteristic of twenty-first-century conflict. I now turn to further address this idea in relation to both Syria and the famine in Yemen, as cases of human suffering in the era of digital war.

The premediation of catastrophe

My consideration of the Yemeni war is that it could be seen as a classic case of suffering caused by a humanitarian crisis, which has elicited repeated expressions of surprise by some in the West when its images of suffering do not translate into their anticipated effects. In particular, it is notable for the persistent scale of the suffering of its civilian population, as warned about and reported over several years.

Like Syria, Yemen is the scene of an ongoing conflict involving both States and non-State armed groups.[72] A UN Human Rights report in August 2018 suggested that "individuals in the Government of Yemen and the coalition, including Saudi Arabia and the United Arab Emirates", may have committed war crimes.[73] The war in/over Yemen has been described by some media as "the worst humanitarian crisis in the world",[74] with consequences including the "worst famine in 100 years".[75] For instance, in October 2020, 325,000 children under the age of 5 were treated for acute malnutrition,[76] and in June 2020 it was estimated that 2 million children under the age of 5 were malnourished.[77]

Yemen has in common with Syria the persistent predictions of its deepening catastrophes of human suffering, and the surprise on the part of a

71 Some six years later, the name "Caesar" was attached to the US National Defense Authorization Act of 2020, in the form of the Caesar Syria Civilian Protection Act, although this Act was not about intervening to change material conditions on the ground, but the pursuit of human rights abuses. The Act commits the US president to "submit[ting] to the appropriate congressional committees a list of foreign persons that the President determines are knowingly responsible for or complicit in serious human rights abuses committed against citizens of Syria or their family members, regardless of whether such abuses occurred in Syria". See: https://tinyurl.com/35xq6ejj.

72 "The Crisis in Yemen: What You Need to Know", *New York Times*, 21 April 2015, available at: www.nytimes.com/interactive/2015/03/26/world/middleeast/yemen-crisis-explained.html?action=click&module=RelatedCoverage&pgtype=Article&region=Footer.

73 UN Human Rights, *Situation of Human rights in Yemen, Including Violations and Abuses since September 2014: Report of the Group of Eminent International and Regional Experts as submitted to the United Nations High Commissioner for Human Rights*, UN Doc A/HRC/42/17, 9 August 2019, paras 97(b)–(c).

74 Robert F. Worth, "How the War in Yemen Became a Bloody Stalemate – and the Worst Humanitarian Crisis in the World", *New York Times*, 31 October 2018, available at: www.nytimes.com/interactive/2018/10/31/magazine/yemen-war-saudi-arabia.html.

75 "Yemen Could Be 'Worst Famine in 100 Years'", *BBC News*, 14 October 2018, available at: www.bbc.co.uk/news/av/world-middle-east-45857729.

76 UNICEF, *Yemen Country Office Humanitarian Situation Report*, October 2020, available at: www.unicef.org/media/89831/file/Yemen-Humanitarian-SitRep-31-October-2020.pdf.

77 UNICEF, *Yemen Five Years On: Children, Conflict and COVID-19*, June 2020, available at: www.unicef.org/yemen/media/4281/file/Yemen%20five%20years%20on_REPORT.pdf.

number of informed commentators over what they see as a lack of visible and coordinated humanitarian relief. For example, on 19 August 2015, the UN World Food Programme (WFP) estimated that "the number of food insecure people in Yemen is now close to 13 million, including 6 million who are severely food insecure and in urgent need of external assistance – that is one in five of the country's population".[78] On the same day, UN Humanitarian Affairs Coordinator Stephen O'Brien called the scale of human suffering in Yemen "nearly incomprehensible …[;] a shocking four out of five Yemenis require humanitarian assistance and nearly 1.5 million people are internally displaced".[79] Twenty million Yemeni people (two thirds of the population) are now considered food insecure, including 10 million that are acutely food insecure.[80] Some 24.1 million people now require humanitarian assistance (closer to five in six), and 3.65 million have been displaced.[81]

Outrage at impotence in the face of news of suffering and death has become part of a new culture of catastrophe, a new norm of confusion at global indifference to the inhumane. "We can no longer just stand by while children are dying in war zones",[82] declares an article from the former UK prime minister, Gordon Brown, in November 2018. "The unspeakable suffering in Yemen has barely elicited more than collective sighs on the global stage",[83] complains another headline in June 2018. Yemen-based journalist Peter Salisbury, in an essay reflecting on events at the end of 2018, concludes: "As the country slips into unimaginable, desperate hunger, it's important to understand that what is happening was utterly, tragically predictable. The people who should have known knew. They just had other priorities."[84]

Meanwhile, there has been a persistent expression of disbelief at the failure of the parties to the Syrian conflict to agree to a pause in the fighting in order to allow for substantive humanitarian relief. For instance, in a blistering indictment in October 2016, Stephen O'Brien called the failure of the UN Security Council to stop the bombing of eastern Aleppo "our generation's shame".[85]

78 WFP, "WFP Warns of Food Crisis in Yemen Amid Challenges in Reaching People and Shortage Of Funding", 19 August 2015, available at: www.wfp.org/news/news-release/wfp-head-warns-growing-food-crisis-yemen-amid-challenges-reaching-people-and-short.

79 Karin Zeitvogel, "Compassion Fatigue Sets In as Yemen Spirals Out of Control", *Washington Diplomat*, 30 September 2015, available at: https://washdiplomat.com/compassion-fatigue-sets-in-as-yemen-spirals-out-of-control/.

80 WFP, "Yemen", available at: www.wfp.org/countries/yemen.

81 OCHA, "Yemen Situation Report", available at: https://reports.unocha.org/en/country/yemen.

82 Gordon Brown, "We Can No Longer Just Stand By While Children Are Dying in War Zones", *The Guardian*, 2 November 2018, available at: www.theguardian.com/commentisfree/2018/nov/02/children-dying-war-zones-targeted-impunity.

83 Kamal Al-Solaylee, "The Unspeakable Suffering in Yemen Has Barely Elicited More Than Collective Sighs on the Global Stage", *Globe and Mail*, 22 June 2018, available at: www.theglobeandmail.com/opinion/article-the-unspeakable-suffering-in-yemen-has-barely-elicited-more-than/.

84 Peter Salisbury, "Yemen's Looming Famine Has Been a Long Time Coming", *Washington Post*, 5 December 2018, available at: www.washingtonpost.com/news/monkey-cage/wp/2018/12/05/yemens-looming-famine-has-been-a-long-time-coming/?utm_term=.6929460263d2.

85 UN Security Council, *The Situation in the Middle East*, UN Doc S/PV.7795, 26 October 2016, p. 6.

Some commentators have employed images of past catastrophes to try to render the siege of Aleppo intelligible through historic comparison. One *New York Times* headline simply read: "Berlin, 1945; Grozny, 2000; Aleppo, 2016".[86] Drone video of the devastation of Aleppo, embedded in the online article, is compared with the state of Berlin at the end of the Second World War, and also the Chechen capital after its siege and assault by Russian forces at the turn of this century.

Implicit in this reporting is the idea that lessons continue to be unlearned but also that the "industrial nature of murder", once seen, no longer has the same impact. This is particularly the case with regard to its inexorable unfolding across multiple intractable or seemingly permanent conflicts: "Today we are assaulted online, on television and in newspapers with big, senseless numbers: At least 140 killed in the Saudi-led bombing of a funeral in Yemen; hundreds slain by car bombs in Baghdad; thousands upon thousands slaughtered in Aleppo."[87] Although characterizations of this century's conflicts as uniquely "permanent" or "perpetual" are overdone,[88] the media "assault" of catastrophe, in the digital media ecology, blunts rather than harnesses outrage.

Thus in Syria, a cycle of image–expectation–inaction became increasingly acute, with a succession of key media exposures around which many assumed would coalesce some kind of international momentum for intervention. These included the thorough arresting of the world's attention through the image of Alan Kurdi's body in September 2015, and the image of the dazed, bloodied and dust-covered 5-year-old Omran Daqneesh sat in an ambulance, after being injured in a military strike in Aleppo in August 2016.[89]

To go further, there is not only an assumption that there should be a response to the publication and circulation of images and video of suffering and death, and especially of children; in addition, saturation coverage is mistaken for effects. Digital contagion is particularly seductive in terms of equating to some kind of effect, such as the fact that images of Alan Kurdi were seen on social media by "20 million people in 12 hours" (see above). The reality, however, is that the situation on the ground is not really dented by this kind of media exposure. For instance, despite the supposed impact of the Alan Kurdi image, a total of 13,622 people died or disappeared attempting to cross the Mediterranean from 1 January 2016 to 31 December 2020.[90]

In sum, then, the famine and the killing of children and other civilians in Yemen and in Syria are stories that were foretold, predicted and premediated. Both of these crises were in the public eye and warned of for years, and yet this

86 Michael Kimmelman, "Berlin, 1945; Grozny, 2000; Aleppo, 2016", *New York Times*, 14 October 2016, available at: www.nytimes.com/2016/10/15/world/middleeast/aleppo-destruction-drone-video.html.

87 *Ibid.*

88 Andrew Hoskins, "About the Project: Forgetting War", available at: https://archivesofwar.gla.ac.uk/forgetting-war/about-the-project/.

89 Elle Hunt, "Boy in the Ambulance: Shocking Image Emerges of Syrian Child Pulled from Aleppo Rubble", *The Guardian*, 18 August 2016, available at: www.theguardian.com/world/2016/aug/18/boy-in-the-ambulance-image-emerges-syrian-child-aleppo-rubble.

90 See the Missing Migrants website, available at: https://missingmigrants.iom.int/.

visibility in the digital era did not, in itself, translate to meaningful, visible action or response. What then sustains the fundamental disjuncture between this persistent mediation of catastrophe in terms of images and scale, and the belief that these should or do have a substantive impact on those suffering or dying?

Digital cover

As I set out above, there is something of a historical shift from the broadcast era to today's digital distorting of the value of exposure. Two of the defining parameters of this debate are, firstly, sensitivity – what should be shown or hidden for fear of upsetting or turning away audiences in a hugely competitive market in which both the value and the price of news have tumbled – and, secondly, digital overload and distraction – there are simply too many images and too many intractable humanitarian catastrophes competing for our attention.

Ariella Azoulay, for instance, argues that complaints over the role of news media in sanitizing views of conflict and suffering, yet at the same time overloading our attention with too many images, reveal a prevalent "phantasmic model" that is founded upon unrealistic expectations of both images and spectators. In this model, "the much sought-after object of vision is a sort of pure object that makes it possible to see war with utter clarity".[91] The other side of this equation is "the passion for a pure spectator who will encounter the image, be appalled by what is revealed, and successfully change the world through her active response to it".[92]

David Rieff points to where he thinks this expectation of the pure image and the pure spectator might come from:

> It was the conceit of journalists … that if people back home could only be told and shown what was actually happening in Sarajevo, if they had to see on their television screens images of what a child who has just been hit by a soft-nosed bullet or a jagged splinter of shrapnel really looks like … then they would want their governments to do something.[93]

Furthermore, Rieff, in an interview with me, asked: "Are we right to talk in terms of images of wars which people don't really care about? The real reason the images don't work is because people don't care about those wars to begin with and nothing is going to talk them into it."[94] And, as Susie Linfield puts it: "For most of history most people have known little, and cared less, about the suffering of those who are unknown or alien."[95]

Thus, it is a mistake to equate the now astonishing availability of digital images and video of suffering with accessibility, knowledge, understanding or

91 Ariella Azoulay, *The Civil Contract of Photography*, Zone Books, New York, 2008, p. 191.
92 *Ibid.*
93 David Rieff, *Slaughterhouse: Bosnia and the Failure of the West*, Simon & Schuster, London, 1996, p. 216.
94 David Rieff, interview with author, Glasgow, 6 March 2018.
95 Susie Linfield, *The Cruel Radiance: Photography and Political Violence*, University of Chicago Press, Chicago, IL, 2010, p. 46.

capacity for response. Rather, as I have stated, the volume and apparent availability of images and video affords the impression that much of human and distant suffering is seen by all, including notably those with the power to intervene.

When sixty-eight medical facilities in Syria are bombed between April and December 2019, there must surely be some reflection on the relationship between the instant availability of billions of images of human suffering and death in the continuous and connective digital glare of social media, and their effect on global actors charged with the preservation of civilian life.[96] It is easy to conclude that any such relationship does not in fact exist and is born from an unrealistic or mythical expectation of some kind of functional journalism.

As information pours through the imploded battlefield, in which individuals are targets, and billions of participants upload their versions of events via Facebook, Instagram, YouTube and WhatsApp, the truth of war is pixelated to oblivion. The granularity of the view of proximity to events paradoxically renders them out of focus. At the same time, despite the apparent curation by the many for the many, and the peer-to-peer connectivity ideal of uninhibited access and the digital culture of sharing, the fundamentals of what is seen and not seen are actually more opaque than those of the ideology of twentieth-century media. The monoliths of Google (which owns YouTube) and Facebook (which owns Instagram and WhatsApp) manage what is seen and not seen through their platform rules, design and algorithms. Yet these organizations have been shown to be shaky guardians of the proliferation and connectivity of content that they have enabled. The very same algorithmic basis of their control of information feeds is exploited by the insurgents of digital war, who hack attention in order to spread misinformation and undermine trust in media more generally.

Two of the principal dimensions of the life of information in the digital media ecology, pervasiveness and connectivity, destabilize traditional modes of watching – the idea of being a spectator long associated with the medium of television. Although some will argue that television – and even television news – defies reports of its demise, this misses the point that we are awash with new visual forms and means of manipulation of content which swamp the traditional, more regulated avenue of broadcast television.

Meanwhile, the displayed virality of social media affords a sense of feeling active, of participation in a swarm-like momentum gathering around an image and behind a cause. This, though, is weightless media: echoing the demand that "something must be done", paradoxically absorbing the response in and by the same digital crowd. The collective are too easily convinced, carried by the velocity of the media that immerses them. This I call a "sharing without sharing",[97] in that individuals feel active in their digital practices of posting, linking, liking, snapping, recording, swiping, scrolling and forwarding media content, but this

96 Union of Medical Care and Relief Organizations, "Humanitarian Catastrophe in Maarat Al Nouman, Syria; At Least 38 Dead, Hundreds Injured", press release, 20 December 2019, available at: https://reliefweb.int/report/syrian-arab-republic/humanitarian-catastrophe-maarat-al-nouman-syria-least-38-dead-hundreds.

97 Andrew Hoskins, "The Restless Past: An Introduction to Digital Memory and Media", in A. Hoskins (ed.), *Digital Memory Studies: Media Pasts in Transition*, Routledge, New York, 2018, p. 2.

action paradoxically functions to disconnect them from that which is passed on or passed over. The work of compassion is not really a collective endeavour, but is fractalized as the presumption of engagement with the content is reduced to the convenience of social media likes or shares. It is this false measure of presence, of virality, that helps to abrogate any kind of responsibility for what is seen.

In sum, there never was a golden age of compassion; rather, the height of the broadcast era afforded the impression of a direct relationship between images, a collective will, and effects on the events portrayed in the images. Today, algorithmically charged outrage is a proxy for effects. It is easy to misconstrue the velocity of linking and liking and "sharing without sharing" as some kind of mass action or mass movement, but in reality, the outrage society oddly numbs itself.

The transition from spectators (part of what was formerly known as the audience) to information-doers of a digital multitude has occurred at the expense of, rather than to the advance of, collective influence. The wars in Syria and Yemen are just two catastrophes that have slowly unfolded in an age of continuous and connective digital glare. If these imploded battlefields of digital war, affording the most proximate and persistent view of human suffering and death in history, cannot ultimately mobilize global action, then it is difficult to imagine what will.

International Review of the Red Cross (2020), **102** (913), 145–180.
Digital technologies and war
doi:10.1017/S1816383121000011

IRRC

AI for humanitarian action: Human rights and ethics

Michael Pizzi, Mila Romanoff and Tim Engelhardt*
Michael Pizzi is a Research Fellow at UN Global Pulse and a Digital Ethics Fellow at the Jain Family Institute.

Mila Romanoff is a Privacy Specialist and Data Governance and Policy Lead at UN Global Pulse.

Tim Engelhardt is a Human Rights Officer at the Office of the UN High Commissioner for Human Rights.

Abstract

Artificial intelligence (AI)-supported systems have transformative applications in the humanitarian sector but they also pose unique risks for human rights, even when used with the best intentions. Drawing from research and expert consultations conducted across the globe in recent years, this paper identifies key points of consensus on how humanitarian practitioners can ensure that AI augments – rather than undermines – human interests while being rights-respecting. Specifically, these consultations emphasized the necessity of an anchoring framework based on international human rights law as an essential baseline for ensuring that human interests are embedded in AI systems. Ethics, in addition, can play a complementary role in filling gaps and elevating standards above the minimum requirements of international human rights law. This paper summarizes the advantages of this framework, while also identifying specific tools and best practices that either already exist and can be adapted to the AI context, or that need to be created, in order to operationalize this human rights framework. As the COVID crisis has laid bare, AI will increasingly shape the global response to the world's toughest problems, especially in the development and humanitarian sector. To ensure that

* The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations.

AI tools enable human progress and contribute to achieving the Sustainable Development Goals, humanitarian actors need to be proactive and inclusive in developing tools, policies and accountability mechanisms that protect human rights.

Keywords: artificial intelligence, AI ethics, machine learning, human rights, humanitarianism, humanitarian organizations.

Introduction

The COVID-19 pandemic currently roiling around the globe has been devastating on many fronts. As the United Nations (UN) Secretary-General recently noted, however, the pandemic has also been a learning opportunity about the future of global crisis response. Specifically, the world is "witnessing first-hand how digital technologies help to confront the threat and keep people connected".[1] Artificial intelligence (AI) is at the forefront of many of these data-driven interventions. In recent months, governments and international organizations have leveraged the predictive power, adaptability and scalability of AI systems to create predictive models of the virus's spread and even facilitate molecular-level research.[2] From contact tracing and other forms of pandemic surveillance to clinical and molecular research, AI and other data-driven interventions have proven key to stemming the spread of the disease, advancing urgent medical research and keeping the global public informed.

The purpose of this paper is to explore how a governance framework that draws from human rights and incorporates ethics can ensure that AI is used for humanitarian, development and peace operations without infringing on human rights. The paper focuses on the use of AI to benefit the UN Sustainable Development Goals (SDGs) and other humanitarian purposes. Accordingly, it will focus on risks and harms that may arise *inadvertently* or *unavoidably* from uses that are intended to serve a legitimate purpose, rather than from malicious uses of AI (of which there could be many).

As the Secretary-General has noted, AI is already "ubiquitous in its applications"[3] and the current global spotlight is likely to expedite its adoption

1 UN General Assembly, *Roadmap for Digital Cooperation: Implementation of the Recommendations of the High-Level Panel on Digital Cooperation. Report of the Secretary-General*, UN Doc. A/74/821, 29 May 2020 (Secretary-General's Roadmap), para. 6, available at: https://undocs.org/A/74/821 (all internet references were accessed in December 2020).

2 See, for example, the initiatives detailed in two recent papers on AI and machine learning (ML) applications in COVID response: Miguel Luengo-Oroz *et al.*, "Artificial Intelligence Cooperation to Support the Global Response to COVID-19", *Nature Machine Intelligence*, Vol. 2, No. 6, 2020; Joseph Bullock *et al.*, "Mapping the Landscape of Artificial Intelligence Applications against COVID-19", *Journal of Artificial Intelligence Research*, Vol. 69, 2020, available at: www.jair.org/index.php/jair/article/view/12162.

3 Secretary-General's Roadmap, above note 1, para. 53.

even further.[4] As the COVID crisis has laid bare, AI will increasingly shape the global response to the world's toughest problems, especially in the fields of development and humanitarian aid. However, the proliferation of AI, if left unchecked, also carries with it serious risks to human rights. These risks are complex, multi-layered and highly context-specific. Across sectors and geographies, however, a few stand out.

For one, these systems can be extremely powerful, generating analytical and predictive insights that increasingly outstrip human capabilities. They are therefore liable to be used as replacements for human decision-making, especially when analysis needs to be done rapidly or at scale, with human overseers often overlooking their risks and the potential for serious harms to individuals or groups of individuals that are already vulnerable.[5] Artificial intelligence also creates challenges for transparency and oversight, since designers and implementers are often unable to "peer into" AI systems and understand how and why a decision was made. This so-called "black box" problem can preclude effective accountability in cases where these systems cause harm, such as when an AI system makes or supports a decision that has a discriminatory impact.[6]

Some of the risks and harms implicated by AI are addressed by other fields and bodies of law, such as data privacy and protection,[7] but many appear to be entirely new. AI ethics, or AI governance, is an emerging field that seeks to address the novel risks posed by these systems. To date, it is dominated by the proliferation of AI "codes of ethics" that seek to guide the design and deployment of AI systems. Over the past few years, dozens of organizations – including international organizations, national governments, private corporations and non-governmental organizations (NGOs) – have published their own sets of principles that they believe should guide the responsible use of AI, either within their respective organizations or beyond them.[8]

4 AI is "forecast to generate nearly $4 trillion in added value for global markets by 2022, even before the COVID-19 pandemic, which experts predict may change consumer preferences and open new opportunities for artificial intelligence-led automation in industries, businesses and societies". *Ibid.*, para. 53.

5 Lorna McGregor, Daragh Murray and Vivian Ng, "International Human Rights Law as a Framework for Algorithmic Accountability", *International and Comparative Law Quarterly*, Vol. 68, No. 2, 2019, available at: https://tinyurl.com/yaflu6ku.

6 See, for example, Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation", *Harvard Journal of Law and Technology*, Vol. 31, No. 2, 2018; Rachel Adams and Nora Ni Loideain, "Addressing Indirect Discrimination and Gender Stereotypes in AI Virtual Personal Assistants: The Role of International Human Rights Law", paper presented at the Annual Cambridge International Law Conference 2019, "New Technologies: New Challenges for Democracy and International Law", 19 June 2019, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3392243.

7 See, for example, Global Privacy Assembly, "Declaration on Ethics and Data Protection in Artificial Intelligence", Brussels, 23 October 2018, available at: http://globalprivacyassembly.org/wp-content/uploads/2019/04/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf; UN Global Pulse and International Association of Privacy Professionals, *Building Ethics into Privacy Frameworks for Big Data and AI*, 2018, available at: https://iapp.org/resources/article/building-ethics-into-privacy-frameworks-for-big-data-and-ai/.

While these efforts are often admirable, codes of ethics are limited in key respects: they lack a universally agreed framework; they are not binding, like law, and hence do not promulgate compliance; they often reflect the values of the organization that created them, rather than the diversity of those potentially impacted by AI systems; and they are not automatically operationalized by those designing and applying AI tools on a daily basis. In addition, the drafters of these principles often provide little guidance on how to resolve conflicts or tensions between them (such as when heeding one principle would undermine another), making them even more difficult to operationalize. Moreover, because tech companies create or control most AI-powered products, this governance model relies largely on corporate self-regulation – a worrying prospect given the absence of democratic representation and accountability in corporate decision-making.

Applying and operationalizing these principles to development and humanitarian aid poses an additional set of challenges. With the exception of several recent high-quality white papers on AI ethics and humanitarianism, guidance for practitioners in this rapidly evolving landscape remains scant.[9] This is despite the existence of several factors inherent in development or humanitarian projects that either exacerbate traditional AI ethics challenges or implicate entirely new ones.

AI governance is quickly emerging as a global priority. As the Secretary-General's Roadmap for Digital Cooperation states clearly and repeatedly, the global approach to AI – during COVID and beyond – must be in full alignment with human rights.[10] The UN and other international organizations have devoted increasing attention to this area, reflecting both the increasing demand for AI and other data-driven solutions to global challenges – including the SDGs – and the ethical risks that these solutions entail. In 2019, both the UN General Assembly[11] and UN Human Rights Council (HRC)[12] passed resolutions calling for the application of international human rights law to AI and other emerging digital technologies, with the General Assembly warning that "profiling, automated decision-making and machine-learning technologies, … without proper

8 For an overview, see Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy and Madhulika Srikumar, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center Research Publication No. 2020-1, 14 February 2020.

9 See Faine Greenwood, Caitlin Howarth, Danielle Escudero Poole, Nathaniel A. Raymond and Daniel P. Scarnecchia, *The Signal Code: A Human Rights Approach to Information During Crisis*, Harvard Humanitarian Initiative, 2017, p. 4, underlining the dearth of rights-based guidance for humanitarian practitioners working with big data. There are a few existing frameworks, however – most notably Data Science & Ethics Group (DSEG), *A Framework for the Ethical Use of Advanced Data Science Methods in the Humanitarian Sector*, April 2020, available at: https://tinyurl.com/yazcao2o. There have also been attempts to guide practitioners on humanitarian law as it applies to lethal autonomous weapons systems, including the Asser Institute's Designing International Law and Ethics into Military AI (DILEMA) project, available at: www.asser.nl/research/human-dignity-and-human-security/designing-international-law-and-ethics-into-military-ai-dilema.

10 Secretary-General's Roadmap, above note 1, para. 50.

11 UNGA Res. 73/179, 2018.

12 HRC Res. 42/15, 2019.

safeguards, may lead to decisions that have the potential to affect the enjoyment of human rights".[13]

There is an urgency to these efforts: while we wrangle with how to apply human rights principles and mechanisms to AI, digital technologies continue to evolve rapidly. The international public sector is deploying AI more and more frequently, which means new risks are constantly emerging in this field. The COVID-19 pandemic is a timely reminder. To ensure that AI tools enable human progress and contribute to achieving the SDGs, there is a need to be proactive and inclusive in developing tools, policies and accountability mechanisms that protect human rights.

The conclusions contained herein are based on qualitative data emerging from multi-stakeholder consultations held or co-hosted by UN Global Pulse along with other institutions responsible for protecting privacy and other human rights, including the Office of the UN High Commissioner for Human Rights (UN Human Rights) and national data protection authorities;[14] multiple interviews and meetings with the diverse panel of AI and data experts that comprise Global Pulse's Expert Group on Governance of Data and AI;[15] guidance and reporting from UN human rights experts; scholarly work on human rights and ethics; and practical guidance for the development and humanitarian sectors issued by organizations like the World Health Organization, the UN Office for the Coordination of Humanitarian Affairs (OCHA),[16] the International Committee of the Red Cross (ICRC),[17] the Harvard Humanitarian Initiative[18], Access Now,[19] Article 19,[20] USAID's Center for Digital Development,[21] and the Humanitarian Data Science and Ethics Group (DSEG).[22]

13 UNGA Res. 73/179, 2018.

14 Consultations include practical workshops on designing frameworks for ethical AI in Ghana and Uganda; on AI and privacy in the global South at RightsCon in Tunis; on a human rights-based approach to AI in Geneva, co-hosted with UN Human Rights; several events at the Internet Governance Forum in Berlin; and a consultation on ethics in development and humanitarian contexts, co-hosted with the International Association of Privacy Professionals and the European Data Protection Supervisor. These various consultations, which took place between 2018 and 2020, included experts from governments, international organizations, civil society and the private sector, from across the globe.

15 See the UN Global Pulse Expert Group on Governance of Data and AI website, available at: www.unglobalpulse.org/policy/data-privacy-advisory-group/.

16 See the OCHA, *Data Responsibility Guidelines: Working Draft*, March 2019, available at: https://tinyurl.com/y64pcew7.

17 ICRC, *Handbook on Data Protection in Humanitarian Action*, Geneva, 2017.

18 F. Greenwood *et al.*, above note 9.

19 Access Now, *Human Rights in the Age of Artificial Intelligence*, 2018, available at: www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf.

20 Article 19, *Governance with Teeth: How Human Rights can Strengthen FAT and Ethics Initiatives on Artificial Intelligence*, April 2019, available at: www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

21 USAID Center for Digital Development, *Reflecting the Past, Shaping the Future: Making AI Work for International Development*, 2018.

22 DSEG, above note 9.

AI in humanitarian aid: Opportunities

Artificial intelligence is not a specific technology. Rather, it is a broad term encompassing a set of tools or capabilities that seek to emulate aspects of human intelligence. As a category, AI generally refers to a system that automates an analytical process, such as the identification and classification of data; in rarer cases, an AI system may even automate a decision. Hence, some prefer the term "automated intelligent system" rather than the more commonly used "artificial intelligence" or "AI". For the purposes of this paper, "AI" will refer primarily to machine learning (ML) algorithms, which are a common component of AI systems defined by the ability to detect patterns, learn from those patterns, and apply those learnings to new situations.[23] ML models may be either supervised, meaning that they require humans to feed them a set of rules to apply, or unsupervised, meaning that the model is capable of learning rules from the data itself and therefore does not require human coders to feed in rules. For this reason, this latter set of models is often described as self-teaching.[24] Deep learning (DL) is, in turn, a more potent subset of ML that uses layers of artificial neural networks (which are modelled after neurons in the human brain) to detect patterns and make predictions.[25]

Algorithmic systems are capable of "execut[ing] complex tasks beyond human capability and speed, self-learn[ing] to improve performance, and conduct [ing] sophisticated analysis to predict likely future outcomes".[26] Today, these systems have numerous capabilities that include natural language processing, computer vision, speech and audio processing, predictive analytics and advanced robotics.[27] These and other techniques are already being deployed to augment development and humanitarian action in innovative ways. Computer vision is being used to automatically identify structures in satellite imagery, enabling the rapid tracking of migration flows and facilitating the efficient distribution of aid in humanitarian crises.[28] Numerous initiatives across the developing world are using AI to provide predictive insights to farmers, enabling them to mitigate the hazards of drought and other adverse weather, and maximize crop yields by sowing seeds at the optimal moment.[29] Pioneering AI tools enable remote

23 Jack M. Balkin, "2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data", Ohio State Law Journal, Vol. 78, No. 5, 2017, p. 1219 (cited in L. McGregor, D. Murray and V. Ng, above note 5, p. 310). See also the European Union definition of artificial intelligence: "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals." European Commission, "A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines", 8 April 2019, available at: https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

24 See "Common ML Problems" in Google's Introduction to Machine Learning Problem Framing course, available at: https://developers.google.com/machine-learning/problem-framing/cases.

25 Tao Liu, "An Overview of the Application of AI in Development Practice", Berkeley MDP, available at: https://mdp.berkeley.edu/an-overview-of-the-application-of-ai-in-development-practice/.

26 L. McGregor, D. Murray and V. Ng, above note 5, p. 310.

27 For good definitions of each of these terms, see Access Now, above note 19, p. 8.

28 See UN Global Pulse's PulseSatellite project, available at: www.unglobalpulse.org/microsite/pulsesatellite/.

diagnosis of medical conditions like malnutrition in regions where medical resources are scarce.[30] The list grows longer every day.[31]

Several factors explain the proliferation of AI in these and other sectors. Perhaps the most important catalyst, however, is the data revolution that has seen the exponential growth of data sets relevant to development and humanitarianism.[32] Data are essential fuel for AI development; without training on relevant data sets, an AI model cannot learn. Finding quality data has traditionally been more difficult in developing economies, particularly in least developed countries[33] and in humanitarian contexts, where technological infrastructure, resources and expertise are often rudimentary. According to a recent comprehensive white paper from the DSEG, however, this has begun to change:

> Currently, we are witnessing unprecedented rates of data being collected worldwide, a wider pool of stakeholders producing "humanitarian" data, data becoming more machine readable, and data being more accessible via online portals. This has enabled an environment for innovation and progress in the sector, and has led to enhanced transparency, informed decision making, and effective humanitarian service delivery.[34]

Key challenges for rights-respecting AI

The very characteristics that make AI systems so powerful also pose risks for the rights and freedoms of those impacted by their use. This is often the case with emerging digital technologies, however, so it is important to be precise about what exactly it is about AI that is "new" or unique – and therefore why it requires particular attention A thorough technical analysis of AI's novel characteristics is beyond the scope of this paper, but some of the most frequently cited challenges of AI systems in the human rights conversation are summarized in the following paragraphs.

29 Examples include AtlasAI, EzyAgric, Apollo, FarmForce, Tulaa and Fraym.

30 See, for example, Kimetrica's Methods for Extremely Rapid Observation of Nutritional Status (MERON) tool, a project run in coordination with UNICEF that uses facial recognition to remotely diagnose malnutrition in children.

31 For more examples of AI projects in the humanitarian sector, see International Telecommunications Union, *United Nations Activities on Artificial Intelligence (AI)*, 2019, available at: www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf; accepted papers of the Artificial Intelligence for Humanitarian Assistance and Disaster Response Workshop, available at: www.hadr.ai/accepted-papers; and the list of projects in DSEG, above note 9, Chap. 3.

32 UN Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development, *A World That Counts: Mobilising the Data Revolution for Sustainable Development*, 2014.

33 See UN Department of Economic and Social Affairs, "Least Developed Countries", available at: www.un.org/development/desa/dpad/least-developed-country-category.html.

34 DSEG, above note 9, p. 3.

Lack of transparency and explainability

AI systems are often obscure to human decision-makers; this is also known as the black box problem.[35] Unlike traditional algorithms, the decisions made by ML or DL processes can be impossible for humans to trace, and therefore to audit or otherwise explain to the public and to those responsible for monitoring their use (this also known as the principle of explainability).[36] This means that AI systems can also be obscure to those impacted by their use, leading to challenges for ensuring accountability when systems cause harm. The obscurity of AI systems can preclude individuals from recognizing if and why their rights were violated and therefore from seeking redress for those violations. Moreover, even when understanding the system is possible, it may require a high degree of technical expertise that ordinary people do not possess.[37] This can frustrate efforts to pursue remedies for harms caused by AI systems.

Accountability

This lack of transparency and explainability can severely impede effective accountability for harms caused by automated decisions, both on a governance and an operational level. The problem is twofold. First, individuals are often unaware of when and how AI is being used to determine their rights.[38] As the former UN Special Rapporteur on the Promotion and Protection of Freedom of Opinion and Expression David Kaye has warned, individuals are unlikely to be aware of the "scope, extent or even existence of the algorithmic decision-making processes that may have an impact on their enjoyment of rights". Individual notice about the use of AI systems is therefore "almost inherently unavailable".[39] This is especially true in humanitarian contexts, where impacted individuals are often not able to give meaningful consent to data collection and analysis (e.g., because it is required to receive essential services).[40]

Second, the obscurity of the data economy and its lack of accountability for human rights[41] can make it difficult for individuals to learn of harms to their rights

35 Cynthia Rudin and Joanna Radin. "Why Are We Using Black Box Models in AI When We Don't Need To?", *Harvard Data Science Review*, Vol. 1, No. 2, 2019, available at: https://doi.org/10.1162/99608f92.5a8a3a3d.

36 See Miriam C. Buiten, "Towards Intelligent Regulation of Artificial Intelligence", *European Journal of Risk Regulation*, Vol. 10, No. 1, 2019, available at: https://tinyurl.com/y8wqmp9a; Anna Jobin, Marcello Ienca and Effy Vayena, "The Global Landscape of AI Ethics Guidelines", *Nature Machine Intelligence*, Vol. 1, No. 9, 2019, available at: www.nature.com/articles/s42256-019-0088-2.pdf.

37 See, for example, L. McGregor, D. Murray and V. Ng, above note 5, p. 319, explaining the various risks caused by a lack of transparency and explainability: "as the algorithm's learning process does not replicate human logic, this creates challenges in understanding and explaining the process".

38 David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc. A/73/348, 29 August 2018, para. 40.

39 *Ibid.*, speaking about the application of AI in the online information environment.

40 DSEG, above note 9, p. 7.

41 Isabel Ebert, Thorsten Busch and Florian Wettstein, *Business and Human Rights in the Data Economy: A Mapping and Research Study*, German Institute for Human Rights, Berlin, 2020.

and to seek redress when those harms occur. It can also make it difficult even for knowledgeable experts or fact-finders to audit these systems and diagnose faults. The organizational complexity of most development and humanitarian projects can compound these challenges.[42] When a single project comprises a long chain of actors (including funders, foreign governments, international organizations, contractors, private sector vendors, local government entities, civil society partners and data collectors), who is ultimately responsible when a system spits out a discriminatory decision (or analysis that ultimately sways said decision)?

Unpredictability

A hallmark of ML and DL algorithms is their ability to learn and evolve in unpredictable ways. Put another way, they are able to "progressively identify new problems and develop new answers. Depending on the level of supervision, systems may identify patterns and develop conclusions unforeseen by the humans who programmed or tasked them."[43] Therein lies their essential value; ML algorithms can, in some cases, analyze data that they have not necessarily been trained to analyze, enabling them to tackle new tasks or even operate in new contexts. At the same time, however, a system's functional solutions will not always be logical or even understandable to human interpreters. This characteristic makes it difficult for human designers and implementers to predict – let alone explain – the nature and level of risk posed by a system or its application in a specific context. Moreover, there is a limit to the adaptability of even the most potent ML systems. Many do *not* generalize well to new contexts, resulting in extreme unpredictability when deployed on data that differs significantly from their training data.

Erosion of privacy

The ability of AI systems to analyze and draw inferences from massive quantities of private or publicly available data can have serious implications for many protected facets of the right to privacy. AI systems can reveal sensitive insights into individuals' whereabouts, social networks, political affiliations, sexual preferences and more, all based on data that people voluntarily post online (such as the text and photos that users post to social media) or incidentally produce from their digital devices (such as GPS or cell-site location data).[44] These risks are especially acute in humanitarian contexts, where those impacted by an AI system are likely

42 Lindsey Andersen, "Artificial Intelligence in International Development: Avoiding Ethical Pitfalls", *Journal of Public and International Affairs*, 2019, available at: https://jpia.princeton.edu/news/artificial-intelligence-international-development-avoiding-ethical-pitfalls.

43 D. Kaye, above note 38, para. 8.

44 See HRC, *Question of the Realization of Economic, Social and Cultural Rights in All Countries: The Role of New Technologies for the Realization of Economic, Social and Cultural Rights. Report of the Secretary-General*, UN Doc. A/HRC/43/29, 4 March 2020 (ESCR Report), p. 10. See also Ana Beduschi, "Research Brief: Human Rights and the Governance of AI", Geneva Academy, February 2020, p. 3: "[D]ue to the increasingly sophisticated ways in which online platforms and companies track online

to be among the most marginalized. As a result, data or analysis that would not ordinarily be considered sensitive might become sensitive. For instance, basic identifying information – such as names, home towns and addresses – may be publicly available information in most contexts, but for a refugee fleeing oppression or persecution in their home country, this information could jeopardize their safety and security if it were to end up in the wrong hands.[45] In addition, data-intensive ML can incentivize further data collection, thus leading to greater interferences with privacy and also the risk of de-anonymization. Moreover, the use of AI to analyze mass amounts of personal data is also linked to infringements on other rights, including freedom of opinion and expression, freedom of association and peaceful assembly, and the right to an effective remedy.[46]

Inequalities, discrimination and bias

When the data on which an AI model is trained are incomplete, biased or otherwise inadequate, it may result in the system producing discriminatory or unfair decisions and outputs.[47] Biases and other flaws in the data can infect a system at several different stages: in the initial framing of the problem (e.g., a proxy variable is chosen that is linked to socioeconomic or racial characteristics); when the data are collected (e.g., a marginalized group is underrepresented in the training data); and when the data are prepared.[48] In some cases, the inherent biases of the developers themselves can be unintentionally coded into a model. There have been several high-profile incidents where ML systems have displayed racial or gender biases – for example, an ML tool used by Amazon for CV review that disproportionately rejected women, or facial recognition tools that are worse at recognizing non-white faces.[49] In the humanitarian context, avoiding unwanted bias and discrimination is intimately related to the core humanitarian principle of impartiality,[50] and the stakes for such discrimination can be especially high –

behaviour and individuals' digital footprints, AI algorithms can make inferences about behaviour, including relating to their political opinions, religion, state of health or sexual orientation."

45 This partly explains the pushback against facial recognition and other biometric identification technology. See, for example, The Engine Room and Oxfam, *Biometrics in the Humanitarian Sector*, March 2018; Mark Latonero, "Stop Surveillance Humanitarianism", *New York Times*, 11 July 2019; Dragana Kaurin, *Data Protection and Digital Agency for Refugees*, World Refugee Council Research Paper No. 12, May 2019.

46 ESCR Report, above note 44, p. 10.

47 D. Kaye, above note 38, paras 37–38.

48 Karen Hao, "This Is How AI Bias Really Happens – and Why It's So Hard to Fix", *MIT Technology Review*, 4 February 2019, available at: www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/. For further explanation of the types of biases that are commonly present in a data sets or training models, see DSEG, above note 9.

49 K. Hao, above note 48; Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of Machine Learning Research*, Vol. 81, 2018; Inioluwa Deborah Raji and Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, 2019.

50 "Humanitarian action must be carried out on the basis of need alone, giving priority to the most urgent cases of distress and making no distinctions on the basis of nationality, race, gender, religious belief, class or political opinions." OCHA, "OCHA on Message: Humanitarian Principles", June 2012, available at: www.unocha.org/sites/dms/Documents/OOM-humanitarianprinciples_eng_June12.pdf.

determining, for instance, who receives critical aid, or even who lives and who dies.[51] On a macro level, algorithms (including AI) can have the effect of "deepen[ing] existing inequalities between people or groups, and exacerbate[ing] the disenfranchisement of specific vulnerable demographics". This is because "[a]lgorithms, more so than other types of data analysis, have the potential to create harmful feedback loops that can become tautological in nature, and go unchecked due to the very nature of an algorithm's automation".[52]

Lack of contextual knowledge at the design phase

There is often a disconnect between the design and application stages of an AI project. This is especially critical if the system is to be applied in humanitarian contexts.[53] The tools may be designed without adequate contextual knowledge; often they are developed to be suitable for business and marketing decision-making rather than for humanitarian aid in the developing world. Tools designed without taking into account certain cultural, societal and gender-related aspects can lead to misleading decisions that detrimentally impact human lives. For example, a system conceived or designed in Silicon Valley but deployed in a developing country may fail to take into account the unique political and cultural sensitivities of that country. The developer may be unaware that in country X, certain stigmatized groups are underrepresented or even "invisible" in a data set, and fail to account for that bias in the training model; or a developer working on a tool to be deployed in a humanitarian context may not be aware that migrant communities and internally displaced persons are frequently excluded from censuses, population statistics and other data sets.[54]

Lack of expertise and last-mile implementation challenges

Insufficient expertise or training on the part of those deploying AI and other data-driven tools is associated with a number of human rights risks. This applies in the public sector, generally, where it is widely acknowledged that data fluency is lacking.[55] This may result in a tendency to incorrectly interpret a system's output, overestimate its predictive capacity or otherwise over-rely on its outputs, such as by allowing the system's "decisions" to supersede human judgement.

51 See, for example, this discussion on the implications of automated weapons systems for international humanitarian law: Noel Sharkey, "The Impact of Gender and Race Bias in AI", *ICRC Humanitarian Law and Policy Blog*, 28 August 2018, available at: https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/.

52 DSEG, above note 9, p. 29.

53 Based on our Geneva consultations.

54 For a discussion on the challenges of collecting and analyzing data on migrant populations, see Natalia Baal and Laura Ronkainen, *Obtaining Representative Data on IDPs: Challenges and Recommendations*, UNHCR Statistics Technical Series No. 2017/1, 2017, available at: www.unhcr.org/598088104.pdf.

55 The UN Data Strategy of 2020 strongly emphasizes the need for capacity-building among civil servants across the UN in the areas of data use and emerging technologies.

It may also create a risk that decision- and policy-makers will use AI as a crutch, employing AI analysis to add a veneer of objectivity or neutrality to their choices.

These risks are further exacerbated in the developing-country and humanitarian contexts, where a lack of technical resources, infrastructure or organizational capacity may preclude the successful exploitation of an AI system.[56] These so-called "last-mile implementation" challenges may elevate human rights risks and other failures, especially in humanitarian contexts. For example, shortcomings – whether anticipated or unanticipated – may increase the chance of human error, which can include anything from failing to audit the system to over-relying on, or misinterpreting, its insights. This, in turn, may lead to detrimental impacts, such as the failure to deliver critical aid, or even discrimination and persecution.

Lack of quality data

Trustworthy and safe AI depends on quality data. Without ready access to quality data sets, AI cannot be trained and used in a way that avoids amplifying the above risks. However, the degree of availability and accessibility of data often reflects social, economic, political and other inequalities.[57] In many development and humanitarian contexts, it is far more difficult to conduct quality data collection. This increases the risks that an AI system will produce unfair outcomes.[58] While data quality standards are not new – responsible technologists have long since developed principles and best practices for quality data[59] – there remains a lack of adequate legal frameworks for enabling access to usable data sets. As the Secretary-General commented in his Roadmap, "[m]ost existing digital public goods [including quality data] are not easily accessible because they are often unevenly distributed in terms of the language, content and infrastructure required to access them".[60]

Over-use of AI

The analytical and predictive capabilities of AI systems can make them highly attractive "solutions" to difficult problems, both for resource-strained practitioners in the field and for those seeking to raise funds for these projects. This creates the risk that AI may be overused, including when less risky solutions

56 Michael Chui *et al.*, *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy*, McKinsey Global Institute, September 2018.

57 On the data gap (concerning older persons), see HRC, *Enjoyment of All Human Rights by Older Persons*, UN Doc. A/HRC/42/43, 4 July 2019; HRC, *Human Rights of Older Persons: The Data Gap*, UN Doc. A/HRC/45/14, 9 July 2020.

58 Jasmine Wright and Andrej Verity, *Artificial Intelligence Principles for Vulnerable Populations in Humanitarian Contexts*, Digital Humanitarian Network, January 2020, p. 15.

59 See, for example, relevant sections in OCHA's Data Responsibility Guidelines, above note 16; the ICRC *Handbook on Data Protection in Humanitarian Action*, above note 17; and the Principles for Digital Development, available at: https://digitalprinciples.org/.

60 Secretary-General's Roadmap, above note 1, para. 23.

are available.[61] For one, there is widespread misunderstanding about the capabilities and limitations of AI, including its technical limitations. The popular depiction of AI in the media tends to be of all-powerful machines or robots that can solve a wide range of analytical problems. In reality, AI projects tend to be highly specialized, designed only for a specific use in a specific context on a specific set of data. Due to this misconception, users may be unaware that they are interacting with an AI-driven system. In addition, while AI is sometimes capable of replacing human labour or analysis, it is generally an inappropriate substitute for human decision-making in highly sensitive or high-stakes contexts. For instance, allowing an AI-supported system to make decisions on criminal sentencing, the granting of asylum[62] or parental fitness – cases where fundamental rights and freedoms are at stake, and where impacted individuals may already be traumatized or distressed – can undermine individual autonomy, exacerbate psychological harm and even erode social connections.[63]

Private sector influence

Private sector technology companies are largely responsible for developing and deploying the AI systems that are used in the development and humanitarian sectors, often by way of third-party vendor contracts or public–private partnerships. This creates the possibility that, in certain cases, corporate interests may overshadow the public interest. For example, the profit-making interest may provide a strong incentive to push for an expensive, "high-tech" approach where a "low-tech" alternative may be better suited for the environment and purposes at hand.[64] Moreover, close cooperation between States and businesses may undermine transparency and accountability, for example when access to information is inhibited on the basis of contractual agreements or trade secret protections. The deep involvement of corporate actors may also lead to the delegation of decision-making on matters of public interest. For example, there is a risk that humanitarian actors and States will "delegate increasingly complex and onerous censorship and surveillance mandates" to companies.[65]

61 "Algorithms' automation power can be useful, but can also alienate human input from processes that affect people. The use or over-use of algorithms can thus pose risks to populations affected by algorithm processes, as human input to such processes is often an important element of protection or rectification for affected groups. Algorithms can often deepen existing inequalities between people or groups, and exacerbate the disenfranchisement of specific vulnerable demographics. Algorithms, more so than other types of data analysis, have the potential to create harmful feedback loops that can become tautological in nature, and go unchecked due to the very nature of an algorithm's automation." DSEG, above note 9, p. 29.

62 Petra Molnar and Lex Gill, *Bots at the Gates*, University of Toronto International Human Rights Program and Citizen Lab, 2018.

63 DSEG, above note 9, p. 11.

64 Based on our Geneva consultations. See also Chinmayi Arun, "AI and the Global South: Designing for Other Worlds", in Markus D. Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford, 2020.

65 D. Kaye, above note 38, para. 44.

Perpetuating and deepening inequalities

Deploying complex AI systems to support services for marginalized people or people in vulnerable positions can at times have the perverse effect of entrenching inequalities and creating further disenfranchisement. Biased data and inadequate models are one of the major problems in this regard, as discussed above, but it is important to recognize that these problems can in turn be seen as expressions of deeply rooted divides along socio-economic, gender and racial lines – and an increased deployment of AI carries the real risk of widening these divides. UNESCO has recently made this point, linking it to the effects of AI on the distribution of power when it stated that "[t]he scale and the power generated by AI technology accentuates the asymmetry between individuals, groups and nations, including the so-called 'digital divide' within and between nations".[66] Corporate capture, as just addressed, can be one of the most important contributors to this development. Countering this trend is no easy task and will require political will, collaboration, open multi-stakeholder engagement, strengthening of democratic governance of societies and promoting human rights in order to empower the people to take an active role in shaping the technological and regulatory environment in which they live.

Intersectional considerations

Some of these challenges distinguish AI systems from other technologies that we have regulated in the past, and therefore may require new solutions. However, it is worth noting that some of the underlying challenges are hardly new. In this regard, we may sometimes glean best practices on governing AI from other fields. For example, data privacy and data security risks and standards developed to protect information have been in existence for a long time. It is true that as the technology develops and more data are generated, new protections need to be developed or old ones updated to reflect the new challenges. Data security remains one of the key considerations in humanitarian work given the sensitivity of the data being collected and processed.

In addition, many of the challenges facing AI in humanitarian aid have been addressed by practitioners in the wider "tech for development" field,[67] such as the challenges associated with last-mile implementation problems, as discussed above. Another perennial challenge is that development or humanitarian projects must sometimes weigh the risks of partnering with governments that have sub-par human rights records. This is undoubtedly true for powerful tools like AI. An AI system designed for a socially beneficial purpose – such as the digital contact tracing of individuals during a disease outbreak, used for containment purposes – could potentially be used by governments for invasive surveillance.[68]

66 UNESCO, *Preliminary Study on the Ethics of Artificial Intelligence*, SHS/COMEST/EXTWG-ETHICS-AI/2019/1, 26 February 2019, para. 22.

67 See, for example, the Principles for Digital Development, above note 59.

Additionally, while all the above challenges are quite common and may lead to potential harms, the organizational context in which these AI systems or processes are embedded is an equally important determinant of their risks. Regardless of a system's analytical or predictive power in isolation (whether it involves a simple algorithm or complex neural networks), we can expect drastically different benefits and risks of harms depending on the nature and degree of human interaction with, or oversight of, that system.

The challenges described above are not merely theoretical – there are already countless real-world examples where advanced AI systems have caused serious harm. In some of the highest-profile AI mishaps to date, the implementer was a government agency or other public sector actor that sought to improve or streamline a public service. For example, a recent trend is the use of algorithmic analysis by governments to determine eligibility for welfare benefits or root out fraudulent claims.[69] In Australia, the Netherlands and the United States, systemic design flaws or inadequate human oversight – among other issues – have resulted in large numbers of people being deprived their rights to financial assistance, housing or health.[70] In August 2020, the UK Home Office decided to abandon a decision-making algorithm it had deployed to screen visa applicants over allegations of racial bias.[71]

We know relatively little about the harms that have been caused by the use of AI in humanitarian contexts. As the DSEG observed in its report, there remains "a lack of documented evidence" of the risks and harms of AI "due to poor tracking and sharing of these occurrences" and a "general attitude not to report incidents".[72] While the risks outlined above have been borne out in other contexts (such as social welfare), in humanitarian contexts there is at least evidence about the potential concerns associated with biometrics and the fears of affected peoples.

A recent illustrative case study is that of Karim, a psychotherapy chatbot developed and tested on Syrian refugees living in the Zaatari refugee camp. Experts who spoke to researchers from the Digital Humanitarian Network expressed concern that the development of an AI therapy chatbot, however advanced, reflected a poor understanding of the needs of vulnerable people in that context.[73] In addition to linguistic and logistical obstacles that became

68 See UN Human Rights, *UN Human Rights Business and Human Rights in Technology Project (B-Tech): Overview and Scope*, November 2019, warning of the inherent human rights risks in "[s]elling products to, or partnering with, governments seeking to use new tech for State functions or public service delivery that could disproportionately put vulnerable populations at risk".

69 Philip Alston, *Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc. A/74/493, 11 October 2019.

70 AI Now Institute, *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems*, September 2018, available at: https://ainowinstitute.org/litigatingalgorithms.pdf; P. Alston, above note 69. Note that even a perfectly designed system with humans in the loop can still lead to bad outcomes if it is not the right approach in a given context. For instance, widespread, deeply rooted discrimination in an oppressive environment may actually have the effect of entrenching discrimination further, even if the AI system itself is not biased and there is a human in the loop.

71 Henry McDonald. "Home Office to Scrap 'Racist Algorithm' for UK Visa Applicants", *The Guardian*, 4 August 2020.

72 DSEG, above note 9, p. 3.

evident during the pilot, the experts argued that a machine therapist was not, in fact, better than having no therapist at all – that it actually risked increasing subjects' sense of alienation in the long term.[74] Karim appears to be an example of what, according to the Humanitarian Technologies Project, happens when "there is a gap between the assumptions about technology in humanitarian contexts and the actual use and effectiveness of such technology by vulnerable people".[75]

The above challenges show that piloting unproven AI tools on vulnerable populations may potentially gravely undermine human rights when those tools are ill-suited for the context or when those deploying the tools lack expertise on how to use them.[76]

Approaches to governing AI: Beyond ethics

The above examples illustrate the potential for AI to both serve human interests and to undermine them, if proper safeguards are not put in place and risks are unaccounted for. For these reasons, the technologists designing these systems and humanitarian and development experts deploying AI are increasingly cognizant of the need to infuse human rights and ethical considerations into their work. Accordingly, there is a growing body of technical specifications and standards that have been developed to ensure AI systems are "safe", "secure" and "trustworthy".[77] But ensuring that AI systems serve human interests is about more than just technical specifications. As McGregor, Murray and Ng have argued, a wider, overarching framework should be in place to incorporate risks of harm at every stage of the system's life cycle and to ensure accountability when things go wrong.[78]

Early AI governance instruments, ostensibly developed to serve this guiding role, have mostly taken the form of "AI codes of ethics".[79] These codes tend to consist of guiding principles that the organization is committed to honouring, akin to a constitution for the development and use of AI. As their names suggest, these codes tend to invoke ethical principles like fairness and justice, rather than guaranteeing specific human rights.[80] Indeed, human rights – the universal and binding system of principles and treaties that all States must observe – have been conspicuously absent from many of these documents.[81] According to Philip

73 J. Wright and A. Verity, above note 58, p. 7.
74 *Ibid.*, p. 6.
75 *Ibid.*, p. 9. See also the Humanitarian Technologies Project website, available at: http://humanitariantechnologies.net.
76 See DSEG, above note 9, p. 8, warning against piloting unproven technology in humanitarian contexts.
77 Peter Cihon, *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*, Future of Humanity Institute, University of Oxford, April 2019.
78 "[T]he complex nature of algorithmic decision-making necessitates that accountability proposals be set within a wider framework, addressing the overall algorithmic life cycle, from the conception and design phase, to actual deployment and use of algorithms in decision-making." L. McGregor, D. Murray and V. Ng, above note 5, p. 311.
79 For a summary of AI codes of ethics released by major institutions, see J. Fjeld *et al.*, above note 8.
80 *Ibid.*

Alston, the UN Special Rapporteur on Extreme Poverty and Human Rights, many AI codes of ethics include token references to human rights – for example, including a commitment to respecting "human rights" as a stand-alone principle – but fail to capture the substantive rights provided for by the Universal Declaration of Human Rights (UDHR) and human rights treaties.[82]

The shortcomings of this "ethics-first approach" are increasingly apparent. One of the key gaps is the absence of accountability mechanisms for when ethical principles are violated.[83] Most codes of ethics provide no answer for who bears the cost of an "unethical" use of technology, what that cost should be, or how violations would be monitored and enforced. Moreover, it is not clear how an individual who feels wronged can determine that a wrong has indeed occurred, or what procedure they can follow to seek redress.[84] Unlike human rights law, codes of ethics typically do not make it clear how to balance the interests of disparate groups or individuals, some of whom may benefit from an AI system to the detriment of others. While AI codes of ethics may constitute an important first step towards more binding governance measures, they require further articulation as specific, enforceable rights to have any real impact.

Human rights as the baseline

For these and other reasons, there was broad consensus across the consultations held by UN Global Pulse and UN Human Rights[85] that human rights should form the basis of any effective AI governance regime. International human rights law (IHRL) provides a globally legitimate and comprehensive framework for predicting, preventing and redressing the aforementioned risks and harms. As McGregor *et al.* argue, IHRL provides an "organizing framework for the design, development and deployment of algorithms, and identifies the factors that States and businesses should take into consideration in order to avoid undermining, or violating, human rights".[86] Far from being a stand-alone and static set of "rules", this framework "is capable of accommodating other approaches to algorithmic accountability – including technical solutions – and …

81 See Mark Latonero, *Governing Artificial Intelligence: Upholding Human Rights and Dignity*, Data & Society, 2018, arguing that human rights do not tend to be central to national AI strategies, with a few exceptions that include the EU's GDPR and strategy documents issued by the Council of Europe, the Canada and France-led Global Partnership on AI, and the Australian Human Rights Commission.

82 See P. Alston, above note 69, arguing that most AI ethics codes refer to human rights law but lack its substance and that token references are used to enhance the code's claims to legitimacy and universality.

83 Corinne Cath, Mark Latonero, Vidushi Marda and Roya Pakzad, "Leap of FATE: Human Rights as a Complementary Framework for AI Policy and Practice", in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020, available at: https://doi.org/10.1145/3351095.3375665.

84 *Ibid.*

85 Consultations include meetings and workshops held by Global Pulse and UN Human Rights in Geneva, Berlin and Tunis.

86 L. McGregor, D. Murray and V. Ng, above note 5, p. 313.

can grow and be built on as IHRL itself develops, particularly in the field of business and human rights".[87]

The case for IHRL can be broken down into several discrete aspects that make this framework particularly appropriate to the novel risks and harms of AI. Firstly, unlike ethics, IHRL is universal.[88] IHRL offers a common vocabulary and set of principles that can be applied across borders and cultures, ensuring that AI serves shared human values as embodied in the UDHR and other instruments. There is no other common set of moral or legal principles that resonates globally like the UDHR.[89] In a world where technology and data flow almost seamlessly across borders, and where technology cannot be governed effectively within a single jurisdiction, this universal legitimacy is essential.

Secondly, the international human rights regime is binding on States. Specifically, it requires them to put a framework in place that "prevents human rights violations, establishes monitoring and oversight mechanisms as safeguards, holds those responsible to account, and provides a remedy to individuals and groups who claim their rights have been violated".[90] At the international level, the IHRL regime also offers a set of built-in accountability and advocacy mechanisms, including the HRC and the treaty bodies, which have complaints mechanisms and the ability to review the performance of member States; the Special Procedures of the HRC (namely the working groups and Special Rapporteurs), which can conduct investigations and issue reports and opinions;[91] and, increasingly, the International Court of Justice, which has begun to carve out a bigger role for itself in human rights and humanitarian jurisprudence.[92] Moreover, regional human rights mechanisms have assumed a key role in developing the human rights system, including by providing individuals with the opportunity to bring legal actions against perpetrators of human rights violations.[93]

Thirdly, IHRL focuses its analytical lens on the rights holder and duty bearer in a given context, enabling much easier application of principles to real-world situations.[94] Rather than aiming for broad ideals like "fairness", human rights law calls on developers and implementers of AI systems to focus in on who, specifically, will be impacted by the technology and which of their specific fundamental rights will be implicated. This is an intensely pragmatic exercise that involves translating higher ideals into narrowly articulated risks and harms. Relatedly, many human rights accountability mechanisms also enable individuals

87 *Ibid.*

88 "[Human rights] are considered universal, both because they are universally recognised by virtually each country in the world, and because they are universally applicable to all human beings regardless of any individual trait." Nathalie A. Smuha, "Beyond a Human Rights-based Approach to AI Governance: Promise, Pitfalls, Plea", *Philosophy and Technology*, 2020 (forthcoming).

89 *Ibid.*

90 L. McGregor, D. Murray and V. Ng, above note 5, p. 311.

91 *Ibid.*

92 Lyal S. Sunga, "The International Court of Justice's Growing Contribution to Human Rights and Humanitarian Law," The Hague Institute for Global Justice, The Hague, 18 April 2016.

93 UN Human Rights, "Regional Human Rights Mechanisms and Arrangements", available at: www.ohchr.org/EN/Countries/NHRI/Pages/Links.aspx.

94 C. Cath *et al.*, above note 83.

to assert their rights by bringing claims before various adjudicating bodies. Of course, accessing a human rights tribunal and formulating a viable claim is much easier said than done. But at the very least, human rights provide these individuals with the "language and procedures to contest the actions of powerful actors", be they States or corporations.[95]

Fourthly, in defining specific rights, IHRL also defines the harms that need to be avoided, mitigated and remedied.[96] In doing so, it identifies the outcomes that States and other entities – including development and humanitarian actors – can work towards achieving. For example, the UN's Committee on Economic, Social and Cultural Rights has developed standards for "accessibility, adaptability and acceptability" that States should pursue in their social protection programmes.[97]

Finally, human rights law and human rights jurisprudence provide a framework for balancing rights that come into conflict with each other.[98] This is essential when deciding whether to deploy a technological tool that entails both benefits and risks. In these cases, human rights law provides guidance on when and how certain fundamental rights can be restricted – namely, by applying the principles of legality, legitimacy, necessity and proportionality to the proposed AI intervention.[99] In this way, IHRL also helps identify red lines – that is, actions that are out of bounds.[100] This framework would be particularly helpful for

95 Christian van Veen and Corinne Cath, "Artificial Intelligence: What's Human Rights Got to Do With It?", Data & Society, 14 May 2018, available at: https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5.

96 L. McGregor, D. Murray and V. Ng, above note 5.

97 See ESCR Report, above note 44; "Standards of Accessibility, Adaptability, and Acceptability", *Social Protection and Human Rights*, available at: https://socialprotection-humanrights.org/framework/principles/standards-of-accessibility-adaptability-and-acceptability/.

98 Karen Yeung, Andrew Howes and Ganna Pogrebna, "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing", in Markus D. Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford, 2020, noting that IHRL provides a "[s]tructured framework for reasoned resolution of conflicts arising between competing rights and collective interests in specific cases", whereas AI ethics codes offer "little guidance on how to resolve such conflicts".

99 Limitations on a right, where permissible, must be necessary for reaching a legitimate aim and must be in proportion to that aim. They must be the least intrusive option available, and must not be applied or invoked in a manner that would impair the essence of a right. They need to be prescribed by publicly available law that clearly specifies the circumstances under which a restriction may occur. See ESCR Report, above note 44, pp. 10–11. See also N. A. Smuha, above note 88, observing that similar formulas for balancing competing rights are found in the EU Charter, the European Convention of Human Rights, and Article 29 of the UDHR.

100 Catelijne Muller, *The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law*, Ad Hoc Committee on Artificial Intelligence, Strasbourg, 24 June 2020, para. 75, available at: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da.
McGregor *et al.* draw red lines from "the prohibition of arbitrary rights interference as a core principle underpinning IHRL [that is] relevant to all decisions that have the potential to interfere with particular rights". L. McGregor, D. Murray and V. Ng, above note 5, p. 337. For more on the relationship between "arbitrary" and "necessary and proportionate", see UN Human Rights, *The Right to Privacy in the Digital Age: Report of the Office of the United Nations High Commissioner for Human Rights*, UN Doc. A/HRC/27/37, 30 June 2014, para. 21 ff.; UN Human Rights, *The Right to Privacy in the Digital Age: Report of the United Nations High Commissioner for Human Rights*, UN Doc. A/HRC/39/29, 3 August 2018, para. 10.

humanitarian organizations trying to decide if and when a certain AI capability (such as a facial recognition technology) should be avoided entirely.

The need for a balancing framework is arguably evident in most humanitarian applications of AI. The balancing approach has been incorporated into UN Global Pulse's Risks, Harms and Benefits Assessment, which prompts the implementers of an AI or data analytics project not only to consider the privacy risks and likelihood, magnitude and severity/significance of potential harms, but also to weigh these risks and harms against the predicted benefits of the project. IHRL jurisprudence helps guide the use of powerful AI tools in these contexts, dictating that such use is only acceptable so long as it is prescribed by law, in pursuit of a legitimate aim, and is necessary and proportionate to that aim.[101] In pursuing this balance, decision-makers can look to decades of IHRL jurisprudence for insight on how to resolve tensions between conflicting rights, or between the rights of different individuals.[102] Other examples of tools and guidance[103] that incorporate the balancing framework include the International Principles on the Application of Human Rights to Communication Surveillance[104] and the OCHA Guidance Note on data impact assessments.[105]

Gaps in Implementing IHRL: Private sector accountability

One major limitation of IHRL is that it is only binding on States. Individuals can therefore only bring human rights claims vertically – against the State – rather than horizontally – against other citizens, organizations or, importantly, companies.[106] This would seem to be a problem for AI accountability because the

101 IHRL "provides a clear framework for balancing competing interests in the development of technology: its tried and tested jurisprudence requires restrictions to human rights (like privacy or non-discrimination) to be prescribed by law, pursue a legitimate aim, and be necessary and proportionate to that aim. Each term is a defined concept against which actions can be objectively measured and made accountable." Alison Berthet, "Why Do Emerging AI Guidelines Emphasize 'Ethics' over Human Rights?" *OpenGlobalRights*, 10 July 2019, available at: www.openglobalrights.org/why-do-emerging-ai-guidelines-emphasize-ethics-over-human-rights.

102 "Furthermore, to do so, enforcers can draw on previously undertaken balancing exercises, which advances predictability and legal certainty. Indeed, decades of institutionalised human rights enforcement resulted in a rich jurisprudence that can guide enforcers when dealing with the impact of AI-systems on individuals and society and with the tensions stemming therefrom – be it in terms of conflicting rights, principles or interests." N. A. Smuha, above note 88.

103 For further guidance on how to craft a human rights-focused impact assessment, see UN Human Rights, *Guiding Principles on Business and Human Rights*, New York and Geneva, 2011 (UNGPs), available at: www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf; ESCR Report, above note 44.

104 The Principles are available at: www.eff.org/files/necessaryandproportionatefinal.pdf. For background and legal analysis, see Electronic Frontier Foundation and Article 19, *Necessary and Proportionate: International Principles on the Application of Human Rights to Communication Surveillance*, May 2014, available at: www.ohchr.org/Documents/Issues/Privacy/ElectronicFrontierFoundation.pdf.

105 ICRC, Privacy International, UN Global Pulse and OCHA Centre for Humanitarian Data, "Guidance Note: Data Impact Assessments", Guidance Note Series No. 5, July 2020, available at: https://centre.humdata.org/wp-content/uploads/2020/07/guidance_note_data_impact_assessments.pdf. See this Guidance Note for more examples of impact assessments designed for humanitarian contexts.

106 John H. Knox, "Horizontal Human Rights Law", *American Journal of International Law*, Vol. 102, No. 1, 2008, p. 1.

private sector plays a leading role in developing AI and is responsible for the majority of innovation in this field. Of course, States are required under IHRL to incorporate human rights standards into their domestic laws; these, in turn, would regulate the private sector. But we know from experience that this does not always happen, and that even when States *do* incorporate human rights law into their domestic regulations, they are only able to enforce the law within their respective jurisdictions. Yet many major technology companies operate transnationally, including in countries where human rights protections are weaker or under-enforced.

Nonetheless, human rights law has powerful moral and symbolic influence that can shape public debate, sharpen criticism and help build pressure on companies, and human rights responsibilities of companies that are independent from States' ability or willingness to fulfil their own human rights obligations are increasingly recognized.[107] There are a number of mechanisms and levers of pressure by which private companies are incentivized to comply.

Emerging as an international norm for rights-respecting business conduct are the UN Guiding Principles on Business and Human Rights (UNGPs).[108] The UNGPs are conceptualizing the responsibility of businesses to respect human rights along all their business activities, and they call on companies to carry out human rights due diligence in order to identify, address and mitigate adverse impacts on human rights in the procurement, development and operation of their products.[109] A growing chorus of human rights authorities have reiterated that these same obligations apply to algorithmic processing, AI and other emerging digital technologies[110] – most recently, in the UN High Commissioner for Human Rights' report on the use of technologies such as facial recognition in the context of peaceful protests.[111] UN Human Rights is also in the process of developing extensive guidance on the application of the UNGPs to the development and use of digital technologies.[112] A growing number of leading AI companies, such as

107 See I. Ebert, T. Busch and F. Wettstein, above note 41. And see C. van Veen and C. Cath, above note 95, arguing that "[h]uman rights, as a language and legal framework, is itself a source of power because human rights carry significant moral legitimacy and the reputational cost of being perceived as a human rights violator can be very high". For context on algorithmic systems, see Council of Europe, *Recommendation CM/Rec(2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems*, 8 April 2020.

108 UNGPs, above note 103. Pillar I of the UNGPs outlines how States should regulate companies.

109 *Ibid.*, Pillar II. See also UN Human Rights, *Key Characteristics of Business Respect for Human Rights*, B-Tech Foundational Paper, available at: www.ohchr.org/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf.

110 See Council of Europe, *Addressing the Impacts of Algorithms on Human Rights: Draft Recommendation*, MSI-AUT(2018)06rev3, 2018: "Private sector actors engaged in the design, development, sale, deployment, implementation and servicing of algorithmic systems, whether in the public or private sphere, must exercise human rights due diligence. They have the responsibility to respect internationally recognised human rights and fundamental freedoms of their customers and of other parties who are affected by their activities. This responsibility exists independently of States' ability or willingness to fulfil their human rights obligations." See also D. Kaye, above note 38.

111 HRC, *Impact of New Technologies on the Promotion and Protection of Human Rights in the Context of Assemblies, including Peaceful Protests: Report of the UN High Commissioner for Human Rights*, UN Doc. A/HRC/44/24, 24 June 2020.

Element AI, Microsoft and Telefonica, have also begun applying the UNGPs to their AI products.[113]

A second critique of a human rights-based approach to AI is that prioritizing human rights at every stage of the deployment cycle will hinder innovation. There is some truth to this – emphasizing human rights may occasionally delay or even preclude the deployment of a risky product. However, it might also prevent later, even more costly effects of managing the potential fallout of human rights violations.[114] Moreover, the value of a human rights approach is not merely in ensuring compliance but in embedding human rights in the very conception, development and roll-out of a project. Prioritizing human rights at every stage of the development process should therefore reduce the number of instances where a product ends up being too risky to deploy.

The role of ethics

While human rights should set the outer boundaries of AI governance, ethics has a critical role to play in responsible AI governance. Even many ardent advocates of a human rights-based approach to AI acknowledge the reinforcing role that ethical principles can play in augmenting or complementing human rights. In the context of AI, "ethics" typically refers to the so-called FAccT principles: fairness, accountability and transparency (sometimes also called FATE, where the E stands for "ethics").[115] To some, the FAccT approach contrasts with the rigidity of law, eschewing hard-and-fast "rights" in favour of broader consideration of what impact a system will have on society.[116] In this way, ethics is often seen as more adaptable to technological evolution and the modern world; IHRL principles, by contrast, were developed decades ago, long before the proliferation of AI and ML systems.

Yet while there are important distinctions between a human rights-based and an ethics-based approach, our consultations have revealed that the "human rights versus ethics" divide pervading AI policy may in some sense be a false dichotomy.[117] It is worth underlining that human rights and ethics have essentially the same goals. As Access Now has succinctly observed, any

112 UN Human Rights, *The UN Guiding Principles in the Age of Technology*, B-Tech Foundational Paper, available at: www.ohchr.org/Documents/Issues/Business/B-Tech/introduction-ungp-age-technology.pdf.

113 Examples include Microsoft's human rights impact assessment (HRIA) and Google's Celebrity Recognition HRIA; and see Element AI, *Supporting Rights-Respecting AI*, 2019; Telefonica, "Our Commitments: Human Rights," available at: www.telefonica.com/en/web/responsible-business/human-rights.

114 L. McGregor, D. Murray and V. Ng, above note 5.

115 Microsoft has produced a number of publications on its FATE work. See "FATE: Fairness, Accountability, Transparency, and Ethics in AI", available at: www.microsoft.com/en-us/research/group/fate#!publications.

116 C. Cath *et al.*, above note 83.

117 For useful background on the pros and cons of the AI ethics and human rights frameworks, see Business for Social Responsibility (BSR) and World Economic Forum (WEF), *Responsible Use of Technology*, August 2019, p. 7 (arguing that ethics and human rights should be "synergistic").

"unethical" use of AI will also likely violate human rights (and vice versa).[118] That said, human rights advocates are rightly concerned about the phenomenon of "ethics-washing",[119] whereby the makers of technology – often private companies – self-regulate through vague and unenforceable codes of ethics. Technical experts, for their part, are often sceptical that "rigid" human rights law can be adapted to the novel features and risks of harm of AI and ML. While both of these concerns may be valid, these two approaches can actually complement, rather than undermine, each other.

For example, it can take a long time for human rights jurisprudence to develop the specificity necessary to regulate emerging digital technologies, and even longer to apply human rights law as domestic regulation. In such cases where law does not provide clear or immediate answers for AI developers and implementers, ethics can be helpful in filling the gaps;[120] however, this is a role that the interpretation of the existing human rights provisions and case law can play as well. In addition, ethics can raise the bar above the minimum standards set by a human rights framework or help incorporate principles that are not well established by human rights law.[121] For instance, an organization developing AI tools might commit to guaranteeing human oversight of any AI-supported decision – a principle not explicitly stated in any human rights treaty, but one that would undoubtedly reinforce (and implement) human rights.[122] Other organizations seeking to ensure that the economic or material benefits of AI are equally distributed may wish to incorporate the ethical principles of distributive justice[123] or solidarity[124] in their use of AI.

When AI is deployed in development and humanitarian contexts, the goal is not merely to stave off regulatory action or reduce litigation risk through compliance. In fact, there may be little in the way of enforceable regulation or oversight that applies in development and humanitarian contexts. Rather, these actors are seeking to materially improve the lives and well-being of targeted communities. AI that fails to protect the rights of those impacted may instead actively undermine this essential development and humanitarian imperative. For these reasons, development and humanitarian actors are becoming more ambitious in their pursuit of AI that is designed in rights-respecting, ethical ways.[125]

118 Access Now, above note 19.

119 Ben Wagner, "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping?", in Emre Bayamlioglu, Irina Baraliuc, Liisa Janssens and Mireille Hildebrandt (eds), *Being Profiled: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen*, Amsterdam University Press, Amsterdam, 2018.

120 Based on our Geneva consultations. See also Josh Cowls and Luciano Floridi, "Prolegomena to a White Paper on an Ethical Framework for a Good AI Society", June 2018, available at https://papers.ssrn.com/abstract=3198732.

121 *Ibid.*, arguing that ethics and human rights can be mutually enforcing and that ethics can go beyond human rights. See also BSR and WEF, above note 117.

122 Access Now, above note 19, p. 17.

123 BSR and WEF, above note 117.

124 Miguel Luengo-Oroz, "Solidarity Should Be a Core Ethical Principle of AI", *Nature Machine Intelligence*, Vol. 1, No. 11, 2019.

125 See, for example, the UN Global Pulse "Projects" web page, available at: www.unglobalpulse.org/projects/.

Principles and tools

A human rights-based framework will have little impact unless it is operationalized in the organization's day-to-day work. This requires developing tools and mechanisms for the design and operation of AI systems at every stage of the product lifecycle – and in every application. This section will introduce several such tools that were frequently endorsed as useful or essential in our consultations and interviews.

In his Strategy on New Technology, the UN Secretary-General noted the UN's commitment to both "deepening [its] internal capacities and exposure to new technologies" and "supporting dialogue on normative and cooperation frameworks".[126] The Secretary-General's High-Level Panel on Digital Cooperation made similar recommendations, calling for enhanced digital cooperation to develop standards and principles of transparency, explainability and accountability for the design and use of AI systems.[127] There has also been some early work within the UN and other international organizations on the development of ethical principles and practical tools.[128]

Internal AI principles

Drafting a set of AI principles, based on human rights but augmented by ethics, can be helpful in guiding an organization's work in this area – and, ultimately, in operationalizing human rights. The goal of such a "code" would be to provide guidance to every member of the team in order to ensure that human needs and rights are constantly in focus at every stage of the AI life cycle. More importantly, the principles could also undergird any compliance tools or mechanisms that the organization subsequently develops, including risk assessments, technical standards and audit procedures. These principles should be broad enough that they can be interpreted as guidance in novel situations – such as the emergence of

126 UN, *UN Secretary-General's Strategy on New Technologies*, September 2018, available at: www.un.org/en/newtechnologies/.

127 High-Level Panel on Digital Cooperation, *The Age of Digital Interdependence: Report of the UN Secretary-General's High-Level Panel on Digital Cooperation*, June 2019 (High-Level Panel Report), available at: https://digitalcooperation.org/wp-content/uploads/2019/06/DigitalCooperation-report-web-FINAL-1.pdf.

128 UNESCO issued a preliminary set of AI principles in 2019 and is in the process of drafting a standard-setting instrument for the ethics of AI. A revised first draft of a recommendation was presented in September 2020. Other entities, including the Organization for Economic Cooperation and Development (OECD) and the European Commission, have released their own sets of principles. OECD, *Recommendation of the Council on Artificial Intelligence*, 21 May 2019; European Commission, *Ethics Guidelines for Trustworthy AI*, 8 April 2019, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. At the Council of Europe, the Committee of Ministers has adopted Recommendation CM/Rec(2020)1, above note 107. The Council of Europe is also investigating the possibility of adopting a legal framework for the development, design and application of AI, based on the Council of Europe's standards on human rights, democracy and the rule of law; see Council of Europe, "CAHAI – Ad Hoc Committee on Artificial Intelligence", available at: www.coe.int/en/web/artificial-intelligence/cahai.

a technological capacity not previously anticipated – but specific enough that they are actionable in the organization's day-to-day work.

The UN Secretary-General has recommended the development of AI that is "trustworthy, human-rights based, safe and sustainable and promotes peace".[129] While an organization's guiding principles should be anchored in these four pillars, there is potential for substantial variation depending on the nature and context of an organization's work. Our consultations suggested that an effective set of principles would be rooted in human rights principles – interpreted or adapted into the AI context – along with complementary ethics principles which provide flexibility to address new challenges that arise as the technology develops.

While suggesting a complete set of principles is beyond the scope of this article, there is an emerging consensus that certain challenges deserve special attention. Three of these challenges – non-discrimination, transparency and explainability, and accountability – will be discussed in more detail below. Other commonly cited principles include human-centred design, human control or oversight, inclusiveness and diversity, privacy, technical robustness, solidarity, sustainability, democracy, good governance, awareness and literacy, *ubuntu*, and banning lethal autonomous weapons systems. A table of the principles that appear most frequently in AI ethics guidelines, based on a 2019 analysis by René Clausen Nielsen of UN Global Pulse, is shown in Figure 1.

Of course, adopting a code of ethics does not, in itself, guarantee that an organization will prioritize human rights in developing AI tools. These principles must be operationalized to have any real impact. The foundational step in this operationalization should be a binding policy commitment to human rights adopted at the executive level. Moreover, the implementation of the commitment needs to be accompanied and guided by appropriate management and oversight structures and processes. Further steps that could be taken would include the translation into technical standards that allow for quality control and auditing. For example, some experts have proposed technical standards for algorithmic transparency, or implementing rules that automatically detect potentially unfair outcomes from algorithmic processing.[130] Moreover, the code would have to be developed in a way that facilitates and informs the creation of concrete tools and procedures for mitigating human rights risks at every stage of the AI life cycle. For example, it could be an element of the human rights due diligence tools described below.

129 Secretary-General's Roadmap, above note 1, para. 88. See also Recommendation 3C of the High-Level Panel Report, above note 127, pp. 38–39, which reads: "[A]utonomous intelligent systems should be designed in ways that enable their decisions to be explained and humans to be accountable for their use. Audits and certification schemes should monitor compliance of artificial intelligence (AI) systems with engineering and ethical standards, which should be developed using multi-stakeholder and multilateral approaches. Life and death decisions should not be delegated to machines. … [E]nhanced digital cooperation with multiple stakeholders [is needed] to think through the design and application of … principles such as transparency and non-bias in autonomous intelligent systems in different social settings."

130 See A. Beduschi, above note 44, arguing for technical standards that "incorporat[e] human rights rules and principles".

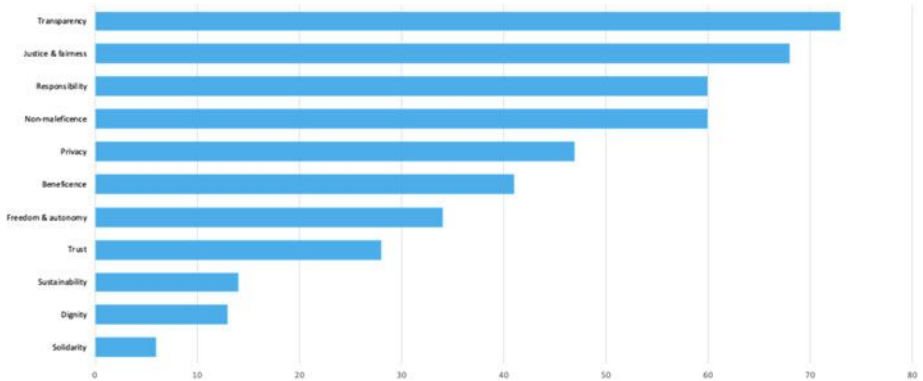


Figure 1. Ethical principles identified in existing AI guidelines. Analysis by René Clausen Nielsen, UN Global Pulse, based on A. Jobin, M. Ienca, and E. Vayena, above note 36.

While each of the aforementioned principles may indeed be essential, our consultations focused on three interrelated ethical principles that are firmly anchored in IHRL and require further elaboration and more careful implementation: non-discrimination, transparency and explainability, and accountability. Organizations using AI for humanitarian aid need to develop policies and mechanisms to ensure that these systems do not have discriminatory impact; that their decisions are capable of being understood and explained, at least to a level adequate for the risks involved; and that there is accountability for harms associated with their operation. This is especially crucial in operations where AI is used to support the vulnerable. While these are not the only governance challenges associated with AI, they offer a starting point for conversations about what makes AI different from other technologies and why it poses unique challenges for human rights.[131]

Non-discrimination

One of the key principles that humanitarian organizations need to ensure is non-discrimination. AI systems tend to reflect existing power relations and dynamics, and their deployment may risk creating new inequalities and dependencies or entrenching those that are already present. Therefore, it is important to note as a starting point that any decision to develop and deploy an AI system in a humanitarian context needs to take a holistic view of how this system will operate in the target environment and how it will affect people's lives, with a strong focus on those in vulnerable positions.

A few solutions were suggested during our consultations and research. Above all, diversity and inclusion are absolutely critical to ensuring that AI

131 For a breakdown of how individual UDHR rights and principles are implicated by the use of AI systems, see Access Now, above note 19.

systems are used in a non-discriminatory manner. This principle should pervade every aspect of AI development and use, from incorporating diverse perspectives in the teams designing and deploying AI systems to ensuring that training data is representative of target populations. Meaningful comprehensive consultations with representatives of affected groups are essential for preventing exclusionary and discriminatory effects of deployed AI solutions.

Second, capacity-building and knowledge sharing are urgently needed. Practitioners that we consulted raised the need for a good-faith intermediary to coordinate knowledge sharing across the world and provide practical advice on how to address bias questions. Such an entity could compile best practices in the AI for development and humanitarian fields and identify areas where experimentation with AI may need to be barred. The intermediary could serve as a discovery resource for organizations using AI that do not know how to interrogate their AI systems and/or lack the resources to do so. Many organizations need someone who can help them troubleshoot potential discrimination concerns by packaging the data and interrogating possible bias.

Third, given that the risks of unwanted discriminatory impact can never be reduced to zero, certain areas may be deemed too risky or uncertain for AI systems to play a central role (e.g., making final determinations). These may include criminal justice, social welfare and refugee/asylum processing, where various pilot projects and cases studies have already flagged problematic discriminatory implications with direct impact on human lives. Our consultations suggested that, in such cases, organizations could make use of red-line bans and moratoria.[132]

Transparency and explainability

Transparency and explainability of AI systems are prerequisites to accountability. However, full transparency into many ML and DL systems is not possible.[133] When a model is unsupervised, it will be capable of classifying, sorting or ranking the data based on a set of rules or patterns that it identifies, and the humans who created this model will not always be able to tell how or why the resulting analysis was arrived at.[134] This means that, in order to make use of this technology, organizations will need to carefully assess if and how these largely

132 A growing number of jurisdictions have issued bans on facial recognition technology, or on the use of such technology in criminal justice contexts. However, some organizations have been more hesitant to embrace red lines. See Chris Klöver and Alexander Fanta, "No Red Lines: Industry Defuses Ethics Guidelines for Artificial Intelligence", trans. Kristina Penner, *Algorithm Watch*, 9 April 2019, available at: https://algorithmwatch.org/en/industry-defuses-ethics-guidelines-for-artificial-intelligence/ (where one source blames the absence of red lines in the EU's ethics guidelines on industry pressure).

133 "Although total explainability of ML-based systems is not currently possible, developers can still provide valuable information about how a system works. Publish easy-to-understand explainers in the local language. Hold community meetings to explain the tool and allow community members to ask questions and provide feedback. Take care to consider literacy levels and the broader information ecosystem. An effective public educational process utilizes the existing ways in which a community receives and shares information, whether that be print, radio, word of mouth, or other channels." L. Andersen, above note 42,.

134 See "Common ML Problems", above note 24.

obscure or unexplainable systems can be used in a way that augments, rather than undermines, human rights.

There are at least two different types of transparency, both of which are essential to ensuring accountability. The first is technical transparency – that is, transparency of the models, algorithms and data sets that comprise an AI system. The second is organizational transparency, which deals with questions such as whether an AI system is being used for a particular purpose, what kind of system or capability is being used, who funded or commissioned the system and for what purpose, who built it, who made specific design decisions, who decided where to apply it, what the outputs were, and how those outputs were used.[135] While the two are related, each type of transparency requires its own set of mechanisms and policies to ensure that a system is transparent and explainable.

To address and ensure the principle of transparency, our consultations and research supported the idea of human-in-the-loop as a foundational principle. Human-in-the-loop is the practice of embedding a human decision-maker into every AI-supported decision.[136] This means that, even in cases where DL is being leveraged to generate powerful predictions, humans are responsible for operationalizing that prediction and, to the extent possible, auditing the system that generated it.[137] In other words, humans hold ultimate responsibility for making decisions, even when they rely heavily on output or analysis generated by an algorithm.[138] However, effective human-in-the-loop requires more than just having a human sign off on major decisions. Furthermore, organizations also need to scrutinize how human decision-makers interact with AI systems and ensure that human decision-makers have meaningful autonomy within the organizational context.[139]

135 See ESCR Report, above note 44, para. 52, arguing that the knowledge and understanding gap between the public and decision-makers can be "a particular problem in the context of the automated decision-making processes that rely on artificial intelligence"; that "[c]omprehensive, publicly available information is important to enable informed decision-making and the relevant consent of affected parties; and that "[r]egulations requiring companies to disclose when artificial intelligence systems are used in ways that affect the exercise of human rights and share the results of related human rights impact assessments may also be a helpful tool". See also L. McGregor, D. Murray and V. Ng, above note 5, arguing that transparency includes why and how the algorithm was created; the logic of the model or overall design; the assumptions underpinning the design process; how performance is monitored; how the algorithm itself has changed over time; the factors relevant to the algorithm's functioning; and the level of human involvement.

136 Sam Ransbotham, "Justifying Human Involvement in the AI Decision-Making Loop", *MIT Sloan Management Review*, 23 October 2017, available at: https://sloanreview.mit.edu/article/justifying-human-involvement-in-the-ai-decision-making-loop/.

137 See L. McGregor, D. Murray and V. Ng, above note 5, arguing that human-in-the-loop acts as a safeguard, ensuring that the algorithmic system supports but does not make the decision.

138 "AI is most exciting when it can both absorb large amounts of data and identify more accurate correlations (diagnostics), while leaving the causational conclusions and ultimate decision-making to humans. This human-machine interaction is particularly important for social-impact initiatives, where ethical stakes are high and improving the lives of the marginalized is the measure of success." Hala Hanna and Vilas Dhar, "How AI Can Promote Social Good", World Economic Forum, 24 September 2019, available at: www.weforum.org/agenda/2019/09/artificial-intelligence-can-have-a-positive-social-impact-if-used-ethically/.

139 One hypothetical raised by a participant at our Geneva event was as follows: a person in a government office is using automated decision-making to decide whose child gets taken away. The algorithm gives

Accountability

Accountability enables those affected by a certain action to demand an explanation and justification from those acting and to obtain adequate remedies if they have been harmed.[140] Accountability can take several different forms.[141] Technical accountability requires auditing of the system itself. Social accountability requires that the public have been made aware of AI systems and have adequate digital literacy to understand their impact. Legal accountability requires having legislative and regulatory structures in place to hold those responsible for bad outcomes to account.

Firstly, there is a strong need for robust oversight mechanisms to monitor and measure progress on accountability mechanisms across organizations and contexts. Such a mechanism could be set up at the national, international or industry level and would need to have substantial policy, human rights and technical capacity. Another idea is for this or another specialized entity to carry out certification or "kitemarking" of AI tools and systems, whereby those with high human rights scores (based on audited practices) are "certified" both to alert consumers and, potentially, open the door to partnerships with governments, international organizations, NGOs and other organizations committed to accountable, rights-respecting AI.[142]

Secondly, while legal frameworks develop, self-regulation will continue to play a significant role in setting standards for how private companies and other organizations operate. However, users and policy-makers could monitor companies through accountability mechanisms and ensure that industry is using its full capacity to ensure human rights.

Thirdly, effective remedies are key elements of accountable AI frameworks. In particular, in the absence of domestic legal mechanisms, remedies can be provided at the company or organization level through internal grievance mechanisms.[143] Whistle-blowing is also an important tool for uncovering abuses and promoting accountability, and proper safeguards and channels should be put in place to encourage and protect whistle-blowers.

Finally, ensuring good data practices is a critical component of AI accountability. Our consultations revealed several mechanisms for data accountability, including quality standards for good data and mechanisms to improve access to quality data, such as mandatory data sharing.

a score of "7". How does this score influence the operator? Does it matter if they're having a good or bad day? Are they pressured to take the score into consideration, either institutionally or interpersonally (by co-workers)? Are they personally penalized if they ignore or override the system?

140 See Edward Rubin, "The Myth of Accountability and the Anti-administrative Impulse", *Michigan Law Review*, Vol. 103, No. 8, 2005.

141 See UN Human Rights, above note 68, outlining the novel accountability challenges raised by AI.

142 High-Level Panel Report, above note 127, Recommendation 3C, pp. 38–39.

143 UNGPs, above note 103, para. 29: "To make it possible for grievances to be addressed early and remediated directly, business enterprises should establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted."

Human rights due diligence tools

It is increasingly recognized that human rights due diligence (HRDD) processes, conducted throughout the life cycle of an AI system, are indispensable for identifying, preventing and mitigating human rights risks linked to the development and deployment of AI systems.[144] Such processes can be helpful in determining necessary safeguards and in developing effective remedies when harm does occur. HRDD gives a rights-holder perspective a central role. Meaningful consultations with external stakeholders, including civil society, and with representatives of potentially impacted individuals and groups, in order to avoid project-driven bias, are essential parts of due diligence processes.[145]

Human rights impact assessments

In order for States, humanitarian organizations, businesses and other actors to meet their respective responsibilities under IHRL, they need to identify human rights risks stemming from their actions. HRDD commonly builds on a human rights impact assessment (HRIA) for identifying potential and actual adverse impacts on human rights related to actual and planned activities.[146] While the HRIA is a general tool, recommended for all companies and sectors by the UNGPs, organizations are increasingly applying the HRIA framework to AI and other emerging digital technologies[147]. The Secretary-General's Roadmap announced plans for UN Human Rights to develop system-wide guidance on HRDD and impact assessments in the use of new technologies.[148] HRIAs should ideally assist practitioners in identifying the impact of their AI interventions, considering such

144 See I. Ebert, T. Busch and F. Wettstein, above note 41. See also Committee on the Elimination of Racial Discrimination, *General Recommendation No. 36 on Preventing and Combating Racial Profiling by Law Enforcement Officials*, UN Doc. CERD/C/GC/36, 17 December 2020, para. 66: "States should encourage companies to carry out human rights due diligence processes, which entail: (a) conducting assessments to identify and assess any actual or potentially adverse human rights impacts; (b) integrating those assessments and taking appropriate action to prevent and mitigate adverse human rights impacts that have been identified; (c) tracking the effectiveness of their efforts; and (d) reporting formally on how they have addressed their human rights impacts."

145 See ESCR Report, above note 44, para. 51. The UNGPs make HRDD a key expectation of private companies. The core steps of HRDD, as provided for by the UNGPs, include (1) identifying harms, consulting with stakeholders, and ensuring public and private actors also conduct assessments (if the system will be used by a government entity); (2) taking action to prevent and mitigate harms; and (3) being transparent about efforts to identify and mitigate harms. Access Now, above note 19, pp. 34–35.

146 D. Kaye, above note 38, para. 68, noting that HRIAs "should be carried out during the design and deployment of new artificial intelligence systems, including the deployment of existing systems in new global markets".

147 Danish Institute for Human Rights, "Human Rights Impact Assessment Guidance and Toolbox", 25 August 2020, available at: www.humanrights.dk/business/tools/human-rights-impact-assessment-guidance-toolbox.

148 "To address the challenges and opportunities of protecting and advancing human rights, human dignity and human agency in a digitally interdependent age, the Office of the United Nations High Commissioner for Human Rights will develop system-wide guidance on human rights due diligence and impact assessments in the use of new technologies, including through engagement with civil society, external experts and those most vulnerable and affected." Secretary-General's Roadmap, above note 1, para. 86.

factors as the severity and type of impact (directly causing, contributing to, or directly linked), with the goal of guiding decisions on whether to use the tool (and if so, how) or not.[149]

Other potentially relevant tools for identifying a humanitarian organization's adverse impact on human rights include data protection impact assessments, which operationalize best practices in data privacy and security; and algorithmic impact assessments, which aim to mitigate the unique risks posed by algorithms. Some tools are composites, such as Global Pulse's Risks, Harms and Benefits Assessment, which incorporates elements found in both HRIAs and data protection impact assessments.[150] This tool allows every member of a team – including technical and non-technical staff – to assess and mitigate risks associated with the development, use and specific deployment of a data-driven product. Importantly, the Risks, Harms and Benefits Assessment provides for the consideration of a product's benefits – not only the risks – and hence reflects the imperative of balancing interests, as provided for by human rights law.

The advantage of these tools is that they are adaptable to technological change. Unlike regulatory mechanisms or red-line bans, HRDD tools are not limited to specific technologies or technological capacities (e.g., facial recognition technology) but rather are designed to "[pre-empt] new technological capabilities and [allow] space for innovation".[151] In addition, well-designed HRDD tools recognize that context specificity is key when assessing human rights risk, hence the need for a case-specific assessment. Regardless of which tool, or combination of tools, makes the most sense in a given situation, it will be necessary to ensure that the assessment has been designed or updated to accommodate AI-specific risks. It may also be useful to adapt tools to specific development and humanitarian sectors, such as public health or refugee response, given the unique risks that are likely to arise in those areas.

It is critical to emphasize that HRIAs should be part of the wider HRDD process whereby identified risks and impacts are effectively mitigated and addressed in a continuous process. The quality of an HRDD process will increase when "knowing and showing" is supported by governance arrangements and leadership actions to ensure that a company's policy commitment to respecting human rights is "embedded from the top of the business enterprise through all its functions, which otherwise may act without awareness or regard for human rights".[152] HRDD should be carried out at all stages of the product cycle and should be used by all parties involved in a project. Equally important is that this framework involves the entire organization – from data scientists and engineers to lawyers and project managers – so that diverse expertise informs the HRDD process.

149 C. Cath *et al.*, above note 83.
150 UN Global Pulse, "Risks Harms and Benefits Assessment", available at: www.unglobalpulse.org/policy/risk-assessment/.
151 Element AI, above note 113, p. 9.
152 UNGPs, above note 103, Commentary to Principle 16, p. 17.

Explanatory models

In addition, organizations could make use of explanatory models for any new technological capability or application.[153] The purpose of an explanatory model is to require technical staff, who better understand how a product works, to explain the product in layman's terms to their non-technical colleagues. This exercise serves both to train data scientists and engineers to think more thoroughly about the inherent risks in what they are building, and to enable non-technical staff – including legal, policy and project management teams – to make an informed decision about whether and how to deploy it. In this way, explanatory models could be seen as a precursor to the risk assessment tools described above.

Due diligence tools for partnerships

An important caveat to the use of these tools is that they are only effective if applied across every link in the AI design and deployment chain, including procurement. Many organizations innovating in this field rely on partnerships with technology companies, governments and civil society organizations in order to build and deploy their products. To ensure proper human rights and ethical standards, it is important that partnerships that support humanitarian and development missions are adequately vetted. The challenge in the humanitarian and development sectors is that most due diligence tools and processes do not (yet) adequately cover AI-related challenges. To avoid potential risks of harm, such procedures and tools need to take into account the technological challenges involved and ensure that partners, particularly private sector actors, are committed to HRDD best practices, human rights and ethical standards. UN Global Pulse's Risks, Harms and Benefits Assessment tool is one example of this.[154]

Moreover, because of the risks that may arise when AI systems are used by inadequately trained implementers, organizations need to be vigilant about ensuring downstream human rights compliance by all implementing partners. As UN Human Rights has observed, most human rights harms related to AI "will manifest in product use", whether intentionally – for instance, an authoritarian government abusing a tool to conduct unlawful surveillance – or inadvertently, through unanticipated discrimination or user error. This means an AI developer cannot simply hand off a tool to a partner with instructions to use it judiciously. That user, and any third party with whom they partner, must commit to thorough, proactive and auditable HRDD through the tool's life cycle.

153 Participants at our Geneva consultations used the term "explanatory models", though this is not yet a widely used term.

154 UN Global Pulse, above note 150. See also OCHA, "Guidance Note: Data Responsibility in Public-Private Partnerships", 2020, available at: https://centre.humdata.org/guidance-note-data-responsibility-in-public-private-partnerships/.

Public engagement

An essential component of effective HRDD is engagement with the populations impacted by an AI tool. Humanitarian organizations should prioritize engagement with rights holders, affected populations, civil society and other relevant stakeholders in order to obtain a comprehensive, nuanced understanding of the needs and rights of those potentially impacted. This requires proactive outreach, including public consultations where appropriate, and also making available accessible communication channels for affected individuals and communities. As Special Rapporteur David Kaye has recommended, "public consultations and engagement should occur prior to the finalization or roll-out of a product or service, in order to ensure that they are meaningful, and should encompass engagement with civil society, human rights defenders and representatives of marginalized or underrepresented end users". In some cases, where appropriate, organizations may choose to make the results of these consultations (along with HRIAs) public.[155]

Audits

Development and humanitarian organizations can ensure that AI tools – whether developed in-house or by vendors – are externally and independently reviewed in the form of audits.[156] Auditability is critical to ensuring transparency and accountability, while also enabling public understanding of, and engagement with, these systems. While private sector vendors are traditionally resistant to making their products auditable – citing both technical feasibility and trade-secret concerns – numerous models have been proposed that reflect adequate compromises between these concerns and the imperative of external transparency.[157] Ensuring and enabling auditability of AI systems would ultimately be the domain of government regulators and private sector developers, and development and humanitarian actors could promote and encourage its application and adoption.[158] For example, donors or implementers could make auditability a prerequisite for grant eligibility.

155 D. Kaye, above note 68, para. 68.
156 *Ibid.*, para. 55.
157 "Private sector actors have raised objections to the feasibility of audits in the AI space, given the imperative to protect proprietary technology. While these concerns may be well founded, the Special Rapporteur agrees … that, especially when an AI application is being used by a public sector agency, refusal on the part of the vendor to be transparent about the operation of the system would be incompatible with the public body's own accountability obligations." *Ibid.*, para. 55.
158 "Each of these mechanisms may face challenges in implementation, especially in the information environment, but companies should work towards making audits of AI systems feasible. Governments should contribute to the effectiveness of audits by considering policy or legislative interventions that require companies to make AI code auditable, guaranteeing the existence of audit trails and thus greater opportunities for transparency to individuals affected." *Ibid.*, para. 57.

Other institutional mechanisms

There are several institutional mechanisms that can be put in place to ensure that human rights are encoded into an organization's DNA. One principle that has already been discussed is human-in-the-loop, whereby human decision-makers are embedded in the system to ensure that no decisions of consequence are made without human oversight and approval. Another idea would be to establish an AI human rights and ethics review board, which would serve a purpose analogous to the review boards used by academic research institutions.[159] The board, which would ideally be composed of both technical and non-technical staff, would be required to review and sign off on any new technological capacity – and ideally, any novel deployment of that capacity – prior to deployment. In order to be effective as a safeguard, the board would need real power to halt or abort projects without fear of repercussion. Though review boards could make use of the HRDD tools introduced above, their review of a project would constitute a separate, higher-level review than the proactive HRDD that should be conducted at every stage of the AI life cycle. Entities should also consider opening up to regular audits of their AI practices and make summaries of these reports available to their staff, and, where appropriate, to the public. Finally, in contexts where the risks of a discriminatory outcome include grave harm to individuals' fundamental rights, the use of AI may need to be avoided entirely – including through red-line bans.

Capacity-building and knowledge sharing

The challenge of operationalizing human rights and ethical principles in the development of powerful and unpredictable technology is far beyond the capabilities of a single organization. There is an urgent need for capacity-building, especially in the public and NGO sectors. This is true both of organizations deploying AI and those charged with overseeing it. Many data protection authorities, for instance, may lack the resources and capacity to take on this challenge in a competent and comprehensive way.[160] Humanitarian agencies may need help applying existing laws and policies to AI and identifying gaps that need to be filled.[161] In addition, the staff at organizations using AI may need to expand training and education in the ethical and human rights dimensions of AI and the technical operations of systems, in order to ensure trust in the humans designing and operating these systems (as opposed to just the system itself).

AI governance is a fundamentally transnational challenge, so in addition to organization-level capacity-building, effective AI governance will require international cooperation. At the international level, a knowledge-sharing portal

159 Based on our consultations.
160 Based on our consultations.
161 Element AI, above note 113.

operated by traditional actors like the UN, and/or by technical organizations like the Institute of Electrical and Electronics Engineers, could serve as a resource for model HRDD tools, technical standards and other best practices.[162] At the country level, experts have suggested that governments create an "AI ministry" or "centre of expertise" to coordinate efforts related to AI across the government.[163] Such an entity would allow each country to establish governance frameworks that are appropriate for the country's cultural, political and economic context.

Finally, a key advantage of the human rights framework is the existence of accountability and advocacy mechanisms at the international level. Organizations should look to international human rights mechanisms, including the relevant HRC working groups and Special Rapporteurs, for exploration and articulation of the emerging risks posed by AI and best practices for mitigating them.[164]

Conclusion

As seen in various contexts, including the ongoing COVID-19 pandemic, AI may have a role to play in supporting humanitarian missions, if developed and deployed in an inclusive and rights-respecting way. To ensure that the risks of these systems are minimized, and their benefits maximized, human rights principles should be embedded from the start. In the short term, organizations can take several critical steps. First, an organization developing or deploying AI in humanitarian contexts could develop a set of principles, based in human rights and supplemented by ethics, to guide its work with AI. These principles should respond to the specific contexts in which the organization works and may vary from organization to organization.

In addition, diversity and inclusivity are absolutely critical to preventing discriminatory outcomes. Diverse teams should be involved in an AI project from the earliest stages of development all the way through to implementation and follow-up. Further, it is important to implement mechanisms that guarantee adequate levels of both technical and organizational transparency. While complete technical transparency may not always be possible, other mechanisms – including explanatory models – can help educate and inform implementers, impacted populations and other stakeholders about the benefits and risks of an AI intervention, thereby empowering them to provide input and perspective on whether and how AI should be used and also enabling them to challenge the ways in which AI is used.[165] Ensuring that accountability mechanisms are in place is also key, both for those working on systems internally and for those

162 Several UN processes that are under way may serve this purpose, including UNESCO's initiative to create the UN's first standard-setting instrument on AI ethics, and the UN Secretary-General's plans to create a global advisory body on AI cooperation.

163 Element AI, above note 113.

164 See M. Latonero, above note 81, calling for UN human rights investigators and Special Rapporteurs to continue researching and publicizing the human rights impacts of AI systems.

165 Access Now, above note 19.

potentially impacted by an AI system. More broadly, engagement with potentially impacted individuals and groups, including through public consultations and by facilitating communication channels, is essential.

One of the foremost advantages of basing AI governance in human rights is that the basic components of a compliance toolkit already (mostly) exist. Development and humanitarian practitioners should adapt and apply established HRDD mechanisms, including HRIAs, algorithmic impact assessments, and/or UN Global Pulse's Risks, Harms and Benefits Assessment. These tools should be used at every stage of the AI life cycle, from conception to implementation.[166] Where it becomes apparent that these tools are inadequate to accommodate the novel risks of AI systems, especially as these systems develop more advanced capabilities, they can be evaluated and updated.[167] In addition, organizations could demand similar HRDD practices from private sector technology partners and refrain from partnering with vendors whose human rights compliance cannot be verified.[168] Practitioners should make it a priority to engage with those potentially impacted by a system, from the earliest stages of conception through implementation and follow-up. To the extent practicable, development and humanitarian practitioners should ensure the auditability of their systems, so that decisions and processes can be explained to impacted populations and harms can be diagnosed and remedied. Finally, ensuring that a project uses high-quality data and that it follows best practices for data protection and privacy is necessary for any data-driven project.

166 OCHA, above note 154.
167 N. A. Smuha, above note 88.
168 For more guidance on private sector HRDD, see UNGPs, above note 19, Principle 17.

International Review of the Red Cross (2020), **102** (913), 181–198.
Digital technologies and war
doi:10.1017/S1816383121000047

IRRC

Freedom of assembly under attack: General and indiscriminate surveillance and interference with internet communications

Ilia Siatitsa*
Dr Ilia Siatitsa is a Programme Director and Legal Officer at Privacy International.

Abstract

Every day across the world, as people assemble, demonstrate and protest, their pictures, their messages, tweets and other personal information are amassed without adequate justification. Arguing that they do so in order to protect assemblies, governments deploy a wide array of measures, including facial recognition, fake mobile towers and internet shutdowns. These measures are primarily analyzed as interferences with the right to privacy and freedom of expression, but it is argued here that protest and other assembly surveillance should also be understood as an infringement of freedom of assembly. This is necessary not only to preserve the distinct nature of freedom of assembly that protects collective action, but also to allow for better regulation of surveillance and interference with internet communications during assemblies.

Keywords: freedom of assembly, protest, digital age, mass surveillance, interference, internet communications.

* Many thanks to Valentina Cadelo and Tomaso Falchetta for their input on the latest version of this article. The views expressed in this article reflect those of the author.

Introduction

The ability to assemble, dissent and protest peacefully is a key element in every society, democratic or otherwise.[1] In 2019 alone, there were more than 100 protests in numerous countries around the globe.[2] Digital technologies have to a certain degree enabled and facilitated these movements as they have been used to coordinate conversations, raise awareness, encourage participation and generate support.[3] At the same time, these same technologies and other means have been increasingly used to surveil and suppress such movements.

A peaceful assembly, including the right to protest,[4] is understood here as "a gathering of persons for a purpose such as expressing oneself, conveying a position on a particular issue or exchanging ideas".[5] The emphasis here is put on the collective exercise of an individual right irrespective of the means used to that end, whether social media is the main platform of expression or assembling physically.[6] The #MeToo movement is an example of the use of the online space to mobilize women's activities and whole populations on a global scale.[7] While the boundaries of when an online campaign is or becomes part of an assembly will depend on the particularities of each specific case, it is better to retain an inclusive approach at a definitional level.[8] Certain forms of expression online, including online protests, will also be protected by freedom of assembly, while others may primarily enjoy other human rights protections, such as freedom of expression.[9]

The close connection between surveillance and interference with internet communications and freedom of assembly has been highlighted in, *inter alia*, the 2020 report of the United Nations (UN) High Commissioner for Human Rights.

1 African Charter on Human and Peoples' Rights, 1520 UNTS 245, 27 June 1981 (ACHPR), Art. 11; American Convention on Human Rights, San José, 22 November 1969 (ACHR), Art. 15; International Covenant on Civil and Political Rights, 999 UNTS 171, 16 December 1966 (ICCPR), Art. 21; European Convention for the Protection of Human Rights and Fundamental Freedoms (as amended by subsequent protocols), CETS No. 5, Rome, 4 November 1950 (ECHR), Art. 11; Universal Declaration of Human Rights, UNGA Res. 217A, 10 December 1948 (UDHR), Art. 20(1).

2 UN Office of the High Commissioner for Human Rights (UN Human Rights), "Press Briefing Note on Protests and Unrest around the World", 25 October 2019, available at: www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25204&LangID=E (all internet references were accessed in January 2021).

3 See, among many, Zeynep Tufekci, *Twitter and Tear Gas: The Power and Fragility of Networked Protest*, Yale University Press, New Haven, CT, 2017, p. 29.

4 "Protest" and "assembly" are used hereinafter almost interchangeably, though they do not have the exact same meaning. On freedom of assembly, see the sources cited in above note 1.

5 Human Rights Committee, General Comment No. 37, "Article 21: Right of Peaceful Assembly", UN Doc. CCPR/C/GC/37, 27 July 2020 (General Comment 37), para. 12.

6 While the physical gathering of persons used to be considered a key component of an assembly, in the digital age, there is an increasing consensus that assembly includes all forms of collective expression, even if occurring only in the digital/online space. See *ibid.*, para. 13; Clément Voule, *Rights to Freedom of Peaceful Assembly and of Association: Report of the Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association*, UN Doc. A/HRC/41/41, 17 May 2019.

7 C. Voule, above note 6, para. 23.

8 See Michael Hamilton, "The Meaning and Scope of 'Assembly' in International Human Rights Law", *International and Comparative Law Quarterly*, Vol. 69, No. 3, 2020, p. 527.

9 See further below on the relationship between assembly and expression. See also *ibid.*, p. 528.

The report focused on the "impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests".[10] However, the accent has primarily been on how digital technologies are used to interfere with privacy and expression, which as a result negatively affects the right to assemble. There has been less analysis of direct interference with freedom of assembly itself.

It is argued here that there is further room to analyze how surveillance and other interferences with internet communications, carried out by governments, may directly infringe on people's freedom of peaceful assembly. This understanding not only enables us to preserve the distinct nature of freedom of assembly as a right that protects collective action, but also allows for better regulation of surveillance and internet communications interference in protests and other forms of assembly.[11]

In the following pages, the author provides a brief overview of common protest surveillance and interference with internet communications; outlines the close relationship between the right to privacy, freedom of expression and freedom of assembly; and provides a legal analysis on why some of those measures directly infringe on freedom of assembly.

Surveillance of peaceful assembly and interference with internet communications

People take over the streets and online spaces in order to assemble for a collective cause and to show solidarity, but also to criticize or protest against their government's policies and measures. State authorities have been responding with the deployment of technology-enabled surveillance, censorship and violent oppression. This article focuses on the first two.[12] Governments have been invoking their positive obligations to protect freedom of assembly, as well as their prerogative to limit protests in the name of public order and national security, as justifications to impose general and indiscriminate surveillance and interference with internet communications.[13]

10 See, among others, UN Human Rights, *Impact of New Technologies on the Promotion and Protection of Human Rights in the Context of Assemblies, including Peaceful Protests: Report of the United Nations High Commissioner for Human Rights*, UN Doc. A/HRC/44/24, 24 June 2020. But even earlier, see HRC Res. 15/21, "The Rights to Freedom of Peaceful Assembly and of Association", 30 September 2010; Maina Kiai and Jeff Vize, "Three Years after Tunisia: Thoughts and Perspectives on the Rights to Freedom of Assembly and Association from United Nations Special Rapporteur Maina Kiai", *Journal of Global Ethics*, Vol. 10, No. 1, 2014.

11 The present article does not cover violent assemblies or situations described by the International Committee of the Red Cross as "other situations of violence". Nor does it cover situations of armed conflict (whether international or non-international) to which international humanitarian law applies.

12 The use of force to respond to peaceful protests is not covered by this article.

13 See analysis below in the third part of this article. States have a positive obligation to take reasonable and appropriate measures to facilitate, protect and enable lawful demonstrations to proceed peacefully. See, *inter alia*, HRC Res. 44/20, "The Promotion and Protection of Human Rights in the Context of Peaceful Protests", 17 July 2020, para. 4; HRC Res. 25/38, "The Promotion and Protection of Human Rights in the Context of Peaceful Protests", 28 March 2014, para. 4; HRC Res. 24/5, "The Rights to Freedom of Peaceful Assembly and of Association", 26 September 2013, preambular para. 8. See also

General and indiscriminate surveillance of peaceful assembly

Digital technologies have significantly expanded the capabilities of authorities to surveil assemblies, including protests. Technologies are used to monitor the planning and organization of protests, to conduct surveillance during protests and even to continue surveillance after protests. This information is obtained in bulk and indiscriminately from public and private spaces,[14] irrespective of whether the persons involved are suspected of committing a crime.[15]

"Safe and confidential communications play a key role in the planning and holding of peaceful protests",[16] and yet through the availability of data and tools to process it, public authorities are increasingly collecting and analyzing the personal information of those planning or organizing protests, as well as of protesters themselves, just for the mere fact of planning or participating in an assembly. Many of these surveillance methods are invisible to protesters and can be used without the knowledge, consent or participation of those surveilled.

Each new protest has been a testament to the fact that the list of tools used to surveil protests is only becoming longer. Online, authorities may monitor social media communications and collect all information posted in relation to the protest indiscriminately.[17] This includes accessing and collecting information from both public and private digital spaces. They even infiltrate private online groups by creating false accounts to monitor conversations, and impersonate protest organizers in order to influence discussions and planning and even arrest dissenters.[18] They may further request that user data be provided by social media platforms and mobile phone applications that track movement, including information on who has carried out an internet search about a protest and on

Pieter van Dijk, Fried van Hoof, Arjen van Rijn and Leo Zwaak (eds), *Theory and Practice of the European Convention on Human Rights*, 4th ed., Intersentia, Antwerp, 2006, pp. 836–837.

14 General Comment 37 underlines that the right to privacy may be infringed upon even when an assembly takes place in public: see General Comment 37, above note 5, para. 62. A similar approach has been followed by the European Court of Human Rights (ECtHR), which has recognized in its jurisprudence that individuals have a reasonable expectation of privacy, despite the fact that their actions might have taken place in public. See, among others, ECtHR, *Uzun v. Germany*, Appl. No. 35623/05 (Fifth Section), 2 September 2010, paras 48–53.

15 In this context, the concept of "a person of interest" in protests has been expanding to include also "influencers" of protests – a term which has been borrowed from marketing and which includes persons whose voice seems to attract attention and mobilize people. See Lina Dencik, Arne Hintz and Zoe Carey, "Prediction, Pre-Emption and Limits to Dissent: Social Media and Big Data Uses for Policing Protests in the United Kingdom", *New Media & Society*, Vol. 20, No. 4, 2018, p. 1445.

16 UN Human Rights, above note 10, para. 24; Human Rights Council, *Joint Report of the Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association and the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions on the Proper Management of Assemblies*, UN Doc. A/HRC/31/66, 4 February 2016, para. 75.

17 On the extent of intelligence that can be acquired by collecting, analyzing and combining information, see Privacy International, "Social Media Intelligence", 23 October 2018, available at: https://privacyinternational.org/explainer/55/social-media-intelligence. See also L. Dencik, A. Hintz and Z. Carey, above note 15.

18 They may also spread false information by impersonating organizers, or even directly endanger protesters by using, *inter alia*, a technique called "boxing" whereby they maliciously publish personal information in order to encourage physical harm to organizers and protesters. UN Human Rights, above note 10, para. 27.

discussions between the organizers or the protesters.[19] Finally, they use hacking techniques[20] either to infiltrate the social media accounts of organizers and protesters in order to get their contacts and private messages, or to infiltrate their devices by tricking them into, for example, downloading malicious software that gives the authorities unhindered access to contacts, messages, pictures, videos and all other personal information on their phones.[21]

General and indiscriminate surveillance intensifies during protests. Fake mobile phone towers, facial recognition software and sentiment analysis software[22] – and more recently, the deployment of military-grade drones reportedly equipped with some of these tools – are all used in concert to ensure that if authorities so decide, not a single person remains anonymous during a protest. For instance, interference with protesters' mobile phones is facilitated by a variety of devices impersonating mobile phone traffic towers that intercept and track all mobile phones in their vicinity. Such devices typically collect International Mobile Subscriber Identity (IMSI) and International Mobile Equipment Identity (IMEI) data that are unique to each mobile phone and SIM card – this is where they get one of their names, IMSI catchers.[23] However, they can do more than that: once connected, some devices also have the capability to block or intercept data transmitted and received by mobile phones, including the content of calls, text messages and websites visited.[24] They can potentially indiscriminately capture the mobile activity of thousands of people.

Authorities do not necessarily need to monitor mobile phones to capture everyone that was at an assembly. It is becoming a regular practice for authorities to make audio-visual recordings of assembly participants and often combine it with facial recognition technology, in real time (live facial recognition) or at a

19 Council of Europe, *Report by the Committee of Experts on Cross-Border Flow of Internet Traffic and Internet Freedom on Freedom of Assembly and Association on the Internet*, 10 December 2015.

20 Hacking is understood here as an act or series of acts which interfere with a system, causing it to act in a manner unintended or unforeseen by the manufacturer, user or owner of that system. Hacking can present grave and unique threats to privacy and security. For further information, see Privacy International, "Hacking Necessary Safeguards", 2018, available at: https://privacyinternational.org/demand/government-hacking-safeguards.

21 Access Now, "OHCHR Call for Input: The Promotion and Protection of Human Rights in the Context of Peaceful Protests", 2019, available at: www.accessnow.org/cms/assets/uploads/2020/06/OHCHR-Call-for-Input-Use-of-ICTs-in-Protests-October-15.pdf.

22 "Sentiment analysis" is used here to describe technology that aims to analyze various data, such as language, text and biometric data, in order to deduce the emotions of individuals. The term is also applied to "voice of the customer" materials such as reviews and survey responses.

23 IMSI catchers are known by a multitude of different names, including cell site simulators, cell grabbers, mobile device identifiers and man-in-the-middle devices, or by their specific brand names, such as StingRay or DRTbox. Jennifer Valentino-DeVries, "How 'Stingray' Devices Work", *Wall Street Journal*, 21 September 2011, available at: https://blogs.wsj.com/digits/2011/09/21/how-stingray-devices-work/.

24 See Christopher Soghoian and Stephanie K. Pell, "Your Secret Stingray's No Secret Anymore: The Vanishing Government Monopoly over Cell Phone Surveillance and Its Impact on National Security and Consumer Privacy", *Harvard Journal of Law & Technology*, Vol. 28, No. 1, 2014; Adrian Dabrowski, Nicola Pianta, Thomas Klepp, Martin Mulazzani and Edgar Weippl, "IMSI-Catch Me If You Can: IMSI-Catcher-Catchers", *Proceedings of the 30th Annual Computer Security Applications Conference*, ACM Press, 2014, available at: http://dl.acm.org/citation.cfm?doid=2664243.2664272.

later point.[25] In the process, the authorities may indiscriminately collect images of everyone at a protest. The technology allows the comparison of the digital representation of a face captured in a digital image with other images in a database to determine whether a given passer-by was a person of interest.[26]

Additionally, some authorities have also been reported to be using military-grade surveillance equipment that could have been equipped with IMSI catchers, facial recognition cameras and other tools to monitor protestors.[27] Other reported technology deployed during assemblies or protests includes automated number plate recognition software, credit card monitoring, mobile phone extraction technology used during stop and search or remotely,[28] sentiment recognition software, body-worn cameras, data from telecommunications providers, and cloud analytics. These are all used in concert to surveil protests, along with a series of aggregation tools that can combine data from all these sources into one record.

Interference with internet communications in peaceful assemblies

On top of these surveillance measures, other tactics used by authorities to interfere with assemblies (before, during and after) include filtering of content related to protests; blocking of websites or platforms used to plan, organize and mobilize protests; closing accounts that belong to organizers, activists or journalists; and shutting down of the Internet and communications networks.

Internet shutdowns describe complete shutdowns of telecommunications and mobile services and internet traffic.[29] These are measures that intend to prevent or disrupt access to or dissemination of information online.[30] Shutdowns may affect an entire country or multiple countries, specific regions, towns, or

25 "Facial recognition technology" is used here to describe any system that has been built to analyze images of individuals for the purpose of identifying them. Such systems can scan distinct, specific facial features, such as face shape, to create a detailed biometric map of a face. UN Human Rights, above note 10, para. 30.

26 The images in a watch list may come from a range of sources and do not just include images of people suspected of criminal wrongdoing. Shaun Walker, "Face Recognition App Taking Russia by Storm May Bring End to Public Anonymity", *The Guardian*, 17 May 2016, available at: www.theguardian.com/technology/2016/may/17/findface-face-recognition-app-end-public-anonymity-vkontakte. One company, called Clearview AI, trained its facial recognition system by using images found on people's social media profiles, without their consent. The Clearview AI facial recognition tool enabled police to link protesters to their respective social media accounts, making it harder for protesters to remain anonymous. Harmon Leon, "This Controversial Company Is Selling Your Social Media Photos to Law Enforcement", *The Observer*, 2 November 2020, available at: https://observer.com/2020/02/clearview-ai-social-media-photos-law-enforcement/.

27 Such technology combines data from mobile phones, license plate readers and real-time arrest records. In aggregate, this data makes it faster and easier for police to track and arrest suspects.

28 See Privacy International, "Mobile Phone Extraction", available at: https://privacyinternational.org/sites/default/files/2019-02/Explainers-MPE.pdf.

29 Also known as kill switches, network shutdowns or blackouts. See Access Now, "#KeepItOn: The Problem", available at: www.accessnow.org/keepiton/#problem. See also HRC Res. 32/13, "The Promotion, Protection and Enjoyment of Human Rights on the Internet", 1 July 2016.

30 David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc. A/HRC/35/22, 30 March 2017, para. 8, n. 6.

specific areas or neighbourhoods. Their duration varies from a couple of hours to months.[31] Internet shutdowns are becoming almost a common practice in times of public unrest and protests.[32] They "involve measures to intentionally prevent or disrupt access to or dissemination of information online in violation of human rights law".[33] At least sixty-five internet shutdowns reportedly took place during protests in 2019.[34]

Additionally, governments have been blocking, throttling or rendering effectively unusable entire websites, social platforms (such as Facebook and Twitter) and mobile applications (such as WhatsApp and Telegram).[35] Another more refined way of blocking and filtering information that the population receives is by blocking keywords or web pages.[36] Occasionally, authorities have been demanding that social platforms block specific users' accounts, claiming that they contain illegal content. This is a common practice targeting key figures of peaceful assemblies and associations in the making.

Other reported reactions include the removal of content related to protests and introducing new legislation that obliges intermediaries (telecommunications companies and service providers) to comply with such requests during protests or otherwise holding them accountable for these protests.[37] All these measures indicate that governments feel entitled to block digital means of communication in an era when they are often the only means of communication available to people.[38]

In addition to the above, governments are gradually introducing legal frameworks that criminalize uses of internet communications technologies, setting indirect barriers to the organization of and participation in assemblies. Increasingly, users of social media and other platforms used to organize protests have been targeted, arrested, prosecuted and even convicted exclusively for their online activities.[39] For instance, in certain States, a number of administrative and

31 Human Rights Committee, *Concluding Observations on the Fifth Periodic Report of Cameroon*, UN Doc. CCPR/C/CMR/CO/5, 30 November 2017, para. 41.

32 D. Kaye, above note 30, para. 8, n. 6, and para. 11. See also HRC Res. 32/13, above note 29, para. 10.

33 D. Kaye, above note 30, para. 8.

34 Access Now, *Targeted, Cut Off, and Left in the Dark: The #KeepItOn Report on Internet Shutdowns in 2019*, 2019, available at: www.accessnow.org/cms/assets/uploads/2020/02/KeepItOn-2019-report-1.pdf.

35 The UN High Commissioner for Human Rights recently reported that "[b]locking of entire websites of human rights organizations and political opposition parties has become increasingly common in many parts of the world, including in countries of the Middle East and North Africa region". UN Human Rights, above note 10, para. 23.

36 Sanja Kelly, Sarah Cook and Mai Truong (eds), *Freedom on the Net 2012: A Global Assessment of Internet and Digital Media*, Freedom House, 2012, pp. 164–176, available at: https://freedomhouse.org/sites/default/files/resources/FOTN%202012%20-%20Full%20Report_0.pdf.

37 Article 19, *The Right to Protest Principles: Background Paper*, 2016, p. 33, available at: www.article19.org/data/files/medialibrary/38581/Protest-Background-paper-Final-April-2016.pdf.

38 Disruption of protests through misinformation and attempts at disruptions of large-scale and public mobilization using the internet have been openly admitted by some States. Gayathry Venkiteswaran, *Freedom of Assembly and Association Online in India, Malaysia and Pakistan: Trends, Challenges and Recommendations*, APC IMPACT, 2016, p. 41, available at: www.apc.org/sites/default/files/FOAA_online_IndiaMalaysiaPakistan.pdf.

39 Victims of such arrests include journalists, human rights defenders (often the "faces" of protests) and other civilians – anyone that organizes, participates or reports on protests. Article 19, above note 37, pp. 33–34.

legislative measures have been taken to target non-violent involvement with protests,[40] with laws that include online acts expressly or implicitly. An example of the latter is the prohibition of the use of social media for the organization of protests in Brazil.[41] Implicitly, the inclusion of broad terms in some laws allows the authorities to target organizers for disturbing public order, incitement to disturb public order, terrorism, or threats to national security.[42] In parallel, there is also an increasing tendency to criminalize activities that could fall under the protection of freedom of assembly, such as certain forms of electronic civic disobedience.[43]

The right to privacy, freedom of expression and freedom of assembly

The impact of general and indiscriminate protest surveillance and interference with internet communications that facilitate the planning and organization of protests has been increasingly considered in recent years. As the UN High Commissioner for Human Rights has concluded, "the use of [new] technologies to surveil or crack down on protesters can lead to human rights violations, including infringement of the right to peaceful assembly".[44] This development is welcome, as often concerns regarding surveillance of assemblies and protests have been primarily identified as infringements of the right to privacy[45] or freedom of expression.[46]

The European Court of Human Rights (ECtHR) has found, for instance, that the retention of data of a peace movement activist who had never been convicted of any offence, concerning a peaceful protest, had been shown to be neither generally necessary nor necessary for the purposes of a particular inquiry. It was therefore a violation of his right to privacy.[47] Also, in a judgment of 25 June 2020, the Economic Community of West African States (ECOWAS) Community Court of Justice ruled that the September 2017 internet shutdown

40 *Ibid*., p. 33.
41 *Ibid*., p. 33.
42 For instance, the Spanish Criminal Code was amended to include a provision criminalizing distribution or public dissemination, through any means, of messages inciting the commission of any crime of disturbance of the peace. Nils Muižnieks, *Report by Nils Muižnieks, Commissioner for Human Rights of the Council of Europe, Following His Visit to Spain from 3 to 7 June 2013*, CommDH(2013)18, 2013, para. 130. See also Council of Europe, above note 19, p. 14.
43 Article 19, above note 37, pp. 25–26; Alex Comninos, *Freedom of Peaceful Assembly and Freedom of Association and the Internet*, APC Issue Paper, 2012, p. 7, available at: www.apc.org/sites/default/files/cyr_english_alex_comninos_pdf.pdf.
44 UN Human Rights, above note 10.
45 UDHR, Art. 12; ACHR, Art. 11; ICCPR, Art. 17; ECHR, Art. 8.
46 UDHR, Art. 19. See also ACHPR, Art. 9; ACHR, Art. 13; ICCPR, Art. 19; ECHR, Art. 10.
47 ECtHR, *Catt v. The United Kingdom*, Appl. No. 43514/15, Judgment (First Section), 24 January 2019. In another case, the Court stated that "Article 10 [freedom of expression] is to be regarded as a *lex generalis* in relation to Article 11 [freedom of assembly], *a lex specialis*, so that it is unnecessary to take it into consideration separately". ECtHR, *Ezelin v. France*, Appl. No. 11800/85, Judgment, 26 April 1991, para. 35.

ordered by the Togolese government during protests was illegal and an affront to the applicants' right to freedom of expression.[48]

Undoubtedly, the three rights of assembly, expression and privacy converge to a great degree when it comes to assemblies, and the lines between them inevitably blur. It is a long-standing principle of the administration of justice that a court often will not examine violations of closely linked rights once it has found a violation of one. This is not unique to freedom of assembly – the ECtHR, for example, has repeatedly declared that it was not necessary to examine whether there had been a violation under another right.[49] There are also cases where the Court found a violation of freedom of assembly and declared that it was unnecessary to examine whether there had been a violation of freedom of expression. Nonetheless, these cases relate to instances where protesters were arrested before, during or immediately after a protest or where the State banned a protest from taking place or unjustifiably restricted the organization of an assembly.[50]

At the UN level, freedom of assembly has been increasingly added next to privacy and freedom of expression when considering the impact of surveillance and interference with internet communications on human rights, precisely in order to underline specific concerns that are raised by the surveillance of peaceful assemblies.[51] There are also resolutions, reports and other documents focusing on freedom of assembly in the digital age that underline the impact of surveillance on assembly.[52]

However, it is maintained here that despite the increasing attention given to freedom of assembly, international and regional bodies (judicial, quasi-judicial, political and independent experts) have not fully explored and captured the full extent to which general and indiscriminate surveillance and interference with internet communications – before, during and after – directly interfere with and potentially infringe on freedom of assembly.

For instance, when examining the impact of assembly surveillance, the UN Human Rights Committee's General Comment 37 focuses on the right to privacy. It states:

> The mere fact that a particular assembly takes place in public *does not mean that participants' privacy cannot be violated*. The right to privacy may be infringed, for example, by facial recognition and other technologies that can identify individual participants in a crowd.[53]

48 Community Court of Justice of ECOWAS, *Amnesty International Togo and Others v. The Togolese Republic*, Judgment, 25 June 2020. Similarly, in December 2012, the ECtHR ruled unanimously that the blanket blocking of entire platforms, in this case the hosting service Google Sites, violates freedom of expression provisions in Article 10 of the ECHR. ECtHR, *Ahmet Yıldırım v. Turkey*, Appl. No. 3111/10, Judgment (Second Section), 18 December 2012, paras 66–68.

49 ECtHR, *Ezelin*, above note 47.

50 See, among others, ECtHR, *Öllinger v. Austria*, Appl. No. 76900/01, Judgment (First Section), 29 June 2006, paras 52–53.

51 See, *inter alia*, HRC Res. 42/15, "The Right to Privacy in the Digital Age", 26 September 2019, preambular para. 30.

52 See, *inter alia*, General Comment 37, above note 5; UNGA Res. 73/179, "The Right to Privacy in the Digital Age", 17 December 2018; HRC Res. 44/20, above note 13; C. Voule, above note 6; UN Human Rights, above note 10.

53 General Comment 37, above note 5, para. 62 (emphasis added).

The in-depth consideration of the impact of surveillance and interference with internet communications on assemblies is key both for ensuring that the distinct nature of freedom of assembly is preserved, and for the better regulation, implementation and enforcement of surveillance measures around protests.

Freedom of assembly incorporates the right of every individual to hold opinions without interference, an element that also describes freedom of expression.[54] However, assembly also encompasses a social component, the sense of acting to pursue a common interest or purpose. It protects the collective nature of protests brought together by a common aim.[55] When freedom of assembly is attacked, the societal network that is united under the specific aim is damaged.[56] As such, freedom of assembly helps to develop and strengthen democratic societies.[57] In equal measure, while surveillance may be primarily an interference with the right to privacy, such interference often provides a gateway to violations of other rights, including freedom of assembly.[58] A violation of the right to privacy, more often than not, is not an end in itself; it rather offers the means for infringing on other rights.[59] In that sense, it often becomes the enabler for infringing on, among others, freedom of assembly.

Mass surveillance and interference with internet communications as an infringement of freedom of assembly

Direct interference with freedom of assembly

Many of the surveillance and internet communications interference measures referred to above can be used to directly interfere with the exercise of freedom of assembly. For example, as mentioned above, so-called IMSI catchers can be used to monitor and intercept ingoing and outgoing communications, but can also edit

54 This *lex specialis* nature is mentioned in ECtHR, *Ezelin*, above note 47.

55 The Human Rights Committee found the right to freedom of assembly to be irrelevant if one is acting alone. Human Rights Committee, *Patrick Coleman v. Australia*, Communication No. 1157/2003, UN Doc. CCPR/C/87/D/1157/2003, Views, 10 August 2006, para. 6.4.

56 One of the distinctive criteria noted by the ECtHR is that in the exercise of the right to freedom of assembly the participants would be seeking not only to express their opinion, but to do so together with others. See, among others, ECtHR, *Navalnyy v. Russia*, Appl. Nos 29580/12 and 4 others, Judgment (Grand Chamber), 17 February 2004, para. 101. See also M. Hamilton, above note 8, pp. 525–526, 534–535.

57 As the ECtHR has underlined, "the participation of citizens in the democratic process is to a large extent achieved through belonging to associations in which they may integrate with each other and pursue common objectives collectively". ECtHR, *Gorzelik and Others v. Poland*, Appl. No. 44158/98, Judgment (Grand Chamber), 17 February 2004, para. 92. See also HRC Res. 38/11, "The Promotion and Protection of Human Rights in the Context of Peaceful Protests", 16 July 2018, p. 11.

58 Most recently, UNGA Res. 73/179, above note 52, para. 9; HRC Res. 42/15, above note 51, preambular para. 12.

59 "[I]n the digital age, technical solutions to secure and to protect the confidentiality of digital communications, which may include measures for encryption, pseudonymization and anonymity, can be important to ensure the enjoyment of human rights, in particular the rights to privacy, to freedom of expression *and to freedom of peaceful assembly* and association." UNGA Res. 73/179, above note 52, preambular para. 26 (emphasis added).

or reroute mobile communications, as well as block service. Governments may use an IMSI catcher to send a message to mobile phones in the area as a way of intimidating protesters or manipulating them into disbanding or conducting some other activity. Similarly, internet shutdowns or placing restrictions on secure and confidential communications may constitute a direct interference with freedom of assembly insofar as they represent an attempt by the government to prevent a protest from being organized or disperse an already ongoing protest. These actions directly hinder the ability of individuals to attend a gathering, to communicate with one another and to organize further.[60]

Freedom of assembly is indeed a qualified right and may be restricted when necessary in a democratic society for a legitimate aim – in the interests of national security, public safety etc.[61] However, as the Human Rights Committee has repeatedly underlined, it may only be limited under strict conditions.[62] Restrictions can be imposed only if prescribed by law and necessary and proportionate in the circumstances, but more often than not these interferences are not even prescribed by law. For instance, in many countries there is no legal framework regulating the use of mass surveillance tools, such as IMSI catchers, and often protesters will not even be aware of their presence during a protest.[63]

Additionally, restrictions to freedom of assembly need to be specific and necessary to achieve a specific legitimate aim;[64] there needs to be a rational connection between the measure and the prescribed aim, meaning that a measure cannot be based on an abstract aspiration that it might facilitate the aim.[65] However, it is arguably impossible to find such a link when imposing general and indiscriminate surveillance measures,[66] or indeed to justify such mass interferences in any circumstances. When protests turn violent or in other situations of violence, certain targeted surveillance and investigatory measures may be taken, but generalized measures against an abstract threat cannot tilt the balance. For instance, governments often claim that internet shutdowns are

60 See Privacy International, *Submission to the Office of the United Nations High Commissioner for Human Rights on the Promotion and Protection of Human Rights in the Context of Peaceful Protests*, October 2019, available at: https://tinyurl.com/2tcqlbn8.

61 ICCPR, Art. 21; ECHR, Art. 11(2).

62 See, among others, Human Rights Committee, *Zinaida Shumilina et al. v. Belarus*, Communication No. 2142/2012, Views, 28 July 2017; Human Rights Committee, *Pavel Levinov v. Belarus*, Communication No. 2082/2011, Views, 14 July 2016.

63 Privacy International, "IMSI Catchers: Facilitating Indiscriminate Surveillance of Protesters", 19 June 2020, available at: https://privacyinternational.org/news-analysis/3948/imsi-catchers-facilitating-indiscriminate-surveillance-protesters.

64 "Such attempts to interfere with the freedom of expression unlawfully pursue an illegitimate objective of undermining the right to peaceful protest". David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc. A/HRC/29/32, 22 May 2015, para. 53.

65 Aharon Barak, "Rational Connection", in *Proportionality: Constitutional Rights and their Limitations*, Cambridge University Press, Cambridge, 2012, pp. 303–316.

66 "General and indiscriminate surveillance measures" here describes systems or technologies that collect, analyze and/or generate data on large numbers of people instead of limiting surveillance to individuals about which there is reasonable suspicion of wrongdoing. See, for instance, Court of Justice of the European Union, *Tele2 Sverige AB v. Post- och telestyrelsen (C-203/15) and Secretary of State for the Home Department v. Tom Watson ao (C-698/15)*, Judgment, 21 December 2016.

necessary for public safety when a peaceful protest is about to turn violent, but such kill switches "may well exacerbate, rather than curtail, tensions";[67] they can stress the gathered crowds, who are no longer able to be informed about what is happening around them.[68] In addition, internet shutdowns affect not only the assemblers but also those who are living in, working in and passing through the area where the assembly takes place.

Following a blockage of Twitter and YouTube, the Commissioner for Human Rights of the Council of Europe underlined that although illegal content could be blocked, applying this measure to entire platforms was a disproportionate response. Accordingly, he requested that such blockages should be lifted.[69] The UN Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association has emphasized that shutdowns and the blocking of entire websites constitute an extreme and disproportionate measure that cannot be justified in any circumstances.[70] He has also called for the prohibition of indiscriminate and untargeted surveillance of those exercising their right to peaceful assembly, in both physical and digital spaces.[71]

Certainty of surveillance amounts to an infringement of freedom of assembly

The certainty of surveillance and interference with communications technologies, particularly due to the general and indiscriminate nature of such measures, arguably infringes on the obligation not to interfere with protests as such, irrespective of whether the information collected is used to further directly interfere with the organization of an assembly, as argued above.

Undoubtedly, when attending a public assembly there may be a reasonable expectation that individuals might be identified, either because police conduct an investigation or because their face appears in a newspaper photograph. However, current surveillance of assemblies, particularly protests, has gone far beyond the expectation of some degree of publicity.[72] It is now becoming a certainty that protesters could be identified through the data collected by the authorities when

67 "Silencing Opposition Is 'Not the Solution', UN Rights Chief Says as Internet Blackout Looms in DR Congo", *UN News*, 17 December 2016, available at: https://news.un.org/en/story/2016/12/548052-silencing-opposition-not-solution-un-rights-chief-says-internet-blackout-looms.

68 On the contrary, the Special Rapporteur has underlined that the Internet may be used to mitigate public safety concerns and help restore public order. For instance, internet communications are key to disseminating accurate information during a crisis. D. Kaye, above note 64, para. 14.

69 *Annual Activity Report 2013 to the Parliamentary Assembly of the Council of Europe: Report of the Thirteenth Sitting*, AS (2014) CR 13, 8 April 2014.

70 UN Human Rights, above note 10, para. 22.

71 Surveillance of protesters should only be conducted on a targeted basis, and only when there is reasonable suspicion that the targets are engaging in or planning to engage in serious criminal offences, based on principles of necessity and proportionality and with judicial supervision. C. Voule, above note 6, para. 57.

72 While the notion of "reasonable expectation of privacy" is found in US law, the jurisprudence of the ECtHR, among others, also seems to recognize the concept, albeit not with an identical understanding. See, for example, John Ip, "The Legality of 'Suspicionless' Stop and Search Powers under the European Convention on Human Rights", *Human Rights Law Review*, Vol. 17, No. 3, 2017.

it is processed and analyzed. As such, the general and indiscriminate surveillance of protests amounts to an unjustified interference not only with the right to privacy but also with the right to peacefully assemble, and in turn violates freedom of assembly.

For instance, the regular audio-visual recording of protests in combination with facial recognition technology requires the collection and processing of facial images of all persons captured by the camera (irrespective of whether the authorities use facial recognition in real time or at a later stage). The permanent record that can be created by these recordings could allow authorities, if they so decide, to identify all those that participated in a protest even at a later time.[73] The UN High Commissioner for Human Rights has recommended that States "[n]ever use facial recognition technology to identify those peacefully participating in an assembly".[74]

Similarly, identity catchers, such as IMSI catchers, can capture the call activity of thousands of people indiscriminately. These are interferences on a mass scale.[75] Additionally, the aggregation of information acquired from different means and methods of surveillance before, during and after protests gives police forces the power to de-anonymize and identify everyone involved in the protest, irrespective of whether they are suspected of having committed a crime.[76] At the same time, the duration of the consequence of surveillance has also radically changed, as there is little indication of how long law enforcement and other agencies involved in protest surveillance will be keeping a record of the collected data.[77]

Inherent to freedom of assembly is the ability to participate in a protest without retribution. Anonymity plays a key role for safe and confidential communications in the planning and holding of protests, as well as for

73 General Comment 37 reiterates that "[t]he wearing of face-coverings or other disguises by assembly participants, such as hoods or masks, or taking other steps to participate anonymously may form part of the expressive element of a peaceful assembly, serve to counter reprisals, or to protect privacy, including in the context of new surveillance technologies". General Comment 37, above note 5, para. 60. Depending on how facial recognition technology develops, it could interfere with this possibility. A company has already claimed to be in the process of developing technology that even bypasses masks. Khari Johnson, "Facial Recognition Is No Match for Face Masks, but Things Are Changing Fast", *VentureBeat*, 8 April 2020, available at: https://venturebeat.com/2020/04/08/facial-recognition-is-no-match-for-face-masks-but-things-are-changing-fast/.

74 UN Human Rights, above note 10, para 53(h).

75 The 2020 Annual Report of the UN High Commissioner for Human Rights confirmed as much, even though it didn't go as far as concluding that blanket measures as such amount to a violation of freedom of assembly. UN Human Rights, above note 10. See also HRC Res. 44/20, above note 13, para. 26.

76 Before a protest, if a person uses social media to support or register with the protest, the police will collect this information; during a protest, if a person takes their mobile phone with them, which most would do, they may be surveilled by drones and IMSI catchers, or if they do not take their mobile, they may be surveilled by facial recognition technology, stop and search of passers-by or the use of a credit card or travel card; and finally, organizers and other "persons of interest" (not suspected of having committed a crime) are often kept under surveillance long after the protest.

77 The ECtHR in the *Catt* case referred to Principle 2 of Recommendation R(87)15 that regulates the use of personal data in the police sector, which states that "the collection of data on individuals solely on the basis that they belong to particular movements or organisations which are not prescribed by law should be prohibited unless absolutely necessary or for the purposes of a particular inquiry". ECtHR, *Catt*, above note 47, para. 124. See also ECtHR, *Segerstedt-Wiberg and Others v. Sweden*, Appl. No. 7124/09, Judgment (Second Section), 6 June 2006, para 107.

participating in protests.[78] Individuals rely on the anonymity of the crowd to protect themselves against retribution, particularly in contexts where any form of dissent is suppressed.[79] This is no longer an option, however, as people's mere participation in a protest today promises the erosion of their privacy, particularly as the cumulative use of these surveillance systems and methods guarantees that the information of individual protesters will be captured by at least one of them, leading to the individual's potential identification. It is argued here that the inevitability of surveillance, as such, becomes a barrier to the organization of and participation in assemblies, including protests, and thus constitutes an unjustified interference that infringes on freedom of assembly.

Infringement of the obligation to facilitate assemblies

General and indiscriminate assembly surveillance and interference with internet communications violate the positive obligations of States to facilitate assemblies and protect assemblers, as well as their positive obligation to take precautionary measures to prevent violations and abuses of the different rights at stake.

States need to secure the effective enjoyment of freedom of assembly.[80] Therefore, they have a positive obligation to take reasonable and appropriate measures to facilitate, protect and enable lawful demonstrations to proceed peacefully.[81] Undoubtedly, in order to fulfil these obligations, they have to take certain measures – for instance, redirecting traffic or providing security.[82] However, the need to adopt such measures is not without limits. The measures must never impair the essence of the right and cannot serve as a justification for measures that violate freedom of assembly, among other rights.[83]

If the network that enables the organization and holding of assemblies is shut down before a demonstration takes place, such a measure directly violates the positive obligation of States to facilitate the exercise of freedom of assembly. Associated activities that happen online in advance of an assembly are equally protected under freedom of assembly.[84] As the ECtHR has repeatedly underlined, "a system of secret surveillance set up to protect national security may undermine or even destroy democracy under the cloak of defending it".[85]

78 See the sources cited in above note 16.
79 General Comment 37, above note 5, para. 60.
80 ECtHR, *Kudrevičius and Others v. Lithuania*, Appl. No. 37553/05, Judgment (Grand Chamber), 15 October 2015, para. 158; ECtHR, *Djavit An v. Turkey*, Appl. No. 20652/92, Judgment (Third Section), 20 February 2003, para. 57.
81 See, *inter alia*, HRC Res. 44/20, above note 13, para. 4; HRC Res. 25/38, above note 13, para. 4; HRC Res. 24/5, above note 13, preambular para. 8. See also P. van Dijk *et al.*, above note 13, pp. 836–837.
82 ECtHR, *Oya Ataman v. Turkey*, Appl. No. 74552/01, Judgment (Second Section), 5 December 2006, para. 39.
83 Human Rights Committee, General Comment No. 31, "The Nature of the General Legal Obligation Imposed on States Parties to the Covenant", UN Doc. CCPR/C/21/Rev.1/Add.13, 26 May 2004 (General Comment 31), para. 6.
84 General Comment 37, above note 5, para. 34.
85 See, *inter alia*, in relation to privacy-related cases, ECtHR, *Big Brother Watch and Others v. the United Kingdom*, Appl. Nos 58170/13, 62322/14, 24960/15, Judgment (First Section, pending referral to the

Undermining the privacy of communications as such infringes on freedom of assembly, because the capacity to use communications technologies securely and privately is vital to the organization and conduct of assemblies.[86] Therefore, any general and indiscriminate surveillance or internet communications interference, including blocking internet connectivity or monitoring social media and other online communications, should also be understood as a violation of the obligation of States to facilitate assemblies.[87]

Infringement of the obligation to ensure a legal framework that safeguards freedom of assembly

Mass surveillance and interference with internet communications infringe on the positive obligation of States to promote an enabling environment for the exercise of the right to peaceful assembly.

Part of this obligation is the overarching obligation to ensure that there is an appropriate, accessible and foreseeable legal and institutional framework that regulates the exercise of freedom of assembly.[88] The legal framework must clearly set out the duties and responsibilities of all those acting in an official capacity – including private companies contracted to provide security – involved in managing assemblies in accordance with international standards, including who can surveil protests or interfere with new technologies, and when they can do so.[89] For instance, the use of IMSI catchers without any framework or of military-grade predator drones, or interference with internet communications by intercepting, redirecting or blocking the use of specific platforms or pages, should all be also understood as *ipso facto* violations of freedom of assembly.[90]

More often than not nowadays, police are deploying surveillance measures and interfering with communications technologies without necessarily abiding by a specific legal framework, either because such a framework does not exist or because the existing one is interpreted too broadly. The absence of a legal framework regulating the use of new technologies for surveillance or interference before, during and after protests, or the existence of one that gives very broad and excessive powers to authorities, should be understood as a direct violation of the obligation to safeguard the exercise of freedom of assembly.

Grand Chamber), 13 September 2018, para. 308; ECtHR, *Roman Zakharov v. Russia*, Appl. No. 47143/06, Judgment (Grand Chamber), 4 December 2015, para. 232.

86 The UN Human Rights Council has underlined that "the possibility of using communications technology securely and privately … is important for the organization and conduct of assemblies". HRC Res. 44/20, above note 13, preambular para. 22. See also the sources cited in above note 16.

87 *Inter alia*, the Human Rights Council reiterated "the importance for all States to promote and facilitate access to the Internet and international cooperation aimed at the development of media and information and communications facilities in all countries". HRC Res. 24/5, above note 13, preambular para. 8.

88 General Comment 37, above note 5, para. 28; see also the obligation to facilitate protests at para. 24.

89 *Ibid*., para. 28.

90 See above on IMSI catchers.

Violation of the obligation to respect freedom of assembly

General and indiscriminate surveillance and interference with internet communications violate the obligation to respect freedom of assembly, due to the chilling effect that their use causes.

As part of the obligation to respect freedom of assembly, States have a negative obligation to refrain from actions that will undermine the enjoyment of this right.[91] General and indiscriminate surveillance and interference with internet communications have the capacity to "chill" the exercise of freedom of assembly, as the monitoring and recording of participants at an assembly may prevent them from joining.[92] In the *Big Brother Watch* case, the ECtHR accepted that any perceived interference with the confidentiality of communications without any limitations may result in a "chilling effect" – that is, a self-restraint – on the lawful exercise of a right, particularly freedom of expression; hence it found a violation of Article 10 of the European Convention on Human Rights.[93] The inevitability of surveillance (see above) should thus be understood as a violation of the obligation to respect freedom of assembly, and not only as an interference with freedom of assembly.

Also, these newer forms of government surveillance, where practices (such as employing facial recognition technologies) lack foreseeability and transparency, exacerbate the negative impact on the exercise of freedom of assembly.[94] As warned by the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, "even a narrow, non-transparent, undocumented, executive use of surveillance may have a chilling effect without careful and public documentation of its use, and known checks and balances to prevent its misuse".[95]

91 Among others, see ECtHR, *Plattform "Ärzte für das Leben" v. Austria*, Appl. No. 10126/82, Judgment, 21 June 1988.

92 Human Rights Council, above note 16, para. 76.

93 ECtHR, *Big Brother Watch*, above note 85, para. 495. See also Bart van der Sloot, "Is the Human Rights Framework Still Fit for the Big Data Era? A Discussion of the ECtHR's Case Law on Privacy Violations Arising from Surveillance Activities", in Serge Gutwirth, Ronald Leenes and Paul De Hert (eds), *Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection*, Springer, Dordrecht, 2016, p. 422.

94 In the context of secret surveillance, the ECtHR has found it "unacceptable that the assurance of the enjoyment of a right guaranteed by the Convention could be thus removed by the simple fact that the person concerned is kept unaware of its violation". ECtHR, *Klass and Others v. Germany*, Appl. No. 5029/71, Judgment (Plenary), 6 September 1978, para. 36. In the context of freedom of expression, Special Rapporteur David Kaye has noted that "[u]nnecessary and disproportionate surveillance may undermine security online and access to information and ideas. Surveillance may create a chilling effect on the online expression of ordinary citizens, who may self-censor for fear of being constantly tracked. Surveillance exerts a disproportionate impact on the freedom of expression of a wide range of vulnerable groups, including racial, religious, ethnic, gender and sexual minorities, members of certain political parties, civil society, human rights defenders, professionals such as journalists, lawyers and trade unionists, victims of violence and abuse, and children." David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc. A/HRC/32/38, 11 May 2016, para. 57.

95 Frank La Rue, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc. A/HRC/23/40, 17 April 2013, para. 52.

Attacking the essence of freedom of assembly

Finally, general and indiscriminate surveillance and interference with internet communications undermine the essence of freedom of assembly.

Human rights instruments that guarantee freedom of assembly permit certain interferences with this right, so long as those interferences abide by certain strictly interpreted principles, including legality, necessity and proportionality, to the extent that they do not undermine the essence or core of this right. As the Human Rights Committee has emphasized, "[i]n no case may the restrictions be applied or invoked in a manner that would impair the essence of a Covenant right".[96]

This obligation is embedded in the core provisions of each human rights instrument, which guarantee that nothing in their provisions may be interpreted as implying that a State or other entity can engage in any act that will lead to the destruction of any of the rights of freedom set forth therein, including freedom of assembly.[97] The ECtHR, on a case relating to measures restricting assembly, held

> that notification, and even authorisation procedures, for a public event *do not in general encroach upon the essence of the right* [of freedom of assembly], *as long as the purpose of regulating the assembly is to allow the authorities to take reasonable and appropriate measures in order to guarantee its smooth conduct*.[98]

It went on to add, though, that "the enforcement of such rules cannot become an end in itself".[99]

In another case, the Court has noted

> that the very essence of the right to freedom of peaceful assembly would be impaired, if the State was not to prohibit a demonstration but was then to impose sanctions on its participants, even one at the lower end of the scale of penalties, for the mere fact of attending it, without committing anything reprehensible, as happened in the applicant's case.[100]

In other words, what this reasoning suggests is that blanket surveillance and other interferences that dissuade individuals from participating in assemblies could be regarded as adversely affecting the essence of freedom of assembly.

96 General Comment 31, above note 83, para. 6. In a different context, the ECtHR has also observed that there exists "the risk that a system of secret surveillance set up to protect national security may undermine or even destroy democracy under the cloak of defending it". ECtHR, *Zakharov*, above note 86, para. 232. UN Human Rights has similarly observed that "any limitation to the right to privacy must not render the essence of the right meaningless and must be consistent with other human rights". UN Human Rights, *The Right to Privacy in the Digital Age: Report of the Office of the United Nations High Commissioner for Human Rights*, UN Doc. A/HRC/27/37, 30 June 2014, para. 23.

97 ICCPR, Art. 2; ECHR, Art. 17; UDHR, Art. 30.

98 ECtHR, *Navalnyy*, above note 56, para. 100 (emphasis added).

99 *Ibid.*, para. 100.

100 The Court therefore concluded that the interference with the applicant's right to freedom of peaceful assembly was not "necessary in a democratic society". ECtHR, *Galstyan v. Armenia*, Appl. No. 26986/03, Judgment (Third Section), 15 November 2007, para. 117; see also ECtHR, *Ashughyan v. Armenia*, Appl. No. 33268/03, Judgment (Third Section), para. 93.

The restrictions imposed upon this right should not unacceptably weaken the protection afforded by it. Freedom of assembly guarantees the right to collectively and peacefully meet, demonstrate or protest without retribution. Read together with the arguments provided in the previous sections, we can conclude that the erosion of participants' anonymity, the inevitability of surveillance, and blanket interference with people's communications for the mere fact of having participated in a gathering adversely affect the essence of freedom of assembly.

Conclusion

New forms of control through the use of surveillance, as well as interference with internet communications, have been increasingly deployed by States to control assemblies, including general and indiscriminate surveillance, internet shutdowns, and the blocking of social media platforms, web pages and mobile applications. Undoubtedly, the new digital reality requires governments to adapt and use the tools at their disposal to assist them in ensuring the safe and free administration of assemblies and movements. However, there has always been one condition – they should always safeguard the enjoyment of human rights in the process.

The use of any such measures should comply with the legal requirements not only of the right to privacy and freedom of expression, but also of the right to freedom of assembly. General and indiscriminate surveillance and blanket interferences with internet communications amount to a direct infringement of the right to freedom of assembly on multiple grounds, and as such should not be used in the context of assemblies – if at all, though that is a separate conversation.[101] General and indiscriminate surveillance and interference with internet communications infringe on freedom of assembly when they are used for direct, unjustified interference with assemblies; they render surveillance inevitable, instead of a possibility; they violate the obligation to facilitate assemblies and the obligation to have a legal framework that facilitates assemblies, as well as the obligation to respect freedom of assembly; and, last but not least, they attack the essence of the right.

Undoubtedly, some of these acts can be and have on occasion been understood as violations of the right to privacy and/or freedom of expression. However, examining their impact on other rights allows for more effective protection of the core values that each protects. This separate legal analysis is needed not only to preserve the distinct nature of freedom of assembly that protects collective action, but also to allow for better regulation of surveillance and interference with internet communications in assemblies, demonstrations and protests. Freedom of assembly is only the beginning – other human rights and their distinct nature stand in line, including freedom of religion and belief and the right to participate in public affairs.

101 See arguments brought forward by Big Brother Watch, Privacy International, Amnesty international and seven other organizations in *Big Brother Watch and Others v. UK*, pending before the Grand Chamber of the ECtHR at the time of writing. ECtHR, *Big Brother Watch*, above note 85.

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

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.

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).

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.

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.

> 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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).

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.

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.*

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.

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.

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.

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.

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.

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.

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.

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").

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.

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.

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.

International Review of the Red Cross (2020), **102** (913), 235–259.
Digital technologies and war
doi:10.1017/S1816383120000466

IRRC

Stepping back from the brink: Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible

Frank Sauer
Frank Sauer is a Senior Research Fellow at Bundeswehr University in Munich. He serves on the International Panel on the Regulation of Autonomous Weapons and is a member of the International Committee for Robot Arms Control, which co-founded the international Campaign to Stop Killer Robots.
Email: frank.sauer@unibw.de.

Abstract
This article explains why regulating autonomy in weapons systems, entailing the codification of a legally binding obligation to retain meaningful human control over the use of force, is such a challenging task within the framework of the United Nations Convention on Certain Conventional Weapons. It is difficult because it requires new diplomatic language, and because the military value of weapon autonomy is hard to forego in the current arms control winter. The article argues that regulation is nevertheless imperative, because the strategic as well as ethical risks outweigh the military benefits of unshackled weapon autonomy. To this end, it offers some thoughts on how the implementation of regulation can be expedited.

Keywords: artificial intelligence, lethal autonomous weapons systems, arms control winter, regulation, Convention on Certain Conventional Weapons, strategic stability, human dignity.

Introduction

The United Nations (UN) Convention on Certain Conventional Weapons (CCW) is the epicentre of the global debate on autonomy in weapons systems. The CCW's purpose "is to ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately".[1] In CCW parlance, the weapon autonomy issue is called "emerging technologies in the area of lethal autonomous weapons systems" (LAWS). In November 2019, CCW States Parties decided to, once again, continue their deliberations on LAWS. For the first time, however, these talks, which had previously been conducted between 2014 and 2016 in informal meetings and since 2017 within the framework of an expert subsidiary body called a Group of Governmental Experts (GGE), were mandated to produce a specific outcome. For ten days in 2020 and for an as-yet unknown number of days in 2021 (when the CCW's next Review Conference is due), the GGE was and is tasked with debating and fleshing out "aspects of the normative and operational framework" on LAWS.[2] In addition, in Annex III of their 2019 report, States Parties adopted eleven guiding principles to take into account going forward.[3] After the first five-day meeting of 2020 was postponed and then conducted in a hybrid format due to the current global COVID-19 pandemic, the second meeting had to be shelved, and it is currently unclear when and how the talks can resume.

While some States – most prominently Russia – have displayed no interest in producing new international law in the CCW, arguing that "concerns regarding LAWS can be addressed through faithful implementation of the existing international legal norms",[4] others – such as Germany – claim that nothing short of "an important milestone" has already been reached with the 2019 report cited above, even describing the adopted eleven guiding principles as a "politically binding regulation".[5]

Meanwhile, the international Campaign to Stop Killer Robots (Killer Robots Campaign, KRC) is criticizing CCW diplomacy as "moving forward at a snail's pace", with low ambitions and negligible outcomes despite widespread

1 United Nations Office in Geneva, "The Convention on Certain Conventional Weapons", available at: https://tinyurl.com/y4orq8q5 (all internet references were accessed in December 2020).

2 UN, *Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects: Revised Draft Final Report*, UN Doc. CCW/MSP/2019/CRP.2/Rev.1, Geneva, 15 November 2019 (CCW Meeting Final Report), p. 5, available at: https://tinyurl.com/y3gjy7mk.

3 *Ibid.*, p. 10.

4 Russian Federation, *Potential Opportunities and Limitations of Military Uses of Lethal Autonomous Weapons Systems: Working Paper Submitted by the Russian Federation*, UN Doc. CCW/GGE.1/2019/WP.1, 15 March 2019, p. 5, available at: https://tinyurl.com/yx9op3n4.

5 German Federal Foreign Office, "Foreign Minister Maas on Agreement of Guiding Principles relating to the Use of Fully Autonomous Weapons Systems", press release, 15 November 2019, available at: www.auswaertiges-amt.de/en/newsroom/news/maas-autonomous-weapons-systems/2277194.

public opposition to LAWS and some thirty countries (twenty-six of which are CCW States Parties) calling for the immediate negotiation of a new, binding legal instrument rather than continuing talks on frameworks and principles, which the KRC tends to consider vague and redundant respectively.[6]

It should be noted up front that the term LAWS itself is problematic. After all, neither "lethality" nor "autonomy" are decisive factors in the debate. The military application of non-lethal force raises concerns as well (take the prohibition against blinding lasers as just one example), and the term "autonomy", philosophically speaking, inappropriately anthropomorphizes machines that have limited agency and are incapable of reasoning and reflecting, as well as being unable to take on responsibility. Nonetheless, at this point the term LAWS is widely used as a shorthand, so the article will stick to this common vocabulary. Also, the article uses the term "regulation" – rather than, for instance, "ban" – because what potential new, binding international law on this issue is commonly understood to eventually codify is not a prohibition of a category of weapons. Instead, it is a positive obligation to retain meaningful human control over the use of military force. And while one might argue that ensuring meaningful human control and prohibiting autonomous weapons (AKA "killer robots") are two sides of the same coin, these two sides nevertheless represent different ways of approaching the issue, as I will argue further below. Lastly, I use the term "technology diffusion" rather than "proliferation" because the latter suggests a distribution from one or only a few points of departure (as in the case of nuclear proliferation) whereas the former suggests an omnidirectional spread from multiple sources, a more fitting picture in this case of widely (and oftentimes even commercially) available hardware and software.

In what follows, I first explain why it is so challenging for everyone involved in the debate to get a conceptual handle on the issue and, for CCW States Parties, to agree on impactful multilateral regulation on LAWS. I argue that finding proper language and a suitable legal framing for the retention of meaningful human control, in light of the enormous military value ascribed to unshackled weapon autonomy, is what makes regulating LAWS so exceptionally difficult. Subsequently, I discuss the implications of inaction, making the case for why retaining human control over the use of force is indeed imperative due to the strategic and ethical risks outweighing the potential military benefits. Lastly, I put forward some suggestions on how regulation could be advanced in practice, the formidable challenge of gathering enough political will amongst CCW States Parties notwithstanding. This is followed by a brief conclusion.

6 KRC, "Alarm Bells Ring on Killer Robots", 15 November 2019, available *at:* www.stopkillerrobots.org/2019/11/alarmbells/; Richard Moyes, "Critical Commentary on the 'Guiding Principles'", Article 36, November 2019, available at: www.article36.org/wp-content/uploads/2019/11/Commentary-on-the-guiding-principles.pdf.

Why regulating weapon autonomy is difficult: Conceptual pitfalls and power politics

From UN Secretary-General António Guterres to prominent members of the artificial intelligence (AI) and tech communities[7] to most States Parties of the CCW, there is near unanimity that LAWS raise various legal, strategic and ethical questions and concerns.[8] Even so, within the CCW States Parties, a consensus on new, binding international law is still a long way off. Regulating weapon autonomy through this multilateral forum is a particularly tough nut to crack. As I will argue in this section, this is due to two reasons. First, weapon autonomy as an issue is comparatively elusive and hard to conceptualize. Second, its perceived military value is exceptionally high, and the current geopolitical landscape is not conducive to new arms control breakthroughs.

Any discussion of the conceptual challenges regarding weapon autonomy has to begin with pointing out a common misunderstanding: the lack of progress in the CCW cannot be attributed to States Parties not having arrived at a shared definition of LAWS yet.[9] Quite to the contrary, it has much more to do with the fact that the attempt to define LAWS was misconceived from the very beginning. This warrants further elaboration.

The first two to three years of the CCW process on LAWS were indeed plagued by confusion and definitional struggles. Considerable effort was required to delineate the LAWS debate from the disputes surrounding remotely piloted aerial vehicles (drones) as well as to avoid anthropomorphizing LAWS as a one-to-one replacement for human soldiers.[10] All stakeholders were seeking – and quite a few lamenting the lack of – a "possible definition of LAWS", sometimes deliberately so in order to justify political heel-dragging. The underlying rationale was that arms control always requires a precise categorization of the *object* in question, such as a landmine, before any regulative action can be taken.

7 Future of Life Institute (FLI), "Autonomous Weapons: An Open Letter from AI and Robotics Researchers", 28 July 2015, available at: https://futureoflife.org/open-letter-autonomous-weapons/; FLI, "An Open Letter to the United Nations Convention on Certain Conventional Weapons", 21 August 2017, available at: https://futureoflife.org/autonomous-weapons-open-letter-2017/.

8 Mary Wareham, "As Killer Robots Loom, Demands Grow to Keep Humans in Control of Use of Force", Human Rights Watch, 2020, available at: www.hrw.org/world-report/2020/country-chapters/killer-robots-loom-in-2020.

9 The need to arrive at a shared definition of LAWS remains a common notion among the CCW States Parties, and some still view it as a prerequisite for the talks to go anywhere. As an example for this line of thought, see the chair's summary of the discussion of the 2019 GGE meeting: "Some delegations chose to address the issue of definitions, with several different views on the need for definitions – working or otherwise – to make further progress in the work of the Group." UN, *Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems: Chair's Summary*, UN Doc. CCW/GGE.1/2019/3/Add.1, 8 November 2019, p. 3, available at: https://tinyurl.com/y68rzkub.

10 Léonard van Rompaey, "Shifting from Autonomous Weapons to Military Networks", *Journal of International Humanitarian Legal Studies*, Vol. 10, No. 1, 2019, pp. 112–119, available at: https://brill.com/view/journals/ihls/10/1/article-p111_111.xml.

In the case of LAWS, however, the old pattern of defining and then regulating a discrete category of military hardware is not applicable.[11] After all, almost any current and future weapons system can conceivably be endowed with autonomous functions, and no one will be able to tell what any given system's level of dependence on human input is by merely inspecting it from the outside. In the past, bilateral nuclear arms control between the United States and the Soviet Union, later Russia, implemented quantitative arms control by developing precisely defined, shared understandings of counting rules and employing them in verification regimes.[12] Similarly, in the realm of multilateral conventional arms control, the now defunct Treaty on Conventional Armed Forces in Europe relied heavily on defining and counting military hardware items.[13] The challenge regarding LAWS, however, is not met by trying to define a category of weapons system – "LAWS", as separated with a list of specific criteria from "non-LAWS" – and then counting and capping its numbers. In fact, in a modern military's system-of-systems architecture, "some AWS [autonomous weapons system] components are intangible and can be geographically distributed, [so] it is far from clear … where and when an AWS begins and ends".[14] Hence, the challenge, broadly speaking, lies in developing a new norm in order to adjust the relationship between humans and machines in twenty-first-century warfighting. A qualitative rather than quantitative approach is required, which, in turn, requires new diplomatic language to grasp the underlying technological developments, something that neither States Parties nor civil society are well versed in yet.

Luckily, the process of conceptualizing the issue and translating it into diplomatic language has begun, and has already made some progress. After almost six years, the codification of a positive obligation of human control over weapons systems is establishing itself more and more at the heart of the debate. This general notion, gaining prominence in the wake of the call for "*meaningful* human control" originally introduced by the NGO Article 36,[15] is being embraced by civil society as well as a consistently growing number of CCW States Parties. Accordingly, the conceptualization is now finding broad acceptance in both academic literature and the diplomatic debate – not least because the United States and the International Committee of the Red Cross (ICRC) have adopted it. This is not some sort of categorical definition of LAWS (versus

11 Elvira Rosert and Frank Sauer, "How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies", *Contemporary Security Policy*, 30 May 2020, available at: https://tinyurl.com/y23o8lo6.

12 Jozef Goldblat, *Arms Control: The New Guide to Negotiations and Agreements*, SAGE Publications, London, 2002, Chap. 5.

13 Treaty on Conventional Armed Forces in Europe, 19 November 1990, available at: www.osce.org/library/14087.

14 Maya Brehm, *Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapon Systems Under International Humanitarian and Human Rights Law*, Geneva Academy Briefing No. 9, May 2017, pp. 15–16.

15 Richard Moyes, "Key Elements of Meaningful Human Control", Article 36, April 2016, available at: www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf. Article 36 is a member of the KRC.

non-LAWS) via a list of criteria. Instead, it is a functional understanding of the phenomenon.[16]

From a functionalist point of view, the LAWS issue is best understood as one of autonomy *in* a weapons system – that is, of the machine rather than a human performing a certain function (or certain functions) during the system's operation.[17] Every military operation concluding with an attack on a target can be systematized along discrete steps of a kill chain or targeting cycle.[18] This includes finding, fixing, tracking, selecting and engaging the target (as well as assessing the effects afterwards). Many weapons systems are already capable of performing some of the targeting cycle functions without human input or supervision – for example, a drone navigating from one waypoint to the next via satellite navigation and thus performing a part of the "finding" function without having to be remotely controlled. An autonomous weapon, however, completes the entire targeting cycle – including the final stages of selecting and engaging the target with force – without human intervention. In the debate about LAWS, the focus rests mainly on those last two functions (which the ICRC calls "critical"[19]) because most of the effects of weapon autonomy currently under discussion derive from giving up human control over them and handing the decision to use force over to a machine.[20]

A peculiarity of the functional approach is that it reminds us that weapons with autonomy in their critical functions already exist. It renders the issue one of the present, not a concern about future weapons technology. That said, so far weapon autonomy, including the critical functions of target selection and engagement, exists only in limited military applications. The Israeli loitering munition Harpy is probably the best example of an already existing weapons system that – albeit only for the very specific task of engaging radar signatures – selects and engages targets without human supervision or control.[21] Harpy is thus considered an autonomous weapons system, as it completes a targeting cycle without human

16 ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Geneva, 2016; US Department of Defense (DoD), Directive 3000.09, "Autonomy in Weapon Systems", 2012 (amended 2017); Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, W. W. Norton, New York, 2018.

17 Vincent Boulanin and Maaike Verbruggen, *Mapping the Development of Autonomy in Weapon Systems*, Stockholm International Peace Research Institute (SIPRI), Stockholm, 2017, available at: www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_0.pdf.

18 International Panel on the Regulation of Autonomous Weapons (iPRAW), *Focus on Human Control*, iPRAW Report No. 5, August 2019, available at: www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf.

19 ICRC, above note 16, p. 7.

20 For the implications of autonomy in earlier stages of the targeting cycle, which are not discussed further here, see Arthur H. Michel, "The Killer Algorithms Nobody's Talking About", *Foreign Policy*, 20 January 2020, available at: https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/.

21 Israel Aerospace Industries, "HARPY: Autonomous Weapon for All Weather", available at: *www.iai.co.il/p/harpy*. A loitering munition is a weapons system that "loiters" in an area for a prolonged period of time, waiting for targets to appear.

intervention. Terminal defence systems capable of firing without human input, such as Phalanx or Patriot, are additional examples.

Especially in the earlier phases of the CCW process, systems like Phalanx gave rise to attempts by some States Parties to classify terminal defence systems as automatic, in order to prevent them from being drawn into the autonomy debate. In this line of reasoning, automatic systems are stationary and are designed to merely repeat a few pre-programmed actions in case of incoming munitions, whilst operating within tightly set parameters and time frames in structured and controlled environments. Autonomous systems, by contrast, are conceived of as having more time and operational room for manoeuvre.[22] Unfortunately, from an engineering point of view, no such clear and commonly agreed upon delineation between automaticity and autonomy exists; in fact, the terms "automatic" and "autonomous" can be, and often are, used interchangeably to describe a process in which a function is being performed by a machine rather than a human.[23]

A functional understanding renders any attempt at an automatic/autonomous delineation superfluous. This is an advantage in terms of conceptual clarity and simplicity. Also, what initially gave rise to the LAWS debate were concerns regarding autonomous targeting of humans, not targeting of missiles, mortar shells or other munitions.[24] Hence, the crux of the matter, also regarding possible regulation, is not whether a system is to be considered automatic or autonomous, but which targets it attacks. I will return to this line of thought in the sections on why regulation is imperative from an ethical point of view and how it is feasible.

Another advantage of the functionalist view is that it allows us to remain largely agnostic regarding the sophistication or the precise characteristics of the underlying technology. Terminal defence systems, to stick with them as an example, have been in use for decades. So, techniques such as machine learning (or whatever is currently *en vogue* in the wide field that is AI) are not necessarily required to lend a weapons system autonomy (or, for that matter, automaticity) in the critical functions of target selection and engagement. That said, AI obviously is a new and powerful enabler. Weapon autonomy is thus not really new, but recent innovations in AI, such as computer vision, are allowing actors to utilize weapon autonomy on a much larger scale. In effect, it is only recently that autonomous targeting has started to leave its former military niche applications and become adoptable across the board.

22 Frank Sauer, "Stopping 'Killer Robots': Why Now Is the Time to Ban Autonomous Weapons Systems", *Arms Control Today*, Vol. 46, No. 8, 2016, pp. 8–9.

23 To give but one example, the J3016 Levels of Automated Driving standard issued by the Society of Automotive Engineers (SAE) "defines six levels of driving automation" and considers level 5 to be "full vehicle autonomy". SAE, "SAE Standards News: J3016 Automated-Driving Graphic Update", 7 January 2019, available at: www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic.

24 I owe this point to an anonymous reviewer.

Autonomous targeting being used only to destroy incoming munitions may feed into a general worry regarding the ever-faster pace of combat,[25] but it is of no humanitarian concern, and it helps to protect soldiers' lives – a fact that would have to be taken into consideration in any possible regulation of LAWS. In stark contrast, autonomy used unmitigatedly in all kinds of weapons systems, in various operational contexts, and against not just incoming munitions but *any and all targets*, including humans, creates more risks than benefits in sum, as will be argued in more detail below.

To sum up, the first reason why regulating weapon autonomy is difficult derives from the fact that CCW States Parties are challenged not to find some common definition of LAWS but instead to collectively stipulate how future targeting processes can be designed so that human control over the use of military force is retained.[26] In other words, the challenge lies not in delineating a specific weapons category but in generally regulating when a machine should make a certain decision or perform a certain function and when a human should do so, especially at the last two stages of the targeting cycle.

This endeavour is further complicated by the fact that, depending on the operational context and the nature of the target, the manner in which human control is implemented can vary. The combat direction system of a navy frigate, for instance, if designed only to fire at incoming anti-ship missiles and operated in autonomous mode only for brief periods of time whilst in the uncluttered environment of the sea, can be considered as remaining under human control "in design and use"[27] even while performing the critical functions of target selection and engagement autonomously. In contrast, an AI-enabled gun designed to accelerate targeting on a main battle tank in an urban environment would require every single shell fired to be triggered by a human with sufficient situational awareness to make an informed decision in order to be considered as remaining under human control in a meaningful sense.

In short, there is no one-size-fits-all standard of meaningful human control[28] because control by design requires a minimum standard of human–machine interaction, whereas control in use is implemented on a case-by-case

25 The general notion of an action–reaction dynamic created by increasing autonomy was first described by Jürgen Altmann: "Because of very fast action and reaction, autonomous weapon systems would create strong pressures for fast attack if both opponents have got them." Jürgen Altmann, "Military Uses of Nanotechnology: Perspectives and Concerns", *Security Dialogue*, Vol. 35, No. 1, 2004, p. 63.

26 Maya Brehm, "Targeting People", Article 36, November 2019, available at: *www.article36.org/wp-content/uploads/2019/11/targeting-people.pdf*; Richard Moyes, "Autonomy in Weapons Systems: Mapping a Structure for Regulation Through Specific Policy Questions", Article 36, November 2019, available at: www.article36.org/wp-content/uploads/2019/11/regulation-structure.pdf; Richard Moyes, "Target Profiles", Article 36, August 2019, available at: https://t.co/HZ1pvMnIks?amp=1; iPRAW, above note 18; Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*, SIPRI and ICRC, June 2020, available at: www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy.pdf.

27 iPRAW, above note 18, pp. 12–13.

28 Daniele Amoroso and Guglielmo Tamburrini, *What Makes Human Control over Weapon Systems "Meaningful"?*, International Committee for Robot Arms Control, August 2019, available at: www.icrac.net/wp-content/uploads/2019/08/Amoroso-Tamburrini_Human-Control_ICRAC-WP4.pdf.

basis.[29] And while defending against incoming munitions remains a worthwhile application of autonomy in a weapon's critical functions, the debate around LAWS suggests that arguably all other targets might require more human involvement and control. This renders the issue of LAWS more abstract, complex, and intellectually and diplomatically challenging than, for instance, conceptualizing a prohibition against anti-personnel landmines.

The second reason why regulating autonomy in weapons systems is difficult is the enormous military significance ascribed to it. This pertains to the five permanent members of the UN Security Council, but also to other countries with technologically advanced militaries such as, to give but two examples, Israel and Australia. The hurdle itself is not new, of course. It is observable in other regulatory processes of the recent past, such as the ones on landmines, cluster munitions and blinding laser weapons, with the latter being achieved within the CCW framework.[30] However, blinding lasers always represented an exotic niche capability that States could forego without great perceived military costs. Landmines and cluster munitions, too, had specific fields of use and were at least partly substitutable. This is not the case with weapon autonomy. Its impact is perceived to be game-changing for militaries in at least two domains of major significance.

First, weapon autonomy promises a whole range of operational and strategic advantages by rendering constant control and communication links obsolete. The militarily beneficial effects of this innovation, proponents argue, are manifold. It allows for a new level of force multiplication (with a single human operating several, dozens or hundreds of systems at once), creates the possibility of "swarming" (opening up new possibilities for overwhelming the enemy and evading counter-fire),[31] reduces personnel costs and increases a system's stealth in the electromagnetic spectrum (offering insurance against communications disruption or hijacking). Most importantly, however, it removes the inevitable delay between a remote human operator's command and the system's response. Swifter reaction times generate a key tactical advantage over a remotely controlled and thus slower-reacting adversarial system. In fact, the promise of gaining the upper hand by allowing for the completion of the targeting cycle at machine speed is arguably the main motivation behind increasing weapon autonomy.[32] Second, weapon autonomy promises to help prevent some of the atrocities of war and render warfare more humane. Since machines know no fear, stress or fatigue and are devoid of negative human emotions, they never panic, overreact or seek revenge, it is argued. Since they lack a self-preservation instinct,

29 I am thankful to an anonymous reviewer for this clarification.

30 E. Rosert and F. Sauer, above note 11.

31 Paul Scharre, *Robotics on the Battlefield, Part II: The Coming Swarm*, Center for a New American Security (CNAS), October 2014, available at: https://tinyurl.com/yy4gxs43; Maaike Verbruggen, *The Question of Swarms Control: Challenges to Ensuring Human Control over Military Swarms*, EU Non-Proliferation and Disarmament Consortium, Non-Proliferation and Disarmament Paper No. 65, December 2019.

32 Michael C. Horowitz, "When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability", *Journal of Strategic Studies*, Vol. 42, No. 6, 2019; Jürgen Altmann and Frank Sauer, "Autonomous Weapon Systems and Strategic Stability", *Survival*, Vol. 59, No. 5, 2017.

they can always delay returning fire. They supposedly allow not only for greater restraint but also – eventually, when technology permits – for better discrimination between civilians and combatants, thus resulting in the potential to apply military force in stricter accordance with the rules of international humanitarian law (IHL). This would add up to an overall ethical benefit – in a utilitarian sense.[33] In sum, the perceived transformative potential of weapon autonomy and the quantity and quality of military benefits ascribed to it render it more significant when compared to specific weapon categories, such as landmines or cluster munitions, that have been subject to humanitarian disarmament in the recent past.

In light of such tempting promises and the already ongoing, expanding efforts begun in the United States and China (as well as Russia, to a slightly lesser extent) to incorporate civilian innovation for the purposes of increasing weapon autonomy,[34] there is currently little appetite in those States to forego some of the conceived military benefits of this, in their view, critical step in military technology.[35] The United States' unwavering position, for instance, is to keep exploring an IHL-compliant use of autonomy in the critical functions of weapons systems.[36] Some middle powers are not keen on regulation at this point either. India, for example, senses an opportunity for leapfrogging and closing the technological gap between itself and the high-tech militaries of the world.[37] In fact, after the campaigns against landmines and cluster munitions, and the current humanitarian disarmament efforts in the areas of the arms trade (Arms Trade Treaty) and nuclear weapons (Treaty on the Prohibition of Nuclear Weapons), some diplomats in Geneva seem outright annoyed by the KRC's push for yet another prohibition treaty.

In addition, geopolitics in general are not conducive to achieving new arms control breakthroughs. The Treaty on the Prohibition of Nuclear Weapons, which will come into force on 22 January 2021, is seen by some as the exception to this rule,

33 Ronald C. Arkin, "Ethical Robots in Warfare", *IEEE Technology and Society Magazine*, Vol. 28, No. 1, 2009; Ronald C. Arkin, "The Case for Ethical Autonomy in Unmanned Systems", *Journal of Military Ethics*, Vol. 9, No. 4, 2010; Ronald C. Arkin, "Governing Lethal Behavior in Robots", *IEEE Technology and Society Magazine*, Vol. 30, No. 4, 2011; United States, *Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems: Working Paper Submitted by the United States of America*, UN Doc. CCW/GGE.1/2019/WP.5, 28 March 2019, available at: https://tinyurl.com/y4xe7tmc.

34 Elsa B. Kania, "In Military-Civil Fusion, China Is Learning Lessons from the United States and Starting to Innovate", *The Strategy Bridge*, 27 August 2019, available at: https://thestrategybridge.org/the-bridge/2019/8/27/in-military-civil-fusion-china-is-learning-lessons-from-the-united-states-and-starting-to-innovate; Elsa B. Kania, *"AI Weapons" in China's Military Innovation*, Brookings Institution, April 2020, available at: www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf; Frank Sauer, "Military Applications of Artificial Intelligence: Nuclear Risk Redux", in Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, SIPRI, Stockholm, 2019.

35 Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power", *Texas National Security Review*, Vol. 1, No. 3, 2018, available at: https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/; Zachary Davis, "Artificial Intelligence on the Battlefield: Implications for Deterrence and Surprise", *Prism*, Vol. 8, No. 2, 2019, pp. 117–121.

36 United States, above note 33. I would like to thank an anonymous reviewer for highlighting this.

37 Shashank R. Reddy, *India and the Challenge of Autonomous Weapons*, Carnegie Endowment for International Peace, June 2016, p. 12, available at: https://carnegieendowment.org/files/CEIP_CP275_Reddy_final.pdf.

but its effects on the multilateral nuclear arms control architecture are still unclear. And at the same time, already existing multilateral and bilateral agreements and treaties are eroding, with some already lost – this list includes the terminated Intermediate-Range Nuclear Forces Treaty, the faltering Joint Comprehensive Plan of Action with Iran, the struggling Open Skies Treaty, and soon, potentially, NewSTART, the only remaining bilateral nuclear arms control treaty between Russia and the United States. Getting a new binding international legal instrument out of the CCW would be challenging in a normal, less frosty geopolitical landscape. The current global arms control winter makes it a daunting feat.

Nevertheless, regulating weapon autonomy in a manner that curbs autonomy in the critical functions and keeps them under human control is sorely needed. After all, the consequences of inaction would be dire because the mid- and long-term strategic and ethical risks of unshackled weapon autonomy far outweigh the desired short-term military gains highlighted above. I will argue this in two steps below, by first focusing on a number of operational and strategic implications and subsequently evaluating the ethical implications of weapon autonomy in regard to human dignity.

Why regulating weapon autonomy is imperative: Strategic implications

The potential impact of unregulated weapon autonomy on military operations, as well as on global peace and strategic stability as a whole, has drawn scholarly attention for quite a while.[38] This body of literature suggests that the implications of regulatory inaction and an ensuing rapid diffusion of weaponized autonomy-enabling technology range from new military vulnerabilities to increased risks of instability and escalation at both the operational and the strategic level.[39] Hence it is in fact especially the great powers that should see it as being not only their responsibility but also in their genuine self-interest[40] to curb this destabilizing chain of effects.

38 See J. Altmann, above note 25; Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Ashgate, Farnham, 2009, Chap. 6; Jean-Marc Rickli, *Some Considerations of the Impact of LAWS on International Security: Strategic Stability, Non-State Actors and Future Prospects*, presentation at CCW Meeting of Experts on LAWS, Geneva, 16 April 2015, available at: https://tinyurl.com/y4fjozpf; Paul Scharre, *Autonomous Weapons and Operational Risk*, CNAS Ethical Autonomy Project, Washington, DC, February 2016, available at: https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf; Wendell Wallach, "Toward a Ban on Lethal Autonomous Weapons: Surmounting the Obstacles", *Communications of the ACM*, Vol. 60, No. 5, 2017, p. 31; Irving Lachow, "The Upside and Downside of Swarming Drones", *Bulletin of the Atomic Scientists*, Vol. 73, No. 2, 2017; J. Altmann and F. Sauer, above note 32; Paul Scharre, "Autonomous Weapons and Stability", PhD thesis, King's College London, March 2020, available at: https://kclpure.kcl.ac.uk/portal/files/129451536/2020_Scharre_Paul_1575997_ethesis.pdf.

39 The following section draws on J. Altmann and F. Sauer, above note 32; F. Sauer, above note 34; Aaron Hansen and Frank Sauer, "Autonomie in Waffensystemen: Chancen und Risiken Für die US-Sicherheitspolitik", *Zeitschrift für Außen- und Sicherheitspolitik*, Vol. 12, No. 2, 2019.

40 For the general argument, see Hedley Bull, *The Anarchical Society: A Study of Order in World Politics*, Macmillan, London, 1977. For the case of AI, see Elsa B. Kania and Andrew Imbrie, "Great Powers

Technology diffusion

To get an idea of the expectable diffusion of technology in the field of LAWS, drones can serve as an indicator.[41] China in particular is not only investing but also exporting in this sector.[42] According to data collected by the New America Foundation,[43] twelve countries have conducted drone strikes and thirty-eight now possess armed drones, as do several non-State actors such as Hamas, Hezbollah, the Houthi rebels and the so-called Islamic State group.

Drone technology spreads comparably quickly because of its dual-use nature. Autonomy is dual-use, too. Weapon autonomy – provided that the platform in question contains the necessary sensors and actuators – mainly comes down to software, which can be transferred and reproduced at close to no cost and is vulnerable to theft via cyber attacks.[44] Consequently, the adoption of software-enabled autonomous functions can be expected to spread rapidly in the existing military hardware ecosystem. Also, the main innovators of autonomy are tech companies and universities, not the defence industry, so it is questionable whether any one country's military can remain the "fast leader"[45] as envisioned by members of the US defence establishment, for example. After all, the US government is not the only one incorporating civilian tech for military purposes by approaching tech firms such as Google, Microsoft and Amazon; China, for example, is doing the same with Tencent, Ali Baba and Baidu.[46] Hence it is highly unlikely that some sort of monopoly in this field, like the one the United States held with stealth technology in the past, is possible.

New operational vulnerabilities

The flip side of the force multiplication effect that militaries hope for with this diffusion-prone technology is scalability, creating the potential for weaker parties to change the power dynamics between themselves and their adversaries. The

Must Talk to Each Other about AI", *Defense One*, 28 January 2020, available at: www.defenseone.com/ideas/2020/01/great-powers-must-talk-each-other-about-ai/162686/?oref=d-river.

41 Frank Sauer and Niklas Schörnig, "Killer Drones: The Silver Bullet of Democratic Warfare?", *Security Dialogue*, Vol. 43, No. 4, 2012; Matthew Fuhrmann and Michael C. Horowitz, "Droning On: Explaining the Proliferation of Unmanned Aerial Vehicles", *International Organization*, Vol. 71, No. 2, 2017; Andrea Gilli and Mauro Gilli, "The Diffusion of Drone Warfare? Industrial, Organizational, and Infrastructural Constraints", *Security Studies*, Vol. 25, No. 1, 2016.

42 Defense Science Board, *The Role of Autonomy in DoD Systems*, 2012, pp. 69–71.

43 New America, "World of Drones", available at: www.newamerica.org/in-depth/world-of-drones/.

44 Sydney J. Friedberg, "Robot Wars: Centaurs, Skynet, and Swarms", *Breaking Defense*, 31 December 2015, available at: http://breakingdefense.com/2015/12/robot-wars-centaurs-skynet-swarms/; Thomas G. Mahnken, *Technology and the American Way of War Since 1945*, Columbia University Press, New York, 2008, p. 123.

45 Robert O. Work, "Robert Work Talks NATO's Technological Innovation and the DoD", *CNAS Brussels Sprouts Podcast*, 11 January 2018, available at: www.cnas.org/publications/podcast/robert-work-talks-natos-technological-innovation-and-the-dod.

46 Defense Science Board, *Summer Study on Autonomy*, 2016, p. 45; Elsa B. Kania, *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*, CNAS, Washington, DC, November 2017, available at: https://s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235805; E. B. Kania, "In Military-Civil Fusion", above note 34.

weaponization of simple, commercially available drones by the so-called Islamic State group and the attack against the Saudi Aramco oil facility give non-autonomous foretastes of what is to come and demonstrate that, advanced aerial defence capabilities notwithstanding, new vulnerabilities are on the rise. Particularly from the point of view of US ground forces, having to face serious threats from above after decades of air dominance represents a paradigm shift.[47] The United States is thus already being forced to rethink its air defence capabilities by intensifying the development of lasers and microwaves. Conventional solutions, such as Stinger missiles, are not only unsuitable for defence against the swarms of small, cheap, disposable drones that autonomy now renders possible, they are also not cost-effective. Whether the new defensive systems can remedy this situation is still open for debate.[48] Suffice it to say that the combination of cheap unmanned systems, autonomy and swarm behaviour creates new risks in general, for troops on the battlefield, for command and control infrastructure and for senior leaders in so-called decapitation scenarios.[49]

As argued above, the possible elimination of the remote control link is a key incentive for having more autonomy in a weapons system – but handing control over to the machine opens up new attack vectors as well. Feeding the system spoofed GPS data is one example; in 2011, Iran was seemingly able to hijack an autonomously navigating US drone in this manner.[50]

What is more, systems relying on machine learning that makes use of deep neural networks,[51] which currently represent the state of the art in the field of computer vision, are also particularly susceptible to manipulation. Some reflective tape on a stop sign, for example, can fool a self-driving car's image recognition system. This susceptibility to error is a tricky but eventually solvable problem in a civilian application such as self-driving cars. Training data is plentiful and easily available, and self-driving cars are designed to operate cooperatively in a tightly regulated environment. The battlefield presents itself very differently – it is characterized by a paucity of data and much greater degrees of unpredictability and vulnerability.[52] After all, an adversary will of course always try to deceive and

47 Kelley Sayler, *A World of Proliferated Drones: A Technology Primer*, CNAS, Washington, DC, 2015, p. 29.

48 Sebastien Roblin, "The U.S. Army Needs More Anti-Aircraft Weapons – and Fast", *War is Boring*, 22 January 2018, available at: http://warisboring.com/the-u-s-army-needs-more-anti-aircraft-weapons-and-fast/.

49 David Barno and Nora Bensahel, "The Drone Beats of War: The U.S. Vulnerability to Targeted Killings", *War on the Rocks*, 21 January 2020, available at: https://warontherocks.com/2020/01/the-drone-beats-of-war-the-u-s-vulnerability-to-targeted-killings/. A decapitation scenario is a scenario in which an attacker aims to destroy or destabilize an opponent's leadership and command and control structure in order to severely degrade or destroy its capacity for (nuclear) retaliation.

50 Sydney J. Friedberg, "Drones Need Secure Datalinks to Survive vs. Iran, China", *Breaking Defense*, 10 August 2012, available at: http://breakingdefense.com/2012/08/drones-need-secure-datalinks-to-survive-vs-iran-china/.

51 For a critical overview, see Gary Marcus, "Deep Learning: A Critical Appraisal", New York University, 2 January 2018, available at: https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf.

52 Michał Klincewicz, "Autonomous Weapons Systems, the Frame Problem and Computer Security", *Journal of Military Ethics*, Vol. 14, No. 2, 2015; Anh Nguyen, Jason Yosinski and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images",

tamper with your systems. Research on adversarial examples[53] suggests that computer vision will leave autonomous weapons systems open to manipulation by tampering with the environment that the machines perceive[54] or even by retraining them if they continue learning during their deployment.[55] Facial recognition for targeting purposes would be quite easy to fool and defeat, too, as the rapid development of countermeasures against domestic surveillance demonstrates.[56]

As the complexity of the software driving a weapons system increases, so does the number of bugs it contains. Such programming errors can have critical effects, including friendly fire.[57] Normal accidents theory[58] suggests that mistakes are basically inevitable. They occur even in domains with extremely high safety and security standards, such as nuclear power plants or manned space travel.[59] The software industry can currently reduce the number of bugs to 0.1–0.5 errors per 1,000 lines of code, which means that complex military systems with several million lines of code, such as the software for the F-35 fighter jet, come with thousands of software errors.[60] The unavoidable reality of regularly having to update the systems complicates this issue further;[61] it is a potential source of new bugs and new errors arising from interactions between newer and older software. Machine learning systems generate specific difficulties because they present themselves as "black boxes" which cannot be debugged the way conventional software can, meaning that they cannot be selectively cleared of specific errors.[62]

Finally, weapon autonomy evokes a new proneness to errors in regard to any remaining interactions with human operators. Here, automation bias comes into play – that is, the uncritical, unfounded trust in the functioning of a

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436; Z. Davis, above note 35, pp. 121–122.

53 Ivan Evtimov *et al.*, "Robust Physical-World Attacks on Deep Learning Models", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, available at: https://arxiv.org/pdf/1707.08945.pdf.

54 See the by now famous example of the turtle mistaken for a rifle, in Anish Athalye, Logan Engstrom, Andrew Ilyas and Kevin Kwok, "Synthesizing Robust Adversarial Examples", *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, 2018, available at: https://arxiv.org/pdf/1707.07397.pdf.

55 Defense Science Board, above note 46, p. 28; Vadim Kozyulin, "International and Regional Threats Posed by the LAWS: Russian Perspective", PIR Center for Policy Studies, April 2016, available at: https://tinyurl.com/y4qslefc; P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 14.

56 Melissa Hellmann, "Special Sunglasses, License-Plate Dresses: How to Be Anonymous in the Age of Surveillance", *Seattle Times*, 12 January 2020, available at: www.seattletimes.com/business/technology/special-sunglasses-license-plate-dresses-juggalo-face-paint-how-to-be-anonymous-in-the-age-of-surveillance/.

57 P. Scharre, above note 38, p. 21.

58 Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, Basic Books, New York, 1984.

59 John Borrie, *Security, Unintentional Risk, and System Accidents*, United Nations Institute for Disarmament Research (UNIDIR), Geneva, 15 April 2016, available at: https://tinyurl.com/yyaugayk; P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38.

60 P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 13.

61 UNIDIR, *The Weaponization of Increasingly Autonomous Technologies in the Maritime Environment: Testing the Waters*, UNIDIR Resources No. 4, Geneva, 2015, p. 8.

62 G. Marcus, above note 51, pp. 10–11.

system.[63] Put simply, autonomous systems might operate incorrectly over periods of time without anyone noticing.[64] A human making a mistake can understand the situation and correct for it, but unmonitored LAWS will not be able to understand and critically reflect in real time the way humans do.[65] This gives rise to risks of unintended military escalation.

Escalation risks and crisis instability

Weapons systems operating without human control generate not only new vulnerabilities but also unpredictability due to unforeseeable interactions with their environment, in turn creating new risks of unintended, unwanted escalation.[66] In that regard, the interaction between two or more autonomous systems is to be considered in particular. High-frequency trading[67] provides a useful analogue, because unforeseen and unwanted interaction processes between two or more autonomously operating trading algorithms occur on a regular basis, sometimes causing so-called "flash crashes" and resulting in financial losses. This can be remedied with regulation of the financial market to an extent, but without internationally binding regulation of autonomy on the battlefield, unforeseeable interactions of LAWS might end in unintended use of force at machine speed, even accidental war before humans can intervene.[68] This risk is not in some distant future. At the Dubai Airshow in 2019, the chief of staff of the US Air Force, General David Goldfein, presented the simulated engagement of an enemy navy vessel with a next-to-fully automated kill chain. The vessel was first picked up by a satellite, then target data was relayed to airborne surveillance as well as command and control assets. A US Navy destroyer was then tasked with firing a missile, the only remaining point at which this targeting cycle involved a human decision, with the rest of the "kill chain … completed machine to machine, at the speed of light".[69] Any machine error in such a system would, if left uncorrected

63 See, for the example of the Patriot missile defence system, John K. Hawley, *Patriot Wars: Automation and the Patriot Air and Missile Defense System*, CNAS, Washington, DC, January 2017, available at: www.cnas.org/publications/reports/patriot-wars.

64 P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 31; Noel Sharkey and Lucy Suchman, "Wishful Mnemonics and Autonomous Killing Machines", *Proceedings of the AISB*, Vol. 136, 2013, pp. 16–17.

65 Defense Science Board, above note 42, p. 15.

66 André Haider and Maria Beatrice Catarrasi, *Future Unmanned System Technologies: Legal and Ethical Implications of Increasing Automation*, Joint Air Power Competence Centre, November 2016, p. 10, available at: www.japcc.org/wp-content/uploads/Future_Unmanned_System_Technologies_Web.pdf; ICRC, *Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon System [s]*, Geneva, 11 April 2016, p. 3, available at: www.icrc.org/en/download/file/21606/ccw-autonomous-weapons-icrc-april-2016.pdf.

67 Gary Shorter and Rena S. Miller, *High-Frequency Trading: Background, Concerns, and Regulatory Developments*, Congressional Research Service, 19 June 2014, available at: *https://fas.org/sgp/crs/misc/R43608.pdf*.

68 P. Scharre, *Autonomous Weapons and Operational Risk*, above note 38, p. 53; J. Altmann and F. Sauer, above note 32, pp. 128–132.

69 "Video: Here's How the US Air Force Is Automating the Future Kill Chain", *Defense News*, 2019, available at: www.defensenews.com/video/2019/11/16/heres-how-the-us-air-force-is-automating-the-future-kill-chain-dubai-airshow-2019/.

by a human due to automation bias, propagate quickly. It stands to reason that the error would propagate "at the speed of light" as well, were the human to be removed. A recent wargaming exercise conducted by the RAND Corporation underlines the risks of crisis instability and unintended escalation; in this exercise, simulated forces were set "on 'full auto' to signal resolve …[,] in one case lead[ing] to inadvertent escalation. Systems set to autonomous mode reacted with force to an unanticipated situation in which the humans did not intend to use force."[70]

Humans are more resistant to mass error than machines. Also, humans, despite being slower and sometimes making mistakes, are better managers than machines. They have the capacity to grasp an unusual situation and understand its context as well as to reflect on a decision, its genesis, its implications and the weight of the responsibility that accompanies it. In terms of crisis management, all this makes humans superior to machines, which so far are only capable of recognizing patterns and executing predefined actions, and which reach superhuman performance only in those narrowly defined scenarios for which they were specifically trained. By removing human control, the distinct role of humans as a versatile fail-safe mechanism is lost.

The prominent case of Lieutenant Colonel Stanislav Petrov renders this evident. The 1983 NATO exercise Able Archer was misunderstood by the Soviets as a cover for an attack with tactical nuclear forces. During this time, a Soviet early-warning satellite registered first one, then a couple more US nuclear intercontinental ballistic missile launches. Petrov, the watch officer in charge at the time, decided (correctly) that this had to be a false alarm and gave the all-clear up the chain of command, thus preventing further, potentially nuclear escalation in this tense situation. Petrov's decision could not have been made by a completely automated system. He later testified that he had arrived at his decision by following a gut feeling, by wondering about the nature of the supposed strike, and by drawing on his past experiences with the early-warning system that he deemed not fully trustworthy.[71] If the human on the destroyer in the next-to-fully automated kill chain presented by General Goldfein were ever to be removed, fully actualizing the key advantage of weapon autonomy that is fighting at machine speed, the "Petrov effect" would be lost. While, in that conventional scenario, this would not mean the inadvertent use of nuclear

70 Yuna H. Wong *et al.*, *Deterrence in the Age of Thinking Machines*, RAND Corporation, 2020, p. xi, available at: www.rand.org/pubs/research_reports/RR2797.html.

71 Bruce G. Blair, *The Logic of Accidental Nuclear War*, Brookings Institution, Washington, DC, 1993, p. 181; Richard Rhodes, *Arsenals of Folly: The Making of the Nuclear Arms Race*, Simon & Schuster, London, 2008, pp. 165–166; David E. Hoffman, *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*, Doubleday, New York, 2009, pp. 6–11, 94–95; Mark Gubrud, "Stopping Killer Robots", *Bulletin of the Atomic Scientists*, Vol. 70, No. 1, 2014; Michael C. Horowitz, Paul Scharre and Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence*, working paper, December 2019, pp. 13–14, available at: https://arxiv.org/ftp/arxiv/papers/1912/1912.05291.pdf; Paul Scharre, "A Million Mistakes a Second", *Foreign Policy*, 12 September 2018, available at: https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/.

weapons, strategic stability is nevertheless already being affected by the effort to increase autonomy in military systems.

Strategic instability

It has recently been suggested that AI as a decision-making aid to humans might help improve the performance of nuclear early-warning and command and control systems, thus reducing the risk of false alarms and inadvertent nuclear use.[72] That said, calls for complete automation in the nuclear realm – that is, for handing over the decision to use nuclear weapons from humans to machines – are practically non-existent.[73] But even with the proverbial push of the button not yet delegated to algorithms, the rush to increase autonomy in military applications and to automatize military processes increases the risk of nuclear stability.[74]

For instance, the increasing capacities of conventional weapons systems – including weapon autonomy – are beginning to affect the strategic level. This development has been described as the increasing "entanglement" of the nuclear and the conventional realm resulting, for example, from "non-nuclear threats to nuclear weapons and their associated command, control, communication, and information (C3I) systems".[75] Simply put, advanced conventional capabilities increasingly allow for nuclear assets to be put at risk. Autonomy in conventional weapons systems is one such advanced capability, thus feeding into this increasing entanglement and, in turn, deteriorating strategic stability.

One specific illustration of this dynamic is the deployment of stealthy unmanned aerial vehicles and the use of "swarming". Perdix is a swarming test program pursued by the US Air Force. In the future, drone swarms of this type might facilitate the search for dispersed mobile missile launchers. Another example is the use of maritime autonomous systems for hunting nuclear-powered ballistic missile submarines, known as SSBNs. The DARPA-funded Anti-Submarine Warfare Continuous Trail Unmanned Vessel is a program that resulted in the development of an autonomous trimaran called Sea Hunter, which

72 M. Horowitz, P. Scharre and A. Velez-Green, above note 71, p. 14; Philip Reiner and Alexa Wehsner, "The Real Value of Artificial Intelligence in Nuclear Command and Control", *War on the Rocks*, 4 November 2019, available at: https://warontherocks.com/2019/11/the-real-value-of-artificial-intelligence-in-nuclear-command-and-control/. On the resulting cyber vulnerabilities, see James Johnson, "The AI–Cyber Nexus: Implications for Military Escalation, Deterrence and Strategic Stability", *Journal of Cyber Policy*, Vol. 4, No. 3, 2019.

73 With the exception of Adam Lowther and Curtis McGiffin, "America Needs a 'Dead Hand'", *War on the Rocks*, 16 August 2019, available at: https://warontherocks.com/2019/08/america-needs-a-dead-hand/.

74 Edward Geist and Andrew J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?*, RAND Corporation, 2018, available at: www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE296/RAND_PE296.pdf; Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su and Moa Peldán Carlsson, *Artificial Intelligence, Strategic Stability and Nuclear Risk*, SIPRI, Stockholm, June 2020, available at: www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

75 James M. Acton (ed.), *Entanglement: Chinese and Russian Perspectives on Non-Nuclear Weapons and Nuclear Risks*, Carnegie Endowment for International Peace, 2017, p. 1, available at: http://carnegieendowment.org/files/Entanglement_interior_FNL.pdf.

is currently being tested by the US Navy. Its ability to detect and pursue SSBNs could potentially limit the second-strike capabilities of other nuclear powers.

These capabilities are just emerging, and neither Perdix nor Sea Hunter, nor their successors, will single-handedly destabilize the global nuclear order. Also, the hypothesis that systems such as Sea Hunter would render the oceans "transparent",[76] virtually nullifying the utility of sea-launched nuclear weapons as a reliable second-strike asset, is hotly debated. Nevertheless, the mere perception of nuclear capabilities becoming susceptible to new risks from the conventional realm is bound to sow distrust between nuclear-armed adversaries. Furthermore, a system like Sea Hunter demonstrates how autonomous weapon technologies are expediting the completion of the targeting cycle, thus putting the adversary under additional pressure and potentially creating "use-them-or-lose-them" scenarios with regard to executing a nuclear second strike.

The entanglement problem, which weapon autonomy is feeding into, is further aggravated by an increasing political willingness to use nuclear means to retaliate against non-nuclear attacks on early-warning and control systems or the weapons themselves. The Trump administration's nuclear posture review[77] signals that the United States may, from now on, respond with nuclear means to significant, non-nuclear strategic attacks (moving away from a "single-purpose" nuclear deterrence framing for nuclear weapons). Russia has already held this position for some time due to the United States' advantage in conventional weapons technology. This does not bode well for stability between the two largest nuclear powers.

To sum up this section, weapon autonomy not only promises military benefits but also creates new vulnerabilities and, more importantly, contributes to an overall accumulation of strategic risk and instability. Increasing operational speed beyond the capability of human cognition removes humans as a valuable fail-safe against unwanted escalation.

Why regulating weapon autonomy is imperative: Ethical implications

The discussion on LAWS in the CCW is slanted towards IHL and the legal implications, as visible, for example, in the eleven guiding principles adopted in the 2019 CCW States Parties meeting report.[78] In the preamble preceding the list of principles, the report states that "international law, in particular the United Nations Charter and International Humanitarian Law … as well as relevant ethical perspectives, should guide the continued work of the Group". Nevertheless, only five out of the eleven guiding principles are legal in nature,

76 Sebastian Brixey-Williams, "Will the Atlantic Become Transparent?", November 2016, available at: https://britishpugwash.org/wp-content/uploads/2016/11/Will-the-Atlantic-become-transparent-.pdf.
77 DoD, *Nuclear Posture Review 2018*, 2018, p. 21, available at: https://tinyurl.com/yc7lu944.
78 CCW Meeting Final Report, above note 2, p. 10.

and not a single one contains a reference to ethical implications. The legal strand of the debate is undoubtedly important, especially since it allows for systematically interrogating the claim that autonomy in weapons renders warfare more IHL-compliant. As it stands, technology is now unable to fulfil this promise of increased IHL compliance,[79] though it might eventually be capable of doing so. But be that as it may, an ethical point of view suggests that the LAWS issue has deeper roots than mere IHL compliance anyway, because it touches upon fundamental norms that go above and beyond the laws of war.[80] Ethical implications were more systematically considered in their own right at the very beginning of the LAWS debate at the UN. In 2013, when the issue was raised in the Human Rights Council, Special Rapporteur Christof Heyns[81] objected against LAWS, arguing that they violate human dignity.

Universal human dignity

That the use of LAWS would be a violation of human dignity has been argued by various scholars of moral philosophy and technology.[82] The notion was picked up by the KRC[83] and lately has also been reiterated by the ICRC.[84] Opposing weapon autonomy on grounds of human dignity has drawn some scrutiny,[85] and the supposed "awkwardness"[86] of this stance is commonly substantiated by pointing out that several meanings of dignity exist and that there is no commonly agreed-upon definition of dignity.

However, being hard to define but relevant and even crucially important is a characteristic of many normative concepts, including many legally codified ones. Cornerstones of IHL such as civilian-ness, which is defined only *ex negativo*, or proportionality, which is not quantifiable and is assessable only on a case-by-case

79 Frank Sauer, Daniele Amoroso, Noel Sharkey, Lucy Suchman and Guglielmo Tamburrini, *Autonomy in Weapon Systems: The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy*, Heinrich Böll Foundation Publication Series on Democracy, Vol. 49, 2018, pp. 23–32, available at: www.boell.de/sites/default/files/boell_autonomy-in-weapon-systems_v04_kommentierbar_1.pdf.

80 The following section draws on Elvira Rosert and Frank Sauer, "Prohibiting Autonomous Weapons: Put Human Dignity First", *Global Policy*, Vol. 10, No. 3, 2019.

81 Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, UN Doc. A/HRC/23/47, 2013, p. 17, available at: https://digitallibrary.un.org/record/755741/files/A_HRC_23_47-EN.pdf.

82 Peter Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making", *International Review of the Red Cross*, Vol. 94, No. 886, 2012; Robert Sparrow, "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems", *Ethics & International Affairs*, Vol. 30, No. 1, 2016.

83 KRC, *Making the Case: The Dangers of Killer Robots and the Need for a Preemptive Ban*, 2016, pp. 21–25.

84 ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, Geneva, 3 April 2018, available at: www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf.

85 Amanda Sharkey, "Autonomous Weapons Systems, Killer Robots and Human Dignity", *Ethics and Information Technology*, Vol. 21, No. 2, 2019.

86 Deane-Peter Baker, "The Awkwardness of the Dignity Objection to Autonomous Weapons", *The Strategy Bridge*, 6 December 2018, available at: https://thestrategybridge.org/the-bridge/2018/12/6/the-awkwardness-of-the-dignity-objection-to-autonomous-weapons.

basis, are examples.[87] Human dignity, too, is contained in various international legal documents. The Universal Declaration of Human Rights refers to it in its preamble, as does the UN Charter. It is also invoked in national bodies of law, as well as court decisions. The key example here is Germany's basic law Article 1(1), which states human dignity's inviolability and prohibits the treatment of humans as objects or means to an end, being referenced in a 2006 landmark decision by the German Constitutional Court. The judges struck down a federal law that would have allowed the German air force to shoot down a hijacked aeroplane that the hijackers may have intended to use as a weapon to kill people on the ground. The Court deemed it unconstitutional to use the aeroplane passengers as mere instruments to try to achieve another, albeit worthy, goal.[88]

The key ethical implication of weapon autonomy in a weapons system's critical functions is thus that allowing algorithms to make kill decisions violates human dignity because the victim is reduced to an object, a mere data point fed to an automated, indifferent killing machine.

It is worth spelling out that this objection is valid even if civilians (or other non-combatants) remain unharmed. After all, narrowing the focus solely to the possibility that LAWS might not be able to make proper – or even better – distinctions between combatants and civilians, a cornerstone of the legal case against LAWS discussed in the CCW, loses sight of the fact that combatants, too, are imbued with human dignity. In other words, weapon autonomy raises a more fundamental concern than the legal strand of the LAWS debate suggests, because "successfully discerning combatants from noncombatants is far from the only issue".[89]

As a general rule, the use of LAWS against humans can be deemed an unacceptable infringement on human dignity because delegating the decision to kill to an algorithm devalues human life.[90] Exceptions from this rule would only be conceivable if they were explicitly not made on the basis of weighing bare lives against each other and then deliberately opting for algorithmic killing. An example for such a boundary case could be a sailor's reliance on weapon autonomy in a narrowly bound scenario of desperate self-defence. If the aforementioned navy frigate[91] were to be under a saturation attack by anti-ship missiles and, potentially, also manned aircraft, then inadvertently endangering human life by relying on autonomous defensive fire for the survival of the ship and its crew could be considered acceptable *ex post*.

Generally speaking again, being killed as the result of algorithmic decision-making matters for the person dying because a machine taking a human life has no

87 I would like to thank an anonymous reviewer for specifying these properties of civilian-ness and proportionality.

88 F. Sauer *et al.*, above note 79, p. 33.

89 Heather M. Roff, "The Strategic Robot Problem: Lethal Autonomous Weapons in War", *Journal of Military Ethics*, Vol. 13, No. 3, 2014, p. 219.

90 Christof Heyns, "Autonomous Weapons in Armed Conflict and the Right to a Dignified Life: An African Perspective", *South African Journal on Human Rights*, Vol. 33, No. 1, 2017, pp. 62–63.

91 See text related to above note 27.

conception of what its action means: "In the absence of an intentional and meaningful decision to use violence, the resulting deaths are meaningless and arbitrary."[92] In other words, the least we can do when killing another human being in war is to recognize this death of a fellow member of our species and put the weight accompanying this decision onto our conscience. The mindlessness of machines killing humans based on software outputs strips the latter of their right to be recognized as humans in the moment of death. This also matters for society at large. Modern warfare, especially in democracies, already decouples societies from warfighting in terms of political and financial costs.[93] A society outsourcing moral costs by no longer even concerning itself with the act of killing, with no individual combatants' psyches burdened by the accompanying responsibility, crosses a moral line. It risks losing touch with fundamental humanitarian values such as the right to a dignified life and respect towards fellow human beings.[94]

To sum up this section, while the legal verdict on weapon autonomy increasing IHL compliance is still out and will be for some time, a more fundamental objection against LAWS based on deontological limits is valid today.

How regulating weapon autonomy is feasible: Fostering a human control norm

The United States and China are demonstrating awareness of the strategic risks of unmitigated weapon autonomy. The US directive on weapon autonomy,[95] albeit attempting to square the circle of using autonomy while not inviting the accompanying risks, can be interpreted this way. China has coined the term "battlefield singularity", a dreaded situation in which war waged at machine speed is too fast for human cognition to keep up.[96] Nevertheless, the current great power rivalry between the United States, China and Russia, all racing for dominance in the field of military AI, is clearly not conducive to regulation of weapon autonomy. With presidents Trump, Xi and Putin in power, a breakthrough is not to be expected any time soon. But political will for regulation can also be generated from the ground up.

Growing political will from the grassroots

Surveys consistently show publics from all over the world rejecting LAWS. Opposition globally increased from 56% in 2016 to 61% in 2018, according to

92 Peter Asaro, "*Jus Nascendi*, Robotic Weapons, and the Martens Clause", in Ryan Calo, A. Michael Froomkin and Ian Kerr (eds), *Robot Law*, Edward Elgar, Cheltenham, 2016, p. 385.

93 F. Sauer and N. Schörnig, above note 41; Sarah E. Kreps, "Just Put It on Our Tab: War Financing and the Decline of Democracy", *War on the Rocks*, 28 May 2018, available at: https://warontherocks.com/2018/05/just-put-it-on-our-tab-21st-century-war-financing-and-the-decline-of-democracy/.

94 Denise Garcia, "Killer Robots: Toward the Loss of Humanity", *Ethics and International Affairs*, April 2015, available at: www.ethicsandinternationalaffairs.org/2015/killer-robots-toward-the-loss-of-humanity/.

95 DoD, above note 16.

96 E. B. Kania, *Battlefield Singularity*, above note 46.

KRC survey data.[97] This conforms with earlier online polling conducted by the Open Roboethics Initiative[98] as well as Heather Roff via IPSOS.[99] Opposition in the United States, China and Russia is at 52%, 59% and 60% respectively.[100] In Europe, the numbers range from 60% in Finland up to 81% in Ireland.[101]

Survey data also suggest that the public's opposition is primarily fuelled not by legal concerns or worries about unwanted escalation or crisis instability but by the notion that delegating decisions over life and death on the battlefield crosses a bright red moral line.[102] So while there is certainly an interesting philosophical debate to be had about the cultural pervasiveness of human dignity as a concept and its relevance to the LAWS issue from utilitarian versus deontological viewpoints, the concern as presented in the preceding section quite clearly resonates with the general public. The notion that there is something fundamentally wrong with having humans killed by mindless machines is thus well suited to creating grassroots pressure on governments in order to muster more political will on the issue. This point is granted even by sceptics of the human dignity argument as a whole: "There could be some campaigning advantages. Saying that something is against human dignity evokes a strong visceral response."[103]

Fostering norm development in the CCW

LAWS keep steadily gathering media attention around the globe.[104] With mounting public pressure and increased scrutiny, there will be a strong incentive for CCW States Parties to produce tangible results for the 2021 Review Conference. The "aspects of the normative and operational framework" that are to be further developed over the course of 2021 could take a more concrete shape in three steps.

First, consensus seems achievable on shared language that adopts the by now widely accepted functionalist view of weapon autonomy as well as a common understanding that some form of positive obligation and affirmation of the principle of human control over weapons systems is required.[105] The CCW's guiding principle (b) already points this way in stating that "[h]uman

97 KRC, "Global Poll Shows 61% Oppose Killer Robots", 22 January 2019, available at: www.stopkillerrobots.org/2019/01/global-poll-61-oppose-killer-robots/.

98 Open Roboethics Institute, "The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll", 9 November 2015, available at: www.openroboethics.org/wp-content/uploads/2015/11/ORi_LAWS2015.pdf.

99 Heather M. Roff, "What Do People Around the World Think about Killer Robots?", *Slate*, 8 February 2017, available at: https://slate.com/technology/2017/02/what-do-people-around-the-world-think-about-killer-robots.html.

100 KRC, above note 97.

101 KRC, "New European Poll Shows Public Favour Banning Killer Robots", 13 November 2019, available at: www.stopkillerrobots.org/2019/11/new-european-poll-shows-73-favour-banning-killer-robots/.

102 KRC, above note 97.

103 A. Sharkey, above note 85, p. 9.

104 R. Charli Carpenter, *"Lost" Causes: Agenda Vetting in Global Issue Networks and the Shaping of Human Security*, Cornell University Press, Ithaca, NY, 2014, pp. 88–121.

105 Stephen D. Goose and Mary Wareham, "The Growing International Movement Against Killer Robots", *Harvard International Review*, Vol. 37, No. 4, 2016, available at: www.jstor.org/stable/26445614.

responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines".[106] The Forum for Supporting the 2020 GGE on LAWS conducted in April 2020 as a webcast by the German Federal Foreign Office, with 320 registered participants representing sixty-three CCW States Parties, underlined the importance of further conceptualizing the human element. Controllability of weapons is arguably a proto-norm already,[107] and a shared terminology – be it "meaningful human control" or some other formulation – could be found to stipulate in a general sense when humans and when machines are to be performing which function in the targeting cycle. The ICRC and the Stockholm International Peace Research Institute (SIPRI) recently presented a conceptual framework that can support this effort of operationalizing human control – that is, of clarifying the "who, what, when and how" of controlling weapons and limiting their autonomy.[108]

Second, since there is no one-size-fits-all standard of meaningful human control, the sharing of best practices and, more importantly, of case studies of specific weapons systems and operational scenarios could allow CCW States Parties to develop a deeper, shared conceptual grasp of the intricacies involved with implementing human control in design and use. The GGE is uniquely suited to facilitate these sorts of deep dives with analyses from multiple stakeholders and a sharing of legal, ethical and operational views. Smaller expert groups such as the International Panel on the Regulation of Autonomous Weapons (iPRAW) and the commission on the responsible use of technologies in the Franco-German Future Combat Air System are already beginning to organize their research toward that end.

Third, a differentiated implementation scheme could be developed that conceives of human control as being exercised in a context-dependent way – that is, contingent on the weapons system, its mission environment, "target profiles"[109] and additional factors such as mission duration.[110] This human control scheme could prescribe minimum standards for controllability by design, for example regarding the ergonomics of human–machine interfaces, and determine "levels of human supervisory control"[111] in use – that is, the tactics, techniques and procedures required to keep human control and responsibility intact during the system's operation.

106 CCW Meeting Final Report, above note 2, p. 10.

107 For the notion of codifying human control as a principle of IHL in general, see Elvira Rosert, *How to Regulate Autonomous Weapons*, PRIF Spotlight 6/2017, Peace Research Institute Frankfurt, 2017, available at: www.hsfk.de/fileadmin/HSFK/hsfk_publikationen/Spotlight0617.pdf.

108 V. Boulanin *et al.*, above note 26. See also Ilse Verdiesen, Filippo Santoni de Sio and Virginia Dignum, "Accountability and Control over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight", *Minds and Machines*, 2020, available at: https://link.springer.com/article/10.1007/s11023-020-09532-9.

109 Moyes, "Target Profiles", above note 26.

110 For this general approach as well as a list of variables to consider, see V. Boulanin *et al.*, above note 26, pp. 30–33.

111 F. Sauer *et al.*, above note 79, pp. 42–45.

It currently seems unlikely that the CCW process, even if it were to complete these three steps, will end up yielding more than "soft law", such as a consensual political declaration or a catalogue of best practices. In fact, a complete breakdown of the CCW process in Geneva is also within the realm of possibility. But even if the CCW turns out not to be the venue from which a legally binding regulation for weapon autonomy emerges, it has already served as an information hub and norm incubator for the last six years – and will continue to do. Especially considering the effect of the COVID-19 crisis on meeting schedules around the globe, it is currently too early to tell if other fora – and if so, which ones – can and should pick up the ball on regulation where the CCW leaves it in 2021, in order to further develop and codify the human control norm as binding international law.

Conclusion

A multilateral regulation of autonomy in weapons systems – that is, codifying a legally binding obligation to retain meaningful human control over the use of force – is difficult yet imperative to achieve. Severe strategic as well as ethical mid- and long-term risks, such as unintended conflict escalation at machine speed and the violation of human dignity, outweigh any short-term military benefits. This analysis has illustrated how regulating weapon autonomy is feasible, presenting a three-step process to facilitate stepping back from the brink: step one, foster the emerging consensus on the notion that a positive obligation to retain human control over weapons systems is prudent and urgently required; step two, further develop the insight that there is no one-size-fits-all standard of meaningful human control; and step three, devise differentiated, context-dependent human control schemes for weapons systems. Given the current geopolitical landscape and the lack of political will to engage in arms control efforts, the taking of these steps will resemble a marathon, not a sprint. After all, the perceived military value of weapon autonomy is exceptionally high, and the issue itself is elusive, requiring an innovative, qualitative approach to arms control.

But history clearly suggests that great powers are not devoid of sensitivity to the accumulation of collective risks – otherwise arms control on nuclear, chemical and biological weapons would never have seen the light of day. The emerging technologies of the twenty-first century present humankind with the opportunity to demonstrate that it has learned from history *before* the risks have manifested themselves to their full extent. Humans do terrible things to each other in war, and there is no technological fix for that. But the international community can at least set rules to curb against uncontrolled escalation and the crossing of fundamental moral lines. If we fail to do so, we will not only lose

the breathing room to ponder and deliberate responses,[112] an essential requirement of political conflict management, as the Cuban Missile Crisis strongly suggests;[113] we will also allow "the ultimate indignity" of war turning into "death by algorithm".[114]

112 Z. Davis, above note 35, p. 122.
113 Frank Sauer, *Atomic Anxiety: Deterrence, Taboo and the Non-Use of U.S. Nuclear Weapons*, Palgrave Macmillan, London, 2015, pp. 91–92.
114 Robert H. Latiff and Patrick J. McCloskey, "With Drone Warfare, America Approaches the Robo-Rubicon", *Wall Street Journal*, 14 March 2013, available at: https://tinyurl.com/y2t7odsh.

International Review of the Red Cross (2020), **102** (913), 261–285.
Digital technologies and war
doi:10.1017/S1816383121000059

IRRC

The changing role of multilateral forums in regulating armed conflict in the digital age

Amandeep S. Gill
Ambassador Amandeep S. Gill is Director of the International Digital Health and AI Research Collaborative at the Global Health Centre of the Graduate Institute for International and Development Studies in Geneva. He was the Executive Director and Co-Lead of the Secretariat of the UN Secretary-General's High-Level Panel on Digital Cooperation until August 2019 and has served as India's Ambassador and Permanent Representative to the Conference on Disarmament in Geneva. Ambassador Gill chaired the Group of Governmental Experts of the Convention on Certain Conventional Weapons on emerging technologies in the area of lethal autonomous weapon systems from 2017 to 2018. He currently serves as a member of the UNESCO Ad Hoc Expert Group drafting a recommendation on the ethics of artificial intelligence, and as a Commissioner on the *Lancet/Financial Times* Commission on "Governing Health Futures 2030: Growing Up in a Digital World".

Abstract

This article examines a subset of multilateral forums dealing with security problems posed by digital technologies, such as cyber warfare, cyber crime and lethal autonomous weapons systems (LAWS).[1] It identifies structural issues that make it difficult for multilateral forums to discuss fast-moving digital issues and respond in time with the required norms and policy measures. Based on this problem analysis, and the recent experience of regulating cyber conflict and LAWS through Groups of Governmental Experts, the article proposes a schema for multilateral governance of

digital technologies in armed conflict. The schema includes a heuristic for understanding human–machine interaction in order to operationalize accountability with international humanitarian law principles and international law applicable to armed conflict in the digital age. The article concludes with specific suggestions for advancing work in multilateral forums dealing with cyber weapons and lethal autonomy.

Keywords: digital technology, conflict, cyber security, autonomous weapons, human–machine interface, distributed governance, multilateral forums, international humanitarian law, accountability.

Introduction

Global security and stability are increasingly intertwined with digital technologies, and even older-generation cyber capabilities are "becoming more targeted, more impactful on physical systems, and more insidious at undermining societal trust".[2] The challenge has grown beyond the kinetic impact of cyber attacks on critical infrastructure – ports, air traffic control, power grids and financial flows – to the "hacking" of public opinion and political institutions. Efforts to develop international norms for responsible behaviour, build capacity and enhance trust remain fragmented, particularly at the inter-governmental level. Norms that are championed by one side or the other often lack comprehensiveness and a critical mass of support from key governments. Frustrated by a lack of international cooperation, many jurisdictions such as the United States and European Union are increasingly imposing unilateral sanctions on specific individuals and entities in response to cyber attacks.[3] Adding to the mistrust are export control, competition policy and investment policy measures targeted at digital companies of peers.[4]

The advent of artificial intelligence (AI) brings another urgent dimension to the regulation of armed conflict in the digital age. The convergence of three trends – namely, increased computing power, big data sets leveraged through the Internet, and low costs of data storage and manipulation – has mainstreamed AI techniques such as machine learning. AI systems can handle tasks that previously

1 In the author's understanding, digital technologies are devices, platforms, data storage and processing architectures, algorithms, computing languages, communication protocols and standards that rely on the representation of information as discrete binary values. Information and communications technology (ICT) is another term that is interchangeably used in this regard.

2 High-Level Panel on Digital Cooperation, *The Age of Digital Interdependence: Report of the UN Secretary-General's High-Level Panel on Digital Cooperation*, June 2019, p. 27.

3 Patryk Pawlak and Thomas Biersteker (eds), *Guardian of the Galaxy: EU Cyber Sanctions and Norms in Cyberspace*, Chaillot Paper 155, EU Institute for Security Studies, October 2019.

4 Ana Swanson, "U.S. Delivers Another Blow to Huawei with New Tech Restrictions", *New York Times*, 15 May 2020; Li Sikun, "China Ready to Target Apple, Qualcomm, Cisco and Boeing in Retaliation against US' Huawei Ban: Source", *Global Times*, 15 May 2020.

seemed reserved for human minds. The March 2016 defeat of Lee Sedol, the eighteen-time world champion in the board game go, by Deep Mind's AI algorithm is a powerful example.[5] In particular, the use of machine learning-based AI systems in armed conflict raises concerns about the loss of human control and supervision over the conduct of war.[6] This could result in harm to civilians and combatants in armed conflict in contravention of international humanitarian law (IHL); it could also bring about a new arms race and lower the threshold for the use of force.

Are multilateral forums fit for purpose for the regulation of armed conflict in the digital age? To answer this question, the following analysis first recalls the historical context of modern multilateral forums dealing with conflict prevention and arms control. It thereafter juxtaposes the established procedures and outcomes of such forums against the nature of digital technologies and the unique characteristics of their use in armed conflict. This is followed by a survey of select forums dealing with the consequences of such use, in particular the Group of Governmental Experts (GGE) on Developments in the Field of Information and Telecommunications in the Context of International Security established by the United Nations General Assembly from time to time since 1998 (UN GGE), and the GGE comprised of all High Contracting Parties to the Convention on Certain Conventional Weapons dealing with lethal autonomous weapons systems (LAWS) since 2016. A framework for understanding human accountability under national and international law in the various stages of the development and use of autonomous digital technologies is abstracted from this survey. The analysis concludes with a mapping of some future directions for multilateral forums dealing with cyber conflict and autonomous weapons, and proposes specific steps for multilateral governance of digital technologies in the international security context by co-opting additional governance actors and taking a tiered approach to the application of norms.

Historical background and context

Historically, the idea of international conferences and forums to negotiate the prevention and regulation of conflict *ante factum* in peacetime is relatively new. We can date it back to the post-war conferences in Geneva of 1863–64 and 1868, which codified the rules of war on land and at sea;[7] the St Petersburg Declaration of 1868, which prohibited the use in conflict of a specific category of

5 Amandeep S. Gill, "The Role of the United Nations in Addressing Emerging Technologies in the Area of Lethal Autonomous Weapons Systems", *UN Chronicle*, Vol. 55, No. 3–4, 2019.

6 The UN Secretary-General has called for a ban on machines with the power and discretion to take lives without human involvement. António Guterres, "Remarks at the 'Web Summit'", Lisbon, 5 November 2018, available at: www.un.org/sg/en/content/sg/speeches/2018-11-05/remarks-web-summit (all internet references were accessed in January 2021).

7 The former resulted in the establishment of the International Committee of the Red Cross (ICRC) and the adoption of the Convention for the Amelioration of the Condition of the Wounded in Armies in the Field of 22 August 1864. The latter helped adapt the principles of the Geneva Convention of 1864 to sea warfare.

munitions;[8] and the series of Peace Conferences held at The Hague in 1899 and 1907. The latter were the "first truly international assemblies meeting in time of peace for the purpose of preserving peace".[9] Their legacy is still with us, for example, in the form of the Permanent Court of Arbitration, which provides a forum for arbitration, enquiry and conciliation among States with regard to the agreements they have entered into, and the famous Martens Clause, which provides a touchstone beyond legal norms for humanitarian protection during armed conflict.[10]

Subsequent forums such as the UN General Assembly and the Military Staff Committee, a subsidiary body of the UN Security Council, reflect the relatively modest provisions of the UN Charter in the field of conflict prevention, disarmament and arms control in comparison, for example, to the Statute of the League of Nations, which was more forward-leaning on collective action for disarmament and world peace.[11] The UN Charter was also pre-atomic, as the provisions of the Charter had been negotiated before the knowledge of atomic weapons became widely known. The development of nuclear weapons and the Cold War arms race further turned attention away from conventional-weapons-related arms control, which tended to play second fiddle to strategic weaponry during the Cold War years and was often seen through a regional rather than a global lens. This gap between the treatment of weapons of mass destruction and conventional weapons persisted even though technology and security trends began to shift in the late 1990s, raising the importance of the latter. The intangible attributes of weapons systems, in particular the growing digitalization of the key components of these systems, also stayed largely unappreciated outside a restricted circle of practitioners of export control regimes such as the Wassenaar Arrangement. The policy-makers' lag with regard to the rapid pace of technology developments in the post-Cold War period is evident also from the shifting use of terms such as "cyber security", "network security", "internet security" and "digital security" – they were unsure of what they were dealing with and tended to echo the most fashionable term of the day.

The digital challenge to multilateral forums

It matters that the "digital turn" was not as dramatic as its nuclear predecessor. It took a while for the practitioners of conflict regulation and arms control to realize that they were dealing with a fundamentally new technology. The internet

8 Declaration Renouncing the Use, in Time of War, of Explosive Projectiles under 400 Grammes Weight, St Petersburg, 29 November and 11 December 1868.

9 James Brown Scott, "Prefatory Note", *The Proceedings of the Hague Peace Conferences: Translation of the Official Texts: The Conference of 1899*, Oxford University Press, Oxford, 1920, p. v.

10 Theodor Meron, "The Martens Clause, Principles of Humanity, and Dictates of Public Conscience", *American Journal of International Law*, Vol. 94, No. 1, 2000.

11 Leland M. Goodrich, Edward Hambro and Anne Patricia Simons, *Charter of the United Nations: Commentary and Documents*, 3rd revised ed., Columbia University Press, New York, 1969.

protocols had been known to the esoteric few since the 1970s, but the switch to the common Transmission Control Protocol/Internet Protocol (TCP/IP) for all internet hosts in 1984 and the invention of the World Wide Web at CERN in Geneva in 1989 led rapidly to the global adoption of digital communications networks. A vast array of applications began to be deployed on the virtual infrastructure created by this distributed and scalable plumbing. The marriage of internet and telephony at the beginning of the twenty-first century liberated users from the yoke of a desk, and social media platforms allowed them to create content at an unprecedented scale. That included, of course, spam, incitement to hate and violence, and malware. Digital networks permitted an unprecedented level of anonymity in the generation and use of content, the Internet's global reach permitted rapid virality of such content, and the high stakes involved in access to and use of the Internet created the incentives for malicious use at scale.

Why is this important in the context of IHL and arms control? Force has traditionally been correlated with physical destruction, injury and death, and few had imagined that non-physical destruction or the physical consequences of non-physical attacks could be so significant as to become issues of international security. Further, if misinformation and propaganda had needed to be regulated, it had been in the narrow context of the ruse in combat or the misuse of symbols of neutrality and protection. No one had imagined that disinformation and distortion of ideas and institutions through digital technology could reach such scale. Other certainties – combatant versus non-combatant, civilian versus military target – were also shaken, if not compromised. Finally, even though irregulars have been used throughout the history of IHL and arms control, practitioners had always assumed that attribution for the use of force was a tractable problem. How to assign accountability when there are so many fingers on the digital trigger, many unwittingly so, and when arguably there might be no trigger at all?

Having examined the conceptual challenges, let us look at the challenges around practice. Traditionally, multilateral forums have regulated conflict by helping to negotiate norms that regulate, restrict or rule out specific means and methods of warfare, as well as by promoting dialogue and confidence-building among potential belligerents. One category of forums – though admittedly small – has allowed States to clarify ambiguous situations, detect cheating and mediate differences before they snowball into conflict.[12] The toolkit that such forums have deployed in architecting norms include measures for information exchange and transparency, declaration of facilities and stocks, measures for verifying the absence of certain types of activities and items, and compliance measures for punishing violations.[13] In technology areas such as space, chemicals, biology and nuclear science, where both civilian and military uses exist, multilateral regimes

12 An example is the International Atomic Energy Agency (IAEA) investigation of ambiguities related to Iran's uranium enrichment centrifuge programme. IAEA, "Verification and Monitoring in Iran", available at: www.iaea.org/newscenter/focus/iran.

13 The Open Skies Treaty of 1992, for example, uses aerial inspections to verify deployment of forces in order to rule out surprise attacks. Arms Control Association, "The Open Skies Treaty at a Glance", fact sheet, available at: www.armscontrol.org/factsheets/openskies.

such as the Missile Technology Control Regime regulate transfers through lists of sensitive items and guidelines for exports to prevent diversion of civilian technologies for military use.[14]

However, digital technologies pose unique challenges for the practitioners of such forums.[15] The question of what exactly constitutes a tool of conflict does not have an obvious answer, and applying the traditional lens of dual-use technologies – space launch vehicles versus missiles, nuclear power reactors versus fissile material production reactors – does not result in an actionable insight. Unlike a battle tank or a ballistic missile, no digital tool can be deemed a weapon independent of its context.[16] Such tools can be endlessly replicated, repurposed and combined with physical systems. There is no discrete set of "processes" such as the two routes of enrichment and reprocessing for producing fissile material for nuclear weapons, and there are no easily identifiable "choke points", such as separation of plutonium from spent fuel, which can be policed.[17] In terms of effect, it is often hard to establish the threshold at which something jumps from the virtual to the physical, from surveillance and harassment to inter-State conflict, and from the local and the national into the international domain. Unlike physical weapons, cyber weapons do not destroy themselves and can be reused to attack the attacker.[18] The actors, too, are hard to separate into discrete sets of State and non-State actors, as States seldom claim official responsibility for cyber attacks. Furthermore, the category of State actors is itself fluid – there is no privileged and static group of possessors behind a high entry threshold.[19] Notions of parity and balance so dear to Cold War practitioners of arms control are also hard to define in matters digital.[20]

At a more basic level, the speed and spread of development of digital technologies overwhelms the ability of policy forums to keep pace with the social, economic and political consequences of technological change.[21] This is further complicated by the bigger role (compared to the State sector) that the private

14 See the "MTCR Guidelines" and "MTCR Annex" sections of the Missile Technology Control Regime website, available at: https://mtcr.info.

15 Colin Picker, "A View from 40,000 Feet: International Law and the Invisible Hand of Technology", *Cardozo Law Review*, Vol. 23, 2001, p. 149.

16 Take malware, for example: law enforcement agencies can use the same lines of code to monitor and thwart the planning of terrorist acts.

17 Funding of research and development (R&D), certain types of datasets or computing capacity could theoretically be considered choke points for the development of lethal autonomous weapons, but practical ways to prevent their use for military purposes are impossible to envision.

18 An example is the NotPetya malware, which used a penetration capability, named EternalBlue, allegedly developed by the US National Security Agency but leaked in early 2017. David Bisson, "NotPetya: Timeline of a Ransomworm", *Tripwire*, 28 June 2017.

19 Unlike the Treaty on the Non-Proliferation of Nuclear Weapons, for example, which designates five States that manufactured and exploded a nuclear device prior to 1967 as nuclear weapon States (Art. IX, para. 3).

20 In the case of nuclear weapons and their delivery systems, elaborate models of parity, stability and balance can be built because their impact is knowable and their capabilities can be estimated, counted and even enshrined in treaty systems such as the Soviet–US strategic treaties. However, cyber capabilities are highly esoteric. Their impact independent of context is hard to estimate, let alone compare.

21 The so-called "Moore's law", which states that density of transistors on an electronic chip will double every year even as costs go down, is an illustration of this frenetic pace.

sector now plays in the development of cutting-edge technology, business applications and the creation of intellectual property.[22] Table 1 summarizes some of the characteristics of traditional multilateral forums which make it challenging to address the policy implications of digital technologies.

Table 1. *Attributes of multilateral forums which pose challenges with regard to digital issues*

Attribute	Challenge
Periodicity, response time	Multilateral forums meet at regular intervals, often annually, even though working groups and preparatory committees may meet inter-sessionally. Meetings last for a short period of time, ranging from less than a week to a few weeks. Treaty negotiation takes many years and treaty review conferences often happen every five years. At the digital end, one year is a long time in technology development and adoption, and irregular meetings for a short duration are insufficient to study impact and plan policy responses.[23]
Agenda-setting	Multilateral forums have structured agendas and inter-sessional programmes with negotiated mandates. Technology issues are often tucked under broad items such as "science and technology developments" or the "impact of emerging technologies". Digital technologies are hard to fit into static agendas and mandates; as the Cambridge Analytica case highlights, they require specificity with regard to the context of their use for governance issues to clearly emerge.[24]

Continued

22 The 2019 *Digital Economy Report* by the UN Conference on Trade and Development (UNCTAD) has an excellent analysis of the policy challenges posed to traditional forums on competition policy and trade rules by the explosive growth of the digital economy. UNCTAD, *Digital Economy Report 2019*, Geneva, 2019, available at: https://unctad.org/en/PublicationsLibrary/der2019_en.pdf.

23 Mark Zuckerberg's testimony before the US Congress is an example of the lag not only between policy-making and digital technology's impact but also between policy-makers' understanding of complex and fast-moving technology-based business models and those who create and run such businesses. Casey Newton, "The 5 Biggest Takeaways from Mark Zuckerberg's Appearance before the Senate", *The Verge*, 10 April 2018. available at: www.theverge.com/2018/4/10/17222444/mark-zuckerberg-senate-hearing-highlights-cambridge-analytica.

24 *Ibid.*

Table 1. *Continued*

Attribute	Challenge
State-centrism	Multilateral forums are State-centric, and State participation is usually limited to diplomatic representatives. Inter-ministerial coordination prior to and inter-agency participation in multilateral meetings is often a luxury for most delegations.[25] Digital technologies are developed largely in the private sector, and this has implications for both understanding and control of technology by governments. Their regulation also generally falls under several ministries and departments, making coordinated action difficult.[26]
Tangibility	Multilateral forums dealing with arms control and conflict focus on tangible weapons and tools, specific production pathways, and clearly delineated industry domains and business models. The impact of these artefacts is also seen in tangible terms – destruction of property, loss of life, damage to the environment etc. Digital technologies are intangible; they may not be part of weapons but can still influence conflict significantly (e.g. through decision support systems) and cannot be isolated from other technology domains. Unlike other weapons, the impact of digital technologies can be mostly (though not necessarily exclusively) socio-psychological, without an underlying physically destructive threat. This attribute can disorient practitioners who are used to counting, comparing and controlling discrete weapons and platforms.
Outcomes	The output of these forums is either legally binding treaties, "hard law" that can be implemented by governments through domestic legislation and/or regulations, or reports and resolutions that have a political impact in the context of inter-State relations. Treaties are policed by verification regimes with inspections, declarations and challenge inspections.

25 "Small Developing Countries Struggle in WTO", *Forbes,* 19 May 2010.
26 For example, unlike traditional banking, which is regulated through ministries of finance, digital payments cut across finance, communications and ICT ministries.

	Political commitments are upheld through dialogue, diplomacy, reciprocity and reputational consequences. Hard law applies to only a narrow aspect of digital technologies, chiefly business-related regulation, consumer and worker protections etc. Mostly, however, digital technologies are governed through "soft law", including "voluntary programs, standards, codes of conduct, best practices, certification programs, guidelines, and statements of principles".[27] Intrusive verification is not standard practice for enforcing these norms.
Inter-disciplinarity	Multilateral forums, by design and in their functioning, tend towards specialized treatment of subjects (trade, disarmament, human rights, environment etc.) independent of technology. While this is changing,[28] most participants in these forums, and the secretariats supporting them, also tend to come from a specialized field. The lack of cross-disciplinary approaches is particularly striking in arms control and disarmament forums, where practitioners normally do not interact with technologists and entrepreneurs.[29] For their part, technology developers and entrepreneurs often lack appreciation of the political and security impact of their innovations. Opportunities for engagement with policy-makers are limited to market regulatory contexts, although in recent years multi-stakeholder forums such as the World Economic Forum have expanded the circle.
Power distribution	Multilateral forums operate with traditional power dynamics – major powers often act as "norm

Continued

27 Wendell Wallach and Gary Marchant, "Toward the Agile and Comprehensive International Governance of AI and Robotics", *Proceedings of the IEEE*, Vol. 107, No. 3, 2019.

28 Recent initiatives on internet governance and digital cooperation have used digital technologies and cross-disciplinarity of impact as their organizing principles, rather than a specific UN domain such as human rights.

29 In the author's experience of negotiations at the Conference on Disarmament or discussions at the UN Disarmament Commission from 2010 to 2017, there was not a single instance of interaction with industry; this is unlike the Nuclear Security Summit process, which was organized outside the UN context. In the Convention on Certain Conventional Weapons (CCW) context, industry interaction was brought in through side events in 2017–18.

Table 1. *Continued*

Attribute	Challenge
	entrepreneurs",[30] while groupings of States act as pressure groups. While private companies and civil society have had an important agenda-setting and opinion-shaping role in some discussions,[31] they take a secondary position to more powerful State and inter-State actors. This power asymmetry sits uneasily with the digital technology reality. For example, digital platforms such as Facebook, Alipay and WhatsApp may have more users ("virtual residents") than the populations of most countries; they operate quasi-global infrastructures, act as cross-border "content policemen" and have market capitalizations that dwarf other sectors and most national GDPs. If norms related to digital technologies are to have an impact, the digital industry has to be a part of the discussion on policy responses and has to cooperate with State actors for their implementation.

Against the background of these challenges, it is useful to look at select multilateral forums at the interface of digital technologies and international security.

Select multilateral forums dealing with international security implications of digital technologies: The UN GGEs and the OEWG on information security

Upon an initiative of the Russian Federation, the item "Developments in the field of information and telecommunications in the context of international security" was put on the agenda of the UN General Assembly in 1998. Resolutions adopted by the First Committee of the General Assembly under this item have over the years created five GGEs to examine the issue and make recommendations. A sixth one is currently under way under the chairmanship of Brazil's Guilherme Patriota, alongside an Open-Ended Working Group (OEWG) on the same subject under

30 Martha Finnemore and Kathryn Sikkink, "International Norm Dynamics and Political Change", *International Organization*, Vol. 52, No. 4, 1998.

31 Two examples are the 1997 Anti-Personnel Landmines Convention, or Ottawa Treaty, and the 2017 Treaty on the Prohibition of Nuclear Weapons. Civil society's role in achieving these treaties was acknowledged through the award of the Nobel Peace Prize to Jody Williams and the International Campaign to Ban Landmines in 1997, and to the International Campaign to Abolish Nuclear Weapons in 2017.

the chairmanship of the Swiss ambassador Jürg Lauber.[32] The limited-membership GGEs, with fifteen to twenty-five government-appointed experts, hold their sessions alternately in Geneva and New York, while the OEWG is open to all UN member States and observers and meets in New York. The value of the UN GGE and OEWG processes is in sustaining dialogue to develop common understandings on the security implications of the use of digital technologies as well as in promoting norms, rules and principles for responsible State behaviour in this area. Opportunities for industry, academia and civil society to engage with these forums, albeit in a limited fashion (such as through written inputs or short statements during plenary debates), have been provided for as part of the resolutions setting them up.[33]

The report of the 2013 UN GGE affirmed that international law, and in particular the UN Charter, is applicable and essential to maintaining peace and stability and promoting an open, secure, peaceful and accessible information and communications technologies (ICT) environment. The 2015 GGE further proposed a set of eleven voluntary and non-binding norms for States aimed at promoting an open, secure, stable, accessible and peaceful ICT environment, while noting a proposal for a Code of Conduct on Information Security.[34] These norms include an obligation (voluntary) not to conduct or knowingly support ICT activity contrary to a State's obligations under international law that intentionally damages critical infrastructure or otherwise impairs the use and operation of critical infrastructure to provide services to the people.[35] They also include an injunction not to target the computer emergency response services of other States. This recalls the notion of protected objects/functions and the principle of distinction in IHL. Unfortunately, consensus broke down in the 2017 GGE, chiefly on the issue of the applicability of international law to cyber conflict.[36]

The challenge today is not only repairing that breach in consensus but also reconciling the outputs of two forums moving in parallel pathways on the same issue: the sixth UN GGE and the OEWG, which were set up in December 2018 through resolutions championed by the United States and Russia respectively. The fault line is not only procedural but also substantive: it lies between approaches that stress the applicability to cyber conflict of existing norms and accountability thereunder, on the one hand, and approaches that reject an

32 The new GGE was set up upon an initiative of the US, while Russia, a previous votary of the GGE idea, switched in 2018 to sponsoring the more participatory methodology of an OEWG open to all member States while still participating in the limited-membership GGE set up under US sponsorship.

33 Details of multi-stakeholder engagement and inputs with regard to the OEWG are available on the OEWG website, available at: www.un.org/disarmament/open-ended-working-group/.

34 UN GGE, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, UN Doc. A/70/174, 22 July 2015, paras 12, 13.

35 Recent multi-stakeholder processes such as the Global Commission on the Stability of Cyberspace (GCSC) have suggested additional norms and protections, including for electoral infrastructure. GCSC, *Advancing Cyberstability: Final Report*, November 2019.

36 Elaine Korzak, "UN GGE on Cybersecurity: The End of an Era?", *The Diplomat*, 31 July 2017, available at: https://thediplomat.com/2017/07/un-gge-on-cybersecurity-have-china-and-russia-just-made-cyberspace-less-safe/.

automatic extension of existing international law, in particular IHL, to cyber conflict and stress instead the need to negotiate new norms, on the other.

It is instructive in this regard to juxtapose the comments of the Russian and US delegates on the pre-draft of the OEWG's report circulated by the chair on 11 March 2020.[37] The US delegate, in comments largely appreciative of the chair's pre-draft but critical of certain inputs, stated:

> We appreciate that the draft report memorializes that all States reaffirmed that international law and, in particular, the Charter of the United Nations, apply to the use of ICTs by States. … The present draft devotes far too much attention (paragraphs 27–30) to proposals made by a minority of States for the progressive development of international law, including through the development of a legally binding instrument on the use of ICTs by States. These proposals lacked specificity and are impractical. The OEWG's mandate is to study how international law applies to the use of ICTs by States, and the report should therefore focus on existing international law. Without a clear understanding of States' views on how existing international law applies to ICTs, it is premature to suggest that international law needs to be changed or developed further.[38]

The Russian delegate, while noting some "positive traits", referred to "many unacceptable approaches" and stated:

> The document exaggeratedly emphasizes the principle of applicability of universally recognized norms and principles of international law set forth in the UN Charter and in the Declaration on principles of international law concerning friendly relations and cooperation among States in accordance with the Charter of the United Nations, dated 1970, to the use of ICTs. At the same time, this principle is not linked to specific modalities of its applicability, namely, who, how and in which circumstances can apply it. These practical aspects demand to be regulated by a specialized international legal instrument that would provide for the modalities of applicability of the existing norms of international law to the use of ICTs, as well as, if necessary, includ[ing] new norms. … We regard as potentially dangerous the attempts to impose the principle of full and automatic applicability of IHL to the ICT environment in peacetime. This statement itself is illogical and contradictory because IHL is applied only in the context of a military conflict while currently the ICTs do not fit the definition of a weapon.[39]

This is not the only arena in which these opposite views on existing law versus new norms play out. In the context of cyber crime, Russia, China and others have

37 OEWG, "Initial 'Pre-draft' of the Report of the OEWG on Developments in the Field of Information and Telecommunications in the Context of International Security", 11 March 2020, available at: https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/03/200311-Pre-Draft-OEWG-ICT.pdf.

38 "United States Comments on the Chair's Pre-draft of the Report of the UN Open Ended Working Group", available at: https://tinyurl.com/yyus5uv7.

39 "Commentary of the Russian Federation on the Initial 'Pre-draft' of the Final Report of the United Nations Open-Ended Working Group", available at: https://tinyurl.com/yxfuudvd.

proposed through the UN General Assembly the negotiation of a comprehensive international convention on countering the use of information and communications technologies for criminal purposes.[40] Many Western countries see this as potentially inimical to human rights and fundamental freedoms, as well as superfluous given the 2001 Budapest Convention on Cybercrime, which was negotiated within the Council of Europe framework; others see it as a crucial step going forward.[41]

The current geopolitical context of discussions on information security in New York makes it difficult to make progress, within the UN setting, on what constitutes a cyber weapon or cyber attack in the context of existing law on the use of force, how existing norms regulating the use of force would apply, and what gaps, if any, need to be addressed through additional norms. Progress is more feasible in smaller settings (or "minilaterals") on what norms apply to cyber conflict and cyber crime, and how those norms apply; attributing responsibility for cyber attacks; building mutual confidence and capacity; and promoting cooperation on enforcement of norms on cyber conflict.[42]

The importance of regional efforts is recognized in the resolution that set up the ongoing UN GGE in 2018, which requires the UN Office of Disarmament Affairs

> to collaborate with relevant regional organizations, such as the African Union, the European Union, the Organization of American States, the Organization for Security and Cooperation in Europe and the Regional Forum of the Association of Southeast Asian Nations, to convene a series of consultations to share views on the issues within the mandate of the group in advance of its sessions.[43]

The rival resolution under the same agenda item setting up the ongoing OEWG similarly calls upon the UN to encourage regional efforts, promote confidence-building and transparency measures, and support capacity-building and the dissemination of best practices.[44]

40 UNGA Res. 74/247, "Countering the Use of Information and Communications Technologies for Criminal Purposes", Draft Resolution, UN Doc. A/C.3/74/L.11, 11 October 2019.

41 "UN Approves Russian-Sponsored, China-Backed Bid on New Cybercrime Convention", *South China Morning Post*, 28 December 2019, available at: www.scmp.com/news/world/united-states-canada/article/3043763/un-approves-russian-sponsored-china-backed-bid-new.

42 For example, a process facilitated by the NATO Cooperative Cyber Defence Centre of Excellence resulted in a compilation of annotated international norms in 2013 known as the Tallinn Manual (updated in 2017 to the Tallinn Manual 2.0). The Tallinn Manual's drafting reflects the view that pre-cyber-era international law applies to cyber operations, both conducted by and directed against States, and that States both have rights and bear obligations under international law. See Michael N. Schmitt and Liis Vihul (eds), *Tallinn Manual 2.0 on International Law Applicable to Cyber Operations*, 2nd ed., Cambridge University Press, Cambridge, 2017. Another example is the proposal by some members of the Shanghai Cooperation Organization for a draft Code of Conduct on Information Security: see "Letter dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, Russian Federation, Tajikistan and Uzbekistan", UN Doc. A/69/723, 13 January 2015.

43 UNGA Res. 73/266, "Advancing Responsible State Behaviour in Cyberspace in the Context of International Security", Draft Resolution, UN Doc. A/C.1/73/L.37, 18 October 2018, op. para. 4.

44 UNGA Res. 73/27, "Developments in the Field of Information and Telecommunications in the Context of International Security", Draft Resolution, UN Doc. A/C.1/73/L.27/Rev.1, 29 October 2018, preambular para. 11.

What is less clear is to what extent UN processes can collaborate with the private sector, which is often the first responder and an involuntary accomplice in cyber attacks, and which is keen to clarify the ambiguity on applying norms in its own interest.[45] Engagement of the private sector could also be essential to avoiding unintended escalation because of a private sector-led response to cyber attacks from another jurisdiction.[46] However, UN forums, for good reasons, cannot treat private sector representatives on a par with member State representatives and often exclude private companies completely. Impatient with progress in multilateral forums, voluntary non-governmental initiatives such as the Paris Call and the Global Commission on the Stability of Cyberspace (GCSC) have taken steps towards crafting common denominators among multiple stakeholders on acceptable behaviour. Such initiatives can complement the efforts of multilateral intergovernmental forums.[47] In recent years, the UN Secretariat and UN agencies have also taken the initiative on involving the private sector more on issues related to technology. The International Telecommunications Union's (ITU) annual AI for Good Summit and the UN Secretary-General's High-level Panel on Digital Cooperation are two examples.

To sum up, the New York-centric efforts on cyber conflict have taken a more purposeful and welcome turn in the past couple of years despite the trying geopolitical circumstances. Challenges remain with regard to the engagement of non-traditional norm shapers, the absence of bottom-up possibilities for norm-making, and the need to depoliticize relatively less controversial aspects of cyber conflict such as assessment, assistance and cooperation.

In Geneva, the focus has traditionally been on expert-driven, in-depth work, aiming to result in legally binding norms on disarmament, arms control, human rights law and IHL. The city, which hosts the "single" multilateral disarmament negotiating forum, the Conference on Disarmament,[48] is also the home of the International Committee of the Red Cross (ICRC) and several other forums, including the World Intellectual Property Organization and the ITU, which plays an important role in the development of standards on digital technologies and capacity-building on cyber security. Geneva's perceived neutrality is an important consideration for multilateral and multi-stakeholder efforts on cyber security; it is therefore not surprising that the UN Institute for Disarmament Research (UNIDIR) holds its annual conference on stability of

45 Tech companies handle millions of attempts to breach cyber security measures on a daily basis. Their servers can be unwitting hosts for distributing malware, and the costs of breaches or non-payment of insurance claims due to ambiguity about the source of attacks can be crippling.

46 The French and Indian experts on the 2017 GGE made proposals to this effect. Source: personal communication with the author.

47 The Final Report of the GCSC proposes, for example, a set of eight norms additional to those proposed by the GGE in 2015, with the proposed norms applying both to State and non-State actors. GCSC, above note 35.

48 The UN General Assembly's First Special Session on Disarmament in 1978 created a "triad" of disarmament forums: the First Committee of the UN General Assembly, in New York; the UN Disarmament Commission, again in New York, as a universal deliberative body; and the Conference on Disarmament, in Geneva, as the "single" multilateral disarmament negotiating forum.

cyber space in Geneva and that the CyberPeace Institute, with its objectives of assisting victims of cyber attacks and establishing accountability for such attacks, has chosen Geneva as its host city.[49] The extensive ecosystem of humanitarian and human rights mechanisms as well as trade and development institutions in Geneva is an important asset for international cooperation on norms in the digital field, which does not respect traditional boundaries across the three UN pillars of trade and development, peace and security, and human rights and humanitarian affairs.

In a 2014 paper, Joseph Nye described in detail a "regime complex for managing global cyber activities", which left out what to many was an obscure forum in Geneva at the juncture of IHL and arms control.[50] This is the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (Convention on Certain Conventional Weapons, CCW). The CCW, negotiated under UN auspices in 1979–80, has its roots in key IHL principles such as proportionality and distinction between civilians and combatants. Currently, the Convention has five Protocols – Protocol I on Non-Detectable Fragments, Protocol II on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices (as amended on 3 May 1996), Protocol III on Prohibitions or Restrictions on the Use of Incendiary Weapons, Protocol IV on Blinding Laser Weapons, and Protocol V on Explosive Remnants of War, which deals with the problem of unexploded and abandoned munitions. The modular design of the Convention allows new instruments to be attached to the framework treaty as technology evolves and new humanitarian concerns emerge.[51]

It would have been hard to imagine in early 2014 that the CCW, which had hitherto mostly dealt with almost nineteenth-century-type weapons systems, would become a front-line forum dealing with the international security and international law implications of emerging digital technologies.[52] The focus, with incidents such as the Stuxnet attack on Iran's uranium enrichment gas centrifuges, had been on malware and cyber conflict. A series of breakthroughs in machine learning in the 2010s propelled another set of digital technologies loosely known as AI to the forefront, and the CCW became the forum for dealing with AI-based lethal autonomous weapons systems.

Regulating lethal autonomy in weapons systems: The CCW case

An important foundation for the CCW discussion was the 2013 report of the UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof

49 Information on the CyberPeace Institute is available at: https://cyberpeaceinstitute.org/.
50 Joseph S. Nye Jr., *The Regime Complex for Managing Global Cyber Activities*, Global Commission on Internet Governance Paper Series No. 1, May 2014.
51 For details of the CCW and its Protocols, see UN Geneva, "The Convention on Certain Conventional Weapons", available at: https://tinyurl.com/y4orq8q5.
52 A. S. Gill, above note 5.

Steyns. Special Rapporteurs are independent human rights experts with a considerable degree of autonomy in pursuing their mandates and in seeking evidence and inputs from governments, civil society, academia and industry.[53] Steyns was concerned by the arbitrariness involved in using drones to target non-State actors and how that challenge could be compounded by the use of autonomous technologies. He defined lethal autonomous robotics as "weapon systems that, once activated, can select and engage targets without further human intervention", and underlined concerns about the accountability of such systems and their compatibility with the requirements of IHL and the standards protecting life under international human rights law.[54] The report generated considerable debate but since the subject was felt to be more a question of arms control and the laws of armed conflict than one of human rights, key delegations in Geneva pushed for the consideration to be moved to the CCW in 2014.[55] A series of informal Meetings of Experts from 2014 to 2016 helped build consensus on a mandate in December 2016 for a formal GGE. The CCW GGE is the expert subsidiary body of the CCW, not just for examining issues like the GGEs set up by the First Committee in New York but also for negotiating new protocols when there is agreement to do so.[56] The CCW GGE, along with other GGEs established under the CCW, is thus different in nature and in terms of participation from the GGEs set up by the First Committee in that it is open to all High Contracting Parties to the Convention. In December 2016, a new GGE established under the CCW, the CCW GGE on LAWS, was mandated by the High Contracting Parties to examine "emerging technologies related to lethal autonomous weapons systems".[57]

There was now a forum with a mandate on the issue and with some attributes that mitigated the challenges previously listed. Its modular framework and past practice offered the possibility of graduating from a discussion to the negotiation of a binding instrument. Note that this is not the case with the GGEs established by the First Committee, although an OEWG can shift gears from discussion to negotiation with a fresh mandate.[58] Countries with emerging capabilities in AI systems such as Australia, Brazil, Canada, China, France, Germany, India, Israel, Japan, the Republic of Korea, Russia, South Africa, the United Kingdom and the United States are all party to the CCW. Further, the balance between humanitarian principles and military necessity inherent to

53 UN Office of the High Commissioner for Human Rights, "Special Procedures of the Human Rights Council", available at: www.ohchr.org/en/HRBodies/SP/Pages/Welcomepage.aspx.

54 Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, UN Doc. A/HRC/23/47, 9 April 2013.

55 Unsurprisingly, these included the main possessors and users of armed drones. Source: personal communication with the author.

56 See the reference to the negotiating work of the 2011 GGE on Cluster Munitions in UN Geneva, "GGE Sessions in 2011", available at: https://bit.ly/2ZozQMI.

57 Decision I of the Sixth Review Conference of the High Contracting Parties to the CCW, 12–16 December 2016.

58 This was the case most prominently with the Arms Trade Treaty adopted in 2013 after a seven-year process. Michael Spies, "Towards a Negotiating Mandate for an Arms Trade Treaty", *Disarmament Diplomacy*, No. 91, Summer 2009.

the Convention and the principal tenets of IHL provided the space for States with very differing views to begin engaging on a complex and rapidly evolving technology.

This is not to say that the choice of the forum or its continued use for addressing lethal autonomy is trouble-free. Its periodicity is annual, and the time allocated to discussions has varied from one to two weeks per annum.[59] Despite rules allowing the participation of academia, civil society and humanitarian organizations, the CCW is still State-centric. Therefore, direct engagement of AI technologists and industry in rule-making is not possible. Further, while anchoring the discussion in IHL comforts key stakeholders, it puts others who look at the issue essentially from a human rights perspective at unease.[60]

The challenge of interdisciplinarity has been mitigated to some extent with side events held by the ICRC, NGOs, academia and governments, which helped bring in fresh perspectives, including those of entrepreneurs.[61] Discussions in 2017 were focused by design on raising awareness of the technology, sifting through the hype and the dystopian fantasies, teasing out the cross-domain connections, bringing together legal, ethical, military, technical and entrepreneurial perspectives, and moving beyond binary mindsets of civilian-military objects and dual-use technologies.[62] The ethics panel was particularly useful in engaging the human rights and faith-based communities.[63] With regard to tangibility, testimonies and reports by independent experts, think tanks and organizations such as the ICRC, UNIDIR and SIPRI, as well as working papers submitted by the more active national delegations, helped build awareness of the nature of autonomous technologies. The ITU's AI for Good summits, which coincidentally started in 2017 across the road from the Palais des Nations, engaged many national delegations and helped cross-pollinate governance thinking around AI.[64]

59 In 2017, the CCW GGE on LAWS met for a week of five working days; this went up to ten working days over two weeks in 2018 and then came down to seven working days in 2019. The 2020 meetings are back to two weeks, while the time allotted in 2021 could go up to four weeks. The time for the meetings is negotiated as part of the annual mandate for the GGE and is subject to budgetary and political considerations. 2017 was curtailed because of arrears in payments by High Contracting Parties. See meeting recordings on the UN Digital Recordings Portal, available at: https://conf.unog.ch/digitalrecordings/; CCW, *Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects: Final Report*, UN Doc. CCW/MSP/2019/9, 13 December 2019.

60 This was the position of many delegations such as those of Sierra Leone, Costa Rica, Mexico, the Holy See and Honduras. For example, Ambassador Yvette Stevens of Sierra Leone argued at the CCW GGE on LAWS session of 13–17 November 2017 that the Human Rights Council should remain seized of the matter in parallel.

61 See, for example, the 2017 schedule of side events of the CCW GGE on LAWS.

62 CCW GGE, *Food-for-Thought Paper*, UN Doc. CCW/GGE.1/2017/WP.1, 4 September 2017. See also CCW GGE, *Provisional Programme of Work*, UN Doc. CCW/GGE.1/2017/2, 4 September 2017, with the integration of cross-disciplinary inputs into the discussions through four expert panels.

63 See meeting records on the UN Digital Recordings Portal, available at: https://conf.unog.ch/digitalrecordings/.

64 Se the AI for Good Global Summit Reports for the years 2017–19, available at: https://aiforgood.itu.int/reports/.

What about outcomes? In 2017, the CCW GGE on LAWS adopted a set of conclusions and recommendations by consensus.[65] One of these was that the CCW is the appropriate framework for dealing with the issue, while another was that IHL applies fully to the potential development and use of LAWS. This was an important early assurance, although it did not settle the question of whether further legal norms were needed. The consensus conclusions also highlighted three issues for future work: characterization of the systems under consideration – the so-called definitional issue; aspects of human–machine interaction, critical to the concern about potential violations of IHL; and possible options for addressing the humanitarian and international security consequences of such systems. Divergent views persisted on the definitions, risks and possible benefits of LAWS – as well as approaches to regulation and control, including the idea of a pre-emptive ban – but the chair's summary ended up becoming a practical device for capturing this diversity of views on future outcomes without blocking progress on substance.[66]

It is worth comparing the relative ease with which applicability of IHL was accepted in the LAWS context compared with the continued difficulty in the context of cyber weapons in the UN GGE and the OEWG. In the CCW, as was said in response to concerns expressed about the IHL reference by the Chinese delegate in 2017, the context was that of armed conflict and lethality – even if some delegations and the ICRC argued for a broader approach of "autonomous weapons systems".[67] Thus, in the context of the objectives and purposes of the CCW, IHL was clearly relevant regardless of views on whether further clarifications on the application of IHL principles to LAWS and/or new norms were needed or not.

A key concept in discussions on LAWS in 2017 was that of meaningful human control.[68] Conceptually attractive because it provided a way to avoid the negotiation of additional norms for ensuring compliance with IHL, it was for that reason also seen as problematic and even otherwise subject to different interpretations. At the April 2018 meeting of the CCW GGE on LAWS, it became possible to look at the broader notion of human involvement and intervention from the perspective of different parts of the technology development cycle. This allowed the GGE to move beyond the conceptual discussion on "meaningful human control" or similar concepts like "appropriate human judgement" and to look at the quality and depth of the human–machine interaction essential for meaningful compliance with IHL in each phase of technology development, deployment and use. Equally, it allowed governance

65 CCW GGE on LAWS, *Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2017/CRP.1, 20 November 2017.

66 A. S. Gill, above note 5.

67 See meeting records for 17 November 2017 on the UN Digital Recordings Portal, available at: https://conf.unog.ch/digitalrecordings/.

68 CCW GGE, *Examination of Various Dimensions of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, in the Context of the Objectives and Purposes of the Convention: Submitted by the Netherlands*, UN Doc. CCW/GGE.1/2017/WP.2, 9 October 2017.

choices at various levels for ensuring human responsibility and accountability to become clearer. The "sunrise slide" shown in Figure 1 captures this discussion and points the way to a distributed technology governance scheme that integrates multilateral norms with national regulatory approaches and industry self-regulation.[69]

We have previously seen how the absence of an agreed definition of cyber weapons is a barrier to progress on legal norms. The issue of defining LAWS could have been a showstopper in the CCW, but it was set aside for a step-by-step examination. At its April 2018 meeting, the CCW GGE on LAWS used a "LEGO exercise" to stack up different attributes that would be needed to characterize LAWS without picking specific attributes early or focusing only on technical characteristics. For one, an early definitional approach could have left out potential pathways to such weapons systems while the technology is still under development.[70] Thus, common ground was built up on the concepts and characteristics that would be required for an eventual definition without searching for the perfect dividing line between what is autonomous (future) and what is automated (present). Additionally, an understanding was reached that there needed to be a continued review of technology relevant to LAWS both for understanding what it is that delegations were dealing with but also its potential impact on regional and international security.[71]

An important substantive outcome reached in August 2018 was a set of "Possible Guiding Principles".[72] This negotiated outcome would not have been possible without the consensus reached the previous year on paragraph 16 of the CCW GGE on LAWS' November 2017 report, which noted the cardinal principle that IHL continues to apply fully to all weapons systems, including the potential development and use of LAWS.[73] That the Guiding Principles were not the end point but an essential foundation for further work was underlined in the consensus 2018 report, which included four options for a policy response: a legally binding instrument with prohibitions and regulations on LAWS; a political declaration underlining human control and accountability, with elements of transparency and technology review; identifying practical measures and best

69 At first glance the "sunrise slide" could be seen as excluding R&D as well as testing and evaluation from the purview of international regulation. Apart from the fact that the extent of international regulation is still to be determined – and in that sense the three regulatory regimes of industry standards, national regulations and international norms are "sliding doors" – this visualization does not exclude the penetration of international norms into domestic regimes on R&D, testing and evaluation, just as IHL weapons reviews have been internalized in domestic practice.

70 Potential carve-outs under some definitions for certain countries have also been an issue in disarmament and arms control negotiations before, and the CCW experience with the failed negotiations on a protocol on cluster munitions in 2011 must have weighed on the minds of some delegates.

71 CCW GGE on LAWS, *Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2017/CRP.3, 23 October 2018, paras 24–26.

72 *Ibid.*, para. 21.

73 CCW GGE on LAWS, above note 65, para. 16(b).

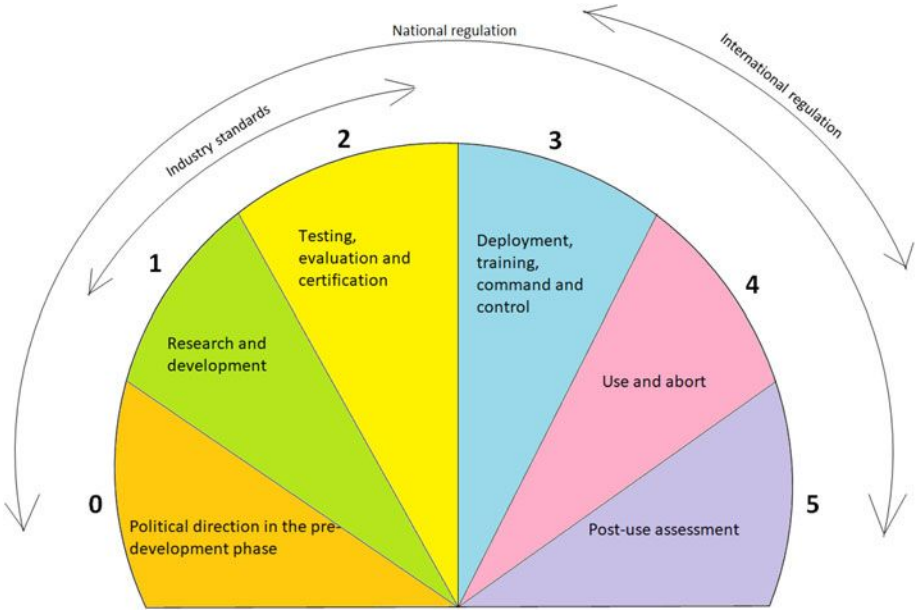


Figure 1. Touchpoints for reinforcing human involvement and oversight and for distributed governance of emerging technologies in the area of lethal autonomous weapons systems.

practices for improving compliance with international law; and a "do-nothing" approach, since IHL is fully applicable.[74]

The ten principles agreed in 2018 are applicability of IHL, non-delegation of human responsibility, accountability for use of force in accordance with international law, weapons reviews before deployment, incorporation of physical, non-proliferation and cyber security safeguards, risk assessment and mitigation during technology development, non-harm to civilian research and development (R&D) and use, the need to adopt a non-anthropomorphic perspective on AI, and the appropriateness of the CCW as a framework for dealing with the issue.[75] These are accompanied by broad understandings on the human–machine interface, on the issue of a potential definition of LAWS, and on continued review of technology pertinent to LAWS. As highlighted in Figure 1, the understandings on the human–machine interface are built around touchpoints from political direction in the pre-development phase all the way to post-use assessment. Significantly, the CCW GGE on LAWS endorsed that accountability threads together these various human–machine touchpoints in the context of the CCW. The GGE also agreed on the need to move in step with technology and

74 CCW GGE on LAWS, *Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2018/3, 23 October 2018.

75 *Ibid.*, section III.A, "Possible Guiding Principles".

build in partnership with industry and other stakeholders a common scientific and policy vernacular across the globe.[76]

The 2019 session of the GGE added an important additional principle to the above ten, building on previous work on the human–machine interface and on the applicability of IHL:

> Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular International Humanitarian Law (IHL). In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole.[77]

It can be argued that principles might not be enough, and that they would have to be clearly linked to practice and to accountability. This criticism is valid, but given the challenges of building consensus on a complex and fast-evolving issue, a principles-first approach allows agreements to be accumulated over time without prejudice to any framework that might eventually be chosen for an international agreement.

Future directions for multilateral forums on cyber security and autonomous weapons

It is safe to make three predictions in the context of digital technologies in armed conflict. Digital technologies will continue to force formerly separate fields of technology development to merge and create new use scenarios both in the civilian and military domains.[78] It will be difficult to outlaw cyber weapons, and in fact the attack surface for malicious actors will continue to expand, for example, as use of big data for predictive personal health and precision public health expands.[79] Autonomy in technology systems will continue to grow and AI will become more of a general-purpose technology used right across the military, from recruitment and training to decision-making on force.

What directions for multilateral forums are possible in this landscape? These forums' uniqueness lies in their convening ability and their inclusiveness. Their competitive advantage is in providing a neutral platform for negotiating common principles and norms in order to clarify what is acceptable and what is

76 A. S. Gill, above note 5.

77 CCW GGE on LAWS, *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, section III, para. 16.

78 As the areas of genomics, emerging infectious diseases and digital technologies converge, new challenges will also emerge for multilateral forums not covered in this survey, such as the 1972 Biological and Toxin Weapons Convention. Eleonore Pauwels, "The Internet of Living Things", *Scientific American Blog*, 25 July 2017.

79 Michael Snyder and Wenyu Zhou, "Big Data and Health", *Lancet Digital Health*, Vol. 1, 29 August 2019.

not. Today's digital challenges require these forums to expand their platform boundaries and lower entry barriers in order to bring in multiple actors, particularly technologists, industry, academia and civil society, also with a view to mitigating rising techno-nationalism. Their secretariats and office bearers need to demonstrate creativity and agility in engaging industry, technology developers and civil society. In particular, articulating a shared vision at every step is essential to drawing these different actors together.

Beyond getting the right people into the room, multilateral forums need to think about shifting their approach to the elaboration of norms, and the kind of norm-making they should prioritize in the digital age.[80] The premium placed on international treaties needs to change; instead, a flexible palette of legal, political and technical norms should be prioritized. This is not an either/or choice between general principles and binding measures, as the CCW example demonstrates. The former can lead to the latter. Further, it is not even essential that every digital issue be subject to the same set of principles. Forums can put in place their own mechanisms for discovery of foundational principles in their own context (recalling the "within the objectives and purposes" phrasing of the CCW GGE on LAWS' mandate) and then think about what approaches to implementation of these principles are feasible and effective in that context.[81] This will lead to a secondary palette of measures to "police" the norms and align action not only across nations but also across industry bodies, national governments and international bureaucracies. Again, this does not mean giving up on intrusive verification regimes wherever the stakes for compliance are very high; rather, it entails mixing and matching such measures with practical experience sharing, peer commentary and peer review to generate normative pressure.

To illustrate the above approach with the example of the CCW, the eleven principles agreed thus far in the CCW GGE on LAWS over 2017–19 can be treated as core norms.[82] These can then form the nucleus of a policy response which should include three other critical elements:

1. A set of practical understandings on national mechanisms to review potentially autonomous lethal weapons systems with regard to obligations under IHL and other applicable international law, and to rule out those systems which cannot comply with such obligations.

80 For a theoretical perspective, see Eric Talbot Jensen and Ronald T. P. Alcala (eds), *The Impact of Emerging Technologies on the Law of Armed Conflict,* Oxford University Press, New York, 2019, especially the chapter by Rebecca Crootof, "Regulating New Weapons Technology".

81 An example is the recent adoption of five principles of AI ethics by the US Department of Defense (DoD): see DoD, "DoD Adopts Ethical Principles for Artificial Intelligence", press release, 24 February 2020, available at: www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/. This should hopefully be followed by concrete measures for operationalizing the principles and clarifying their interface with legal obligations regarding AI use in weapons systems.

82 See CCW GGE on LAWS, above notes 74 and 77.

2. A set of practical understandings on enhancing the quality of the human–machine interface in AI systems that are developed and deployed in the military domain. This would allow non-lethal system components such as decision support systems that could have a bearing on the scope and intensity of conflict to also be considered.
3. A regular technology discussion (not a review mechanism) at the CCW so that policy measures can be considered for updating in light of future technology breakthroughs. The participation of technologists and industry representatives can be formalized in this context, potentially creating a new model for multilateral forums for the participation of these stakeholders.

The implementation of the above understandings would be a national prerogative in accordance with the distributed technology governance scheme mentioned previously. Nonetheless, the voluntary experience sharing in Geneva on weapons reviews and the human–machine interface should help generate peer pressure and move all tiers of governance to higher levels of diligence and responsibility. The whole scheme can be anchored in the CCW framework creatively through a decision of the next Review Conference in 2021.

What about the cyber security forums? While it is difficult now to fundamentally re-architecture the existing forums or walk back the recent recriminations on the applicability of international law, it is possible to enlarge the scope of compromises by creating additional substantive tracks. Alongside the elaboration of core norms in the UN GGE and the OEWG, there could be parallel tracks on characteristics and use scenarios, on traceability and on cooperation and assistance. These tracks should be co-led by industry and academia representatives from different regions. On the lines of the proposed technology discussion mechanism in the LAWS context, these tracks could be a model for engaging industry and subject-matter experts on the governance of digital technologies in armed conflict.

The discussion on characteristics in the context of specific instances of cyber attacks faced by industry – in the first phase these could be focused on cyber crime or ransomware – would allow different approaches to definitions to emerge. It could also uncover common evidentiary standards, approaches to assessment of context and attenuating circumstances, and criteria for damage assessment based on financial losses or down time of critical information infrastructure.

The objective of the second track would be to address the thorny issue of attribution in a non-threatening and cooperative manner. Practically, the discussion could focus on developing an independent traceability capacity, with a roster of multidisciplinary experts who will, with the assistance of designated capacities in governments and industry, establish clear, transparent and commonly accepted standards and methodologies for collecting and analyzing evidence related to the origins and conduct of attacks, and for assessment of context and of damage resulting from the attacks.[83] When the right normative

83 While different in context, traceability will also be an important aspect of ensuring accountability in the LAWS context through national or industry-level auditability of design methodologies, training data, testing and evaluation results.

framework is in place, traceability findings made in a neutral and apolitical setting should help facilitate peaceful resolution of disputes without finger-pointing, eliminate malware sources for which States do not claim knowledge or responsibility, plug vulnerabilities, and facilitate settlement of claims, including insurance claims for damages. Creating this capacity is a much more practical approach than creating an attribution and enforcement mechanism for which UN organizations do not have the requisite capabilities or political support.

The third track would develop methodologies for exchange of national experiences, information sharing and capacity-building through global and regional networks. While cyber crime knows no borders, protection of systems is still a strongly national exercise. Lack of trust and competing cultures and regulation further complicate information sharing among computer emergency response teams (CERTs) and between law enforcement and industry. De-politicization of the CERT function and development of technology-based trust tools or neutral third parties would facilitate information sharing on threats, coordinated disclosure of vulnerabilities and collaborative responses. When the normative framework is in place, this track could also provide an independent international assistance capacity with a roster of geographically diverse and independent experts who can provide assistance themselves or call on designated capacities in governments and industry to channel assistance to victims of massive cyber attacks.[84] This capacity should complement the efforts of networks of CERTs which act as first responders in case of attacks, and should focus on damage management and additional assistance to victims, going beyond the current mandate of national and regional CERTs.

The strategic aspects common to these three tracks are opening up the governance discussion to more actors; depoliticizing vulnerability assessment, assistance and cooperation; and de-emphasizing top-down, State-centric norm-making. In light of the schema proposed previously in this article, a tiered approach to norms is also intrinsic to these three tracks. Standards for robust software development, assessing and reporting vulnerabilities, non-riposte to cyber attacks etc. can be primarily developed at the industry level with the participation of national governments as observers. Norms for traceability, cooperation and assistance can be jointly developed for implementation through national laws, while State-centric intergovernmental forums can continue to deal with the applicability of international law and the development of new international norms. Like sliding parallel doors, these three modalities of regulation can together provide flexible responses to the challenge posed by fast-moving digital technologies to multilateral governance of digital conflict.

84 Such panels have proved to be reasonably successful in other contexts; a roster of experts has been maintained by the UN Secretary-General for chemical and biological attacks. See UN Office for Disarmament Affairs, "Secretary-General's Mechanism for Investigation of Alleged Use of Chemical and Biological Weapons", available at: www.un.org/disarmament/wmd/secretary-general-mechanism/.

Concluding summary

Armed conflict is taking new forms in the digital age. Both attribution and accountability are harder to establish, and the boundaries between States, the subject of regulation in multilateral forums thus far, and non-State actors are blurred. Technology itself is shifting ceaselessly and is weaponizable in unpredictable ways. Multilateral norms and norm-making practices need to adjust with agility and effectiveness. This article has highlighted a set of challenges, such as over-structured agendas, State-centrism and limited output modalities, that multilateral forums need to overcome if they are to be successful in dealing with the impact of digital technologies on international security. It has looked at recent multilateral attempts to regulate existing cyber weapons and emerging lethal autonomous weapons systems. Based on the experience of these negotiations and other trends, it has suggested a future direction for the CCW GGE on LAWS as well as for the forums dealing with cyber weapons. The key signposts for the proposed future work are interdisciplinarity, with strong engagement by the private sector; a tiered approach to norm-making, with international regulation moving in step with national regulation and industry standards; and a modular build-up of obligations anchored in guiding principles.

International Review of the Red Cross (2020), **102** (913), 287–334.
Digital technologies and war
doi:10.1017/S1816383120000387

IRRC

Twenty years on: International humanitarian law and the protection of civilians against the effects of cyber operations during armed conflicts

Laurent Gisel, Tilman Rodenhäuser and Knut Dörmann*

Laurent Gisel is Head of the Arms and Conduct of Hostilities Unit in the Legal Division of the International Committee of the Red Cross (ICRC) in Geneva. Between 2013 and 2020, he was the ICRC's Senior Legal Adviser, Cyber Team Leader and file holder for the rules governing the conduct of hostilities under international humanitarian law, including their application during urban warfare, cyber operations, and outer space operations.

Dr Tilman Rodenhäuser is a Legal Adviser at the ICRC, working on cyber operations during armed conflict, non-State armed groups, and detainee transfers.

* An earlier version of this article has been published by the same authors under the title "The Applicability and Application of International Humanitarian Law to Cyber Warfare", *Chinese Review of International Law*, Vol. 32, No. 4, 2019. It has been substantially updated and broadened for publication in this issue of the *Review*. This article was written in a personal capacity and does not necessarily reflect the views of the ICRC.

Dr Knut Dörmann is Head of the ICRC's Delegation to the EU, NATO and the Kingdom of Belgium (Brussels), and former Chief Legal Officer and Head of the Legal Division (2007–19). Prior to that, he was Deputy Head of the ICRC's Legal Division (2004–07) and a Legal Adviser to the ICRC (1999–2004), including on cyber operations.

Abstract

The use of cyber operations during armed conflicts and the question of how international humanitarian law (IHL) applies to such operations have developed significantly over the past two decades. In their different roles in the Legal Division of the International Committee of the Red Cross (ICRC), the authors of this article have followed these developments closely and have engaged in governmental and non-governmental expert discussions on the subject. In this article, we analyze pertinent humanitarian, legal and policy questions. We first show that the use of cyber operations during armed conflict has become a reality of armed conflicts and is likely to be more prominent in the future. This development raises a number of concerns in today's increasingly cyber-reliant societies, in which malicious cyber operations risk causing significant disruption and harm to humans. Secondly, we present a brief overview of multilateral discussions on the legal and normative framework regulating cyber operations during armed conflicts, looking in particular at various arguments around the applicability of IHL to cyber operations during armed conflict and the relationship between IHL and the UN Charter. We emphasize that in our view, there is no question that cyber operations during armed conflicts, or cyber warfare, are regulated by IHL – just as is any weapon, means or methods of warfare used by a belligerent in a conflict, whether new or old. Thirdly, we focus the main part of this article on how IHL applies to cyber operations. Analyzing the most recent legal positions of States and experts, we revisit some of the most salient debates of the past decade, such as which cyber operations amount to an "attack" as defined in IHL and whether civilian data enjoys similar protection to "civilian objects". We also explore the IHL rules applicable to cyber operations other than attacks and the special protection regimes for certain actors and infrastructure, such as medical facilities and humanitarian organizations.

Keywords: cyber operations, armed conflict, cyber warfare, human cost, international humanitarian law.

: : : : : : :

The use of cyber operations during armed conflicts and the question of how international humanitarian law (IHL) applies to such operations have developed significantly over the past two decades. This finding is true at the operational, legal and political levels. Operationally, the use of cyber operations during armed conflict has become a reality of armed conflicts and is likely to be more prominent in the future. This development raises a number of concerns in

today's ever more cyber-reliant societies, in which malicious cyber operations risk causing significant disruption and harm to humans. At the political and legal levels, through multilateral processes States have achieved agreement on some aspects of the legal and normative framework regulating cyber operations; however, the application of IHL to cyber operations during armed conflict remains the subject of intense discussion. A few States have published positions on how IHL applies to cyber operations during armed conflicts, and a wealth of academic studies exist on the matter, yet key issues remain controversial and lack agreement among States and other experts, or require further analysis. These include the notion of "attack", the question of how civilian data are protected against harmful cyber operations, and which IHL rules apply to cyber operations other than attacks. In their different roles in the Legal Division of the International Committee of the Red Cross (ICRC), the authors of this article have followed these developments and debates closely and have engaged in governmental and non-governmental expert discussions on the applicability and application of IHL to cyber operations during armed conflicts since their beginning.

The ICRC has recently published a comprehensive institutional position on *International Humanitarian Law and Cyber Operations during Armed Conflicts*, which was submitted to the United Nations (UN) Group of Governmental Experts (GGE) and Open-Ended Working Group (OEWG).[1] In this article, we expand on this position and first explain why the potential human cost of cyber operations is a humanitarian concern. We then underline that IHL applies to – and therefore restricts – cyber operations during armed conflicts and examine different States' views on this subject. Third, we analyze when cyber operations may trigger an armed conflict and how this threshold relates to the prohibition of the use of force and the right to self-defence under the UN Charter and customary international law. In the last and most substantial part of the article, we delve into some of the long-standing questions on how IHL applies to cyber operations during armed conflicts and what positions States have taken on some of the key issues.

Operationally, the use of cyber technology has become a reality in today's armed conflicts and is likely to increase in the future. Some States have acknowledged publicly that they have conducted cyber operations in ongoing armed conflicts. In particular, the United States, the United Kingdom and Australia have disclosed that they used cyber operations in their conflict against the Islamic State group.[2] There are also public reports suggesting that Israel used

1 ICRC, *International Humanitarian Law and Cyber Operations during Armed Conflicts*, Position Paper submitted to the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security and the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, 2019, available at: www.icrc.org/en/document/international-humanitarian-law-and-cyber-operations-during-armed-conflicts (all internet references were accessed in August 2020). Also available in the "Reports and Documents" section of this issue of the *Review*.

2 See, in particular, Mike Burgess, Australian Signals Directorate, "Offensive Cyber and the People Who Do It", speech given to the Lowy Institute, 27 March 2019, available at: www.asd.gov.au/speeches/20190327-lowy-institute-offensive-cyber-operations.htm; Paul M. Nakasone, "Statement of General Paul

cyber operations against Hamas – and allegations that Hamas used cyber operations against Israel.[3] Furthermore, cyber operations have affected other countries involved in armed conflicts, such as Georgia in 2008,[4] Ukraine in 2015–17[5] and Saudi Arabia in 2017,[6] though the authors of these cyber attacks remain unknown and attribution of responsibility is contested. It is therefore unclear whether these operations had a nexus to the respective armed conflicts and thus whether IHL applied. Moreover, there have been reports of cyber operations by States in other situations where the legal classification may not be straightforward, including in what is sometimes referred to as a "grey zone".[7] These examples show an increase in military cyber operations over the past decade – a change in warfare that might continue. Indeed, an increasing number of States are said to have or to be developing cyber military capabilities, including the five permanent member States of the UN Security Council.[8] Examples of the use of cyber operations

M. Nakasone, Commander, United States Cyber Command, before the Senate Committee on Armed Services", 14 February 2019, available at: www.armed-services.senate.gov/imo/media/doc/Nakasone_02-14-19.pdf; Jeremy Fleming, GCHQ, "Director's Speech at CyberUK18", 12 April 2018, available at: www.gchq.gov.uk/pdfs/speech/director-cyber-uk-speech-2018.pdf.

3 "Hackers Interrupt Israeli Eurovision WebCast with Faked Explosions", *BBC News*, 15 May 2019, available at: www.bbc.co.uk/news/technology-48280902; Zak Doffman, "Israel Responds to Cyber Attack with an Air Strike on Cyber Attackers in World First", *Forbes*, 6 May 2019, available at: www.forbes.com/sites/zakdoffman/2019/05/06/israeli-military-strikes-and-destroys-hamas-cyber-hq-in-world-first/#1c692f73afb5. While the purported target of the alleged cyber operation by Hamas has not been publicly released, the targeting of Hamas' building by kinetic means was said to be based on intelligence gained as part of the Israeli Defence Forces' cyber defence effort.

4 David Hollis, "Cyberwar Case Study: Georgia 2008", *Small War Journal*, 2010, available at: https://smallwarsjournal.com/blog/journal/docs-temp/639-hollis.pdf.

5 Andy Greenberg, "How an Entire Nation Became Russia's Test Lab for Cyberwar", *Wired*, 20 June 2017, available at: www.wired.com/story/russian-hackers-attack-ukraine/; Andy Greenberg, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History", *Wired*, 22 August 2018, available at: www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/.

6 Blake Johnson *et al.*, "Attackers Deploy New ICS Attack Framework 'TRITON' and Cause Operational Disruption to Critical Infrastructure", *Fireeye Blogs*, 14 December 2017, available at: www.fireeye.com/blog/threat-research/2017/12/attackers-deploy-new-ics-attack-framework-triton.html.

7 For example, there have been various media reports – based on anonymous official sources – that the United States has carried out cyber operations against targets in Russia and Iran, and that Israel has carried out a cyber operation against a port in Iran. See Ellen Nakashima, "U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms", *Washington Post*, 27 February 2019, available at: https://tinyurl.com/yxs8twyv; David E. Sanger and Nicole Perlroth, "U.S. Escalates Online Attacks on Russia's Power Grid", *New York Times*, 15 June 2019, available at: www.nytimes.com/2019/06/15/us/politics/trump-cyber-russia-grid.html; Julian E. Varnes and Thomas Gibbons-Neff, "U.S. Carried out Cyberattacks on Iran", *New York Times*, 22 June 2019, available at: www.nytimes.com/2019/06/22/us/politics/us-iran-cyber-attacks.html; Joby Warrick and Ellen Nakashima, "Officials: Israel Linked to a Disruptive Cyberattack on Iranian Port Facility", *Washington Post*, 18 May 2020, available at: https://tinyurl.com/y4onsrt9. On so-called "grey zones" and cyber technology, see Camille Faure, "Utilisation contemporaine et future des technologies cyber/numériques dans les conflits armés", in Gabriella Venturini and Gian Luca Beruto (eds), *Whither the Human in Armed Conflict? IHL Implications of New Technology in Warfare*, 42nd Round Table on Current Issues of International Humanitarian Law, International Institute of Humanitarian Law, Sanremo, 2020 (forthcoming); Gary Corn, "Punching on the Edges of the Grey Zone: Iranian Cyber Threats and State Cyber Responses", *Just Security*, 11 February 2020, available at: www.justsecurity.org/68622/punching-on-the-edges-of-the-grey-zone-iranian-cyber-threats-and-state-cyber-responses/. On the threshold of application of IHL, see the section below entitled "Cyber Operations that Are Governed by IHL".

8 In addition to the United States and the United Kingdom, France has set out the objective of "acquir[ing] a cyber defence capability" to defend against "foreign States or terrorist groups [which] could attack the

during conflicts include espionage; target identification; information operations to affect the enemy's morale and will to fight; the interruption, deception or obfuscation of the enemy's communication systems aimed at hindering force coordination; and cyber operations in support of kinetic operations.[9] An example of the latter is the disabling of an enemy's military radar stations in support of air strikes.[10] Moreover, as seen in a range of cyber operations over the past decade – which may not necessarily have occurred in the context of armed conflicts – cyber operations against electricity grids, health-care systems, nuclear facilities or other critical infrastructure risk causing significant human harm.[11] Legally, discussions on whether and how international humanitarian law applies to, and restricts, cyber operations during armed conflicts began over two decades ago.[12] The drafting processes for the two *Tallinn Manuals on the International Law Applicable to Cyber Operations* (Tallinn Manuals) have shown that there is

critical infrastructures". France, Agence Nationale de la Sécurité des Système d'Information, *Information System Defence and Security: France's Strategy*, 2011, available at: www.ssi.gouv.fr/uploads/IMG/pdf/2011-02-15_Information_system_defence_and_security_-_France_s_strategy.pdf. The 2015 *White Paper on China's Military Strategy* states that "in response to the increasing development of cyber military capabilities from other states, China will develop a defensive cyber military capacity". See Government of China, *White Paper on China's Military Strategy*, 2015, available at: www.gov.cn/zhengce/2015-05/26/content_2868988.htm. Russia has been less explicit on the subject, but the Russian Federation's Doctrine of Information Security identifies "upgrading the information security system of the Armed Forces of the Russian Federation, other troops, military formations and bodies, including forces and means of information confrontation" as a "key area of ensuring information security in the field of national defence". See Ministry of Foreign Affairs of the Russian Federation, *Doctrine of Information Security of the Russian Federation*, 5 December 2016, available at: https://tinyurl.com/y6yhp7pv. See also Ministry of Defence of the Russian Federation, "Western MD Operators Repelled Cyberattack of the Simulated Enemy in the Course of the Union Shield – 2015", 2015, available at: https://eng.mil.ru/en/news_page/country/more.htm?id=12056193@egNews. For general estimates on the spread of cyber tools, see Anthony Craig, "Understanding the Proliferation of Cyber Capabilities", Council on Foreign Relations, 2018, available at: www.cfr.org/blog/understanding-proliferation-cyber-capabilities. According to the United Nations Institute for Disarmament Research (UNIDIR) Cyber Index, in 2012 forty-seven States had cyber security programmes that gave some role to their armed forces (UNIDIR, *The Cyber Index: International Security Trends and Realities*, UN Doc. UNIDIR/2013/3, Geneva, 2013, p. 1), while in 2020 Digital Watch Observatory recorded twenty-three and thirty States with respectively evidence or indications of offensive cyber capabilities (Digital Watch Observatory, "UN GGE and OEWG", available at: https://dig.watch/processes/un-gge).

9 ICRC, *Avoiding Civilian Harm from Military Cyber Operations during Armed Conflicts*, forthcoming.

10 Sharon Weinberger, "How Israel Spoofed Syria's Air Defense System", *Wired*, 4 October 2007, available at: www.wired.com/2007/10/how-israel-spoo/; Lewis Page, "Israeli Sky-Hack Switched Off Syrian Radars Countrywide", *The Register*, 22 November 2007, available at: www.theregister.co.uk/2007/11/22/israel_air_raid_syria_hack_network_vuln_intrusion/.

11 In November 2018, the ICRC convened an expert meeting to develop a realistic assessment of cyber capabilities and their potential humanitarian consequences in light of their technical characteristics. See Laurent Gisel and Lukasz Olejnik (eds), *ICRC Expert Meeting: The Potential Human Cost of Cyber Operations*, ICRC, Geneva, 2019, available at: www.icrc.org/en/download/file/96008/the-potential-human-cost-of-cyber-operations.pdf. See also Sergio Caltagirone, "Industrial Cyber Attacks: A Humanitarian Crisis in the Making", *Humanitarian Law and Policy Blog*, 3 December 2019, available at: https://blogs.icrc.org/law-and-policy/2019/12/03/industrial-cyber-attacks-crisis/. The World Economic Forum (WEF) *Global Risks Report 2020* ranks cyber attacks among the top ten risks in terms of both likelihood and impact; see WEF, *The Global Risks Report 2020*, 2020, p. 3, available at: www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf.

12 See US Department of Defense Office of General Counsel, *An Assessment of International Legal Issues in Information Operations*, 1999, available at: https://fas.org/irp/eprint/io-legal.pdf; for one of the early academic examinations of these questions, see Knut Dörmann, "Computer Network Attack and

significant consensus among experts that IHL applies in cyberspace and that its basic rules and principles can and must be applied when conducting cyber operations during armed conflict.[13] As seen in the diverging views recorded in the Tallinn Manuals as well as in a growing number of State positions and the rich body of academic publications on cyber-related issues, however, various aspects of how certain rules of IHL apply in this field remain under-explored and disagreement exists on other questions, including some of the most-examined ones (see the section below on "The Limits that IHL Imposes on the use of Cyber Capabilities during Armed Conflicts"). At the political level, recent and ongoing discussions in the UN have shown that finding agreement on the applicability of IHL to cyber operations and furthering the study of how its rules should be interpreted remains challenging.[14] Discussions on questions relating to "information security" started when the Russian Federation introduced a first resolution on the subject at the UN General Assembly in 1998. These discussions have intensified over the course of the last few years. Since 2004, governmental experts have been meeting in six consecutive Groups of Governmental Experts on questions relating to information and telecommunications in the context of international security. In 2018, the UN General Assembly also established the OEWG, which runs in parallel to the GGEs. Both groups are mandated, *inter alia*, to study "how international law applies to the use of information and communications technologies by States".[15] These discussions should build on the important conclusions reached by previous GGEs. In 2013 and 2015, States in the GGE affirmed that "international law and in particular the Charter of the United Nations is applicable" in the information and communication technology environment and cited "the established international legal principles, including, where applicable, the principles of humanity, necessity, proportionality and distinction".[16] Yet, it appears from recent discussions in these UN processes, and as discussed further below, that finding agreement on the applicability of IHL to cyber operations and furthering the study of how its rules should be interpreted is challenging.

International Humanitarian Law", 2001, available at: www.icrc.org/en/doc/resources/documents/article/other/5p2alj.htm.

13 See Michael N. Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Cambridge University Press, Cambridge, 2013 (Tallinn Manual); Michael N. Schmitt and Liis Vihul (eds), *Tallinn Manual 2.0 on International Law Applicable to Cyber Operations*, 2nd ed., Cambridge University Press, Cambridge, 2017 (Tallinn Manual 2.0).

14 See, notably, OEWG, "Initial 'Pre-draft' of the Report of the OEWG on Developments in the Field of Information and Telecommunications in the Context of International Security", 11 March 2020, available at: https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/03/200311-Pre-Draft-OEWG-ICT.pdf.

15 UNGA Res. 73/27, "Developments in the Field of Information and Telecommunications in the Context of International Security", UN Doc. A/RES/73/27, 11 December 2018, para. 5; UNGA Res. 73/266, "Advancing Responsible State Behaviour in Cyberspace in the Context of International Security", UN Doc. A/RES/73/266, 2 January 2019, para. 3.

16 UN General Assembly, "Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security: Note by the Secretary-General", UN Doc. A/70/174, 22 July 2015, paras 24, 28(d).

At the regional level, already in 2009 member States of the Shanghai Cooperation Organization (SCO) had identified "[d]eveloping and using information weapons" and "preparing and conducting information warfare" to be a major threat in the field of international information security, but they remained silent on the applicable legal framework.[17] Discussions on the application of international law, including IHL, have taken place in the Asian–African Legal Consultative Organization (AALCO) (which established an Open-Ended Working Group on International Law in Cyberspace in 2015)[18], the Commonwealth,[19] the European Union,[20] the North Atlantic Treaty Organization (NATO)[21] and the Organization of American States (OAS),[22] among others.

The potential human cost of cyber operations

The development of information and communication technology, including communication over computer networks (cyberspace), offers tremendous benefits and opportunities for States, societies and individuals in the social, economic, development, and information and communication realms, among others. The international community, societies, and each of us individually are increasingly relying on digital tools. This trend – which may be accelerated further by the COVID-19 pandemic spreading at the time of writing this article – increases our dependency on the uninterrupted functioning of these technologies, and thus increases our vulnerability to cyber operations. The rapidly evolving nature of cyberspace and cyber technology, and the potential human cost of cyber operations, therefore necessitates constant monitoring and assessment.

The use of cyber tools as a means or method of warfare offers militaries the possibility of achieving their objectives without necessarily causing direct harm to civilians or physical damage to civilian infrastructure. Depending on the circumstances, cyber operations might enable targeting a military objective while reducing the expected incidental damage to civilian objects compared to the use

17 Agreement on Cooperation in the Field of Ensuring International Information Security among Member States of the Shanghai Cooperation Organization, Yekaterinburg, 16 June 2009 (SCO Agreement); unofficial translation in Ministry of Defense of the Russian Federation, "The State and the Prospects of Russian Military Cooperation on International Information Security (A Collection of Papers)", 2014, pp. 77 ff. See also, for example, J. Fleming, above note 2, p. 5.

18 See AALCO, *International Law in Cyberspace*, Doc. No. AALCO/58/DAR ES SALAAM/2019/SD/17, available at: www.aalco.int/Final%20Cyberspace%202019.pdf.

19 See the Commonwealth Cyber Declaration issued at the Commonwealth Heads of Government Meeting, London, 16–20 April 2018, available at: https://thecommonwealth.org/commonwealth-cyber-declaration.

20 See, for example, EU Council Conclusions, General Affairs Council meeting, Doc. No. 11357/13, 25 June 2013.

21 See, for example, the Wales Summit Declaration issued by the heads of State and government participating in the meeting of NATO in Wales, 5 September 2014, para. 72, available at: www.nato.int/cps/en/natohq/official_texts_112964.htm.

22 See OAS, *Improving Transparency: International Law and State Cyber Operations: Fourth Report*, OAS Doc. CJI/doc. 603/20 rev.1 corr.1, 5 March 2020, available at: www.oas.org/en/sla/iajc/docs/CJI_doc_603-20_rev1_corr1_eng.pdf.

of other means of warfare. In recent intergovernmental discussions, some States emphasized that if employed responsibly and in accordance with international law, "the use of ICTs [information and communications technologies] in military contexts may be preferable to use of kinetic weapons and can be de-escalatory".[23] In contrast, as noted above, SCO member States have warned of the dangers of "[d]eveloping and using information weapons" and "preparing and conducting information warfare".[24]

Conducting highly discriminative cyber operations that comply with IHL and spare civilian populations can be technologically challenging. The interconnectivity that characterizes cyberspace means that whatever has an interface with the Internet can be affected by cyber operations conducted from anywhere in the world. A cyber attack on a specific system may have repercussions on various other systems, regardless of where those systems are located. There is a real risk that cyber tools – either deliberately or by mistake – may cause large-scale and diverse effects on critical civilian infrastructure. The interconnectedness of cyberspace also means that all States should be concerned with its effective regulation: "attacks carried out against one State can affect many others – wherever they are located and irrespective of whether they are involved in the conflict".[25] Cyber operations conducted over recent years – primarily outside armed conflicts – have shown that malware can spread instantly around the globe and affect civilian infrastructure and the provision of essential services.[26] As a result, commentators are warning that industrial cyber attacks present "a humanitarian crisis in the making".[27]

The health-care sector seems particularly vulnerable to cyber attacks.[28] The sector is moving towards increased digitization and interconnectivity, which increases its digital dependency and its attack surface – a development that is likely to continue in the coming years. Too often, these developments have not been matched by a corresponding improvement in cyber security.[29]

23 "UK Response to Chair's Initial 'Pre-draft' of the Report of the OEWG on Developments in the Field of Information and Telecommunications in the Context of International Security", available at: https://front.un-arm.org/wp-content/uploads/2020/04/20200415-oewg-predraft-uk.pdf. See also ICRC, above note 9; Gary Corn, "The Potential Human Costs of Eschewing Cyber Operations", *Humanitarian Law and Policy Blog*, 31 May 2019, available at: https://blogs.icrc.org/law-and-policy/2019/05/31/potential-human-costs-eschewing-cyber-operations/.

24 SCO Agreement, above note 17, Art. 2.

25 Helen Durham, "Cyber Operations during Armed Conflict: 7 Essential Law and Policy Questions", *Humanitarian Law and Policy Blog*, 26 March 2020, available at: https://blogs.icrc.org/law-and-policy/2020/03/26/cyber-armed-conflict-7-law-policy-questions/.

26 Examples include the malware CrashOverride, the ransomware WannaCry, the wiper program NotPetya, and the malware Triton. CrashOverride affected the provision of electricity in Ukraine; WannaCry affected hospitals in several countries; NotPetya affected a very large number of businesses; Triton was aimed at disrupting industrial control systems, and was reportedly used in attacks against Saudi Arabian petrochemical plants. For some discussion, see Laurent Gisel and Lukasz Olejnik, "The Potential Human Cost of Cyber Operations: Starting the Conversation", *Humanitarian Law and Policy Blog*, 14 November 2018, available at: https://blogs.icrc.org/law-and-policy/2018/11/14/potential-human-cost-cyber-operations/.

27 See S. Caltagirone, above note 11.

28 L. Gisel and L. Olejnik (eds), above note 11, pp. 18–22.

This vulnerability became particularly apparent during the COVID-19 pandemic, when hospitals and other health-care facilities in various States had their operations disrupted by hostile cyber operations. In light of the particular importance of the health-care sector for mitigating suffering at all times, but especially during armed conflicts and health crises, the ICRC has called on all States to respect and protect medical services and medical facilities against cyber attacks of any kind, whether in time of peace or in time of conflict, and to reaffirm and recommit to international rules that prohibit such actions.[30] While this call reflects existing obligations under IHL as applicable to cyber operations during armed conflict,[31] it would reaffirm, or arguably strengthen, existing prohibitions under public international law that apply at all times.[32]

Cyber operations against other critical civilian infrastructure, such as electricity, water and sanitation, can also cause significant harm to humans.[33] This infrastructure is often operated by industrial control systems (ICSs). A cyber attack against an ICS requires specific expertise and sophistication, and often, custom-made malware. While ICS attacks have been less frequent than other types of cyber operations, their frequency is reportedly increasing and the severity of the threat has evolved more rapidly than anticipated only a few years ago.[34] Cyber security specialists have pointed out that "due to the potential of cyber-physical attacks to have kinetic effect and cause casualties, it is urgent and of utmost importance for the international community of IT security specialists,

29 See Aaron F. Brantly, "The Cybersecurity of Health", *Council on Foreign Relations Blog*, 8 April 2020, available at: https://tinyurl.com/yxc4oc9j.

30 See "Call by Global Leaders: Work Together Now to Stop Cyberattacks on the Healthcare Sector", *Humanitarian Law and Policy Blog*, 26 May 2020, available at: https://blogs.icrc.org/law-and-policy/2020/05/26/call-global-leaders-stop-cyberattacks-healthcare/. In the specific framework of the above-mentioned OEWG, the ICRC suggested that States could adopt a norm whereby they commit "not to conduct or knowingly support cyber operations that would harm medical services or medical facilities, and to take measures to protect medical services from harm". This suggestion combines a "negative" element, namely that States should not conduct or knowingly support cyber activity that would harm medical services or facilities, and a "positive" element, meaning that States should take measures to protect medical services from harm. See ICRC, "Norms for Responsible State Behavior on Cyber Operations Should Build on International Law", 11 February 2020, available at: www.icrc.org/en/document/norms-responsible-state-behavior-cyber-operations-should-build-international-law.

31 See below section entitled "IHL Rules Protecting Objects Indispensable to the Survival of the Civilian Population, Medical Services, and Humanitarian Relief Operations".

32 For greater detail on how international law applies to such operations, see Kubo Mačák, Laurent Gisel and Tilman Rodenhäuser, "Cyber Attacks against Hospitals and the COVID-19 Pandemic: How Strong are International Law Protections?", *Just Security*, 27 March 2020, available at: www.justsecurity.org/69407/cyber-attacks-against-hospitals-and-the-covid-19-pandemic-how-strong-are-international-law-protections/. See also the Oxford Statement on the International Law Protections against Cyber Operations Targeting the Health Care Sector, May 2020 (Oxford Statement), available at: www.elac.ox.ac.uk/the-oxford-statement-on-the-international-law-protections-against-cyber-operations-targeting-the-hea.

33 L. Gisel and L. Olejnik (eds), above note 11, pp. 23–28. See also Aron Heller, "Israeli Cyber Chief: Major Attack on Water Systems Thwarted", *ABC News*, 28 May 2020, available at: https://abcnews.go.com/International/wireStory/israeli-cyber-chief-major-attack-water-systems-thwarted-70920855.

34 *Ibid.*, p. 25.

governments and humanitarian lawyers to have a conversation about how to regulate the deployment of cyber-physical attacks".[35]

As will be discussed later on in this article, in an armed conflict IHL protects the health-care sector quite comprehensively, and it prohibits attacks on civilian infrastructure, unless such infrastructure has become a military objective.

Thinking beyond the impact of cyber operations on specific infrastructure, there are at least three characteristics of cyber operations that raise further concern.[36]

First, though not impossible, attributing cyber attacks to a State or non-State actor has proven challenging.[37] This hampers the possibility of identifying actors who violate IHL in cyberspace and holding them responsible, which is one way to ensure compliance with IHL. Plausible deniability, and the hope of remaining covered, may also alter the political calculations involved in conducting cyber attacks – and in conducting them in violation of international law.

Second, as noted for example by China, "the proliferation of malicious cyber tools and technology [is] rising".[38] Cyber tools and methods can indeed proliferate in a unique manner that is difficult to control. Today, sophisticated cyber attacks are only carried out by the most advanced and best-resourced actors. Once malware is used, stolen, leaked or otherwise becomes available, however, actors other than those who developed the malware might be able to find it online, reverse engineer it, and reuse it for their own purposes.

Third, cyber operations bear a risk of overreaction by targeted States and a subsequent escalation in violence. For the target of a cyber attack, it is normally difficult to know whether the attacker aims at espionage or at causing other, potentially physical damage. The aim of a cyber operation may only be identified with certainty once the effect or the end goal is achieved. Hence, there is a risk that the target of an operation will anticipate worst-case impact and react in a stronger manner than if it knew that the attacker's intent was solely espionage.

At the time of writing, cyber operations have not caused major human harm. However, significant economic harm has been caused.[39] With regard to the

35 Marina Krotofil, "Casualties Caused through Computer Network Attacks: The Potential Human Costs of Cyber Warfare", 42nd Round Table on Current Issues of International Humanitarian Law, 2019, available at: http://iihl.org/wp-content/uploads/2019/11/Krotofil1.pdf.

36 See also ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, Geneva, 2019 (ICRC Challenges Report 2019), p. 27, available at: www.icrc.org/en/document/icrc-report-ihl-and-challenges-contemporary-armed-conflicts; L. Gisel and L. Olejnik (eds), above note 11, p. 7.

37 For a broader discussion on attribution, including the pertinent international law rules, see the section below entitled "The Issue of Attribution".

38 Statement by Counsellor Sun Lei of the Chinese Delegation at the Thematic Discussion on Information and Cyber Security at the First Committee of the 72nd Session of the UN General Assembly, 23 October 2017, available at: www.china-un.org/eng/chinaandun/disarmament_armscontrol/unga/t1505683.htm.

39 The overall cost of cyber crime alone is measured in trillions of dollars: it was estimated at $3 trillion in 2015 worldwide, and this figure is predicted to double by 2021 (Steve Morgan, "Hackerpocalypse: A Cybercrime Revelation", Herjavec Group, 17 August 2016, available at: www.herjavecgroup.com/hackerpocalypse-cybercrime-report/). NotPetya's impact was estimated at well above $1 billion, with some estimates as high at $10 billion (Fred O'Connor, "NotPetya Still Roils Company's Finances,

potential human cost of cyber operations, much is unknown in terms of technological evolution, the capabilities and the tools developed by the most sophisticated actors – including military ones – and the extent to which the use of cyber operations during armed conflicts might be different from the trends observed so far. In other words, while the risk of human cost does not appear extremely high based on current observations, especially considering the destruction and suffering that conflicts always cause, the evolution of cyber operations requires close attention due to existing uncertainties and the rapid pace of change.

The applicability of IHL to cyber operations during armed conflicts

From a legal point of view, the primary framework imposing limitations on the use of cyber operations during armed conflict and protecting civilian populations against potential harm is international humanitarian law.

IHL does not contain a definition of cyber operations, cyber warfare or cyber war, and neither do other fields of international law. Various definitions of cyber operations have been used in military or other documents by certain States.[40] Other States refer instead to information warfare or information war and define this notion in a manner that includes at least some aspects of what is often understood as cyber warfare.[41] Irrespective of how cyber operations, cyber warfare or information warfare are defined by States and others, the determination of whether IHL applies to such operations has to be made based on the nature, effects and circumstances of such operations.

The ICRC understands "cyber operations during armed conflict" to mean operations against a computer system or network, or another connected device, through a data stream, when used as means or method of warfare in the context of an armed conflict.[42]

Costing Organizations $1.2 Billion in Revenue", *Cybereason*, 9 November 2017, available at: www.cybereason.com/blog/notpetya-costs-companies-1.2-billion-in-revenue; A. Greenberg, above note 5). The financial system is also often affected by cyber attacks: see, for example, Choe Sang-Hun, "Computer Networks in South Korea Are Paralyzed in Cyberattacks", *New York Times*, 20 March 2013, available at: www.nytimes.com/2013/03/21/world/asia/south-korea-computer-network-crashes.html.

40 See, for instance, US Department of Defense, *DOD Dictionary of Military and Associated Terms*.

41 The SCO Agreement, above note 17, defines "information war" as "a confrontation between two or more States in the information space with the aim of damaging information systems, processes and resources, critically important and other structures, undermining political, economic and social systems, psychologically manipulating masses of the population to destabilize society and the State, and also forcing the State to take decisions in the interest of the opposing party". The Russian Federation Armed Forces define information war in the same manner, stating that "the Armed Forces of the Russian Federation follow … international humanitarian law" during military activities in the global information space (Ministry of Defence of the Russian Federation, *Russian Federation Armed Forces' Information Space Activities Concept*, 2011, section 2.1, available at: https://eng.mil.ru/en/science/publications/more.htm?id=10845074@cmsArticle).

42 See ICRC, above note 1.

While there continues to be debate on the question of whether IHL applies to, and therefore restricts, cyber operations during armed conflict, from the outset the ICRC has taken a clear and affirmative position.[43] In the ICRC's view, there is no question that cyber operations during armed conflicts, or cyber warfare, are regulated by IHL – just as is any weapon, means or method of warfare used by a belligerent in a conflict, whether new or old. The fact that cyber operations rely on new and continuously developing technology does not prevent the application of IHL to the use of such technologies as means or methods of warfare. This holds true whether cyberspace is considered as a new domain of warfare similar to air, land, sea and outer space; as a different type of domain because it is man-made while the former are natural; or not as a domain as such.

In our view, there is cogent support for this view in IHL treaties, in the jurisprudence of the International Court of Justice (ICJ), and in the views expressed by a number of States and international organizations.

The very object and purpose of IHL is to regulate future conflicts, meaning those that occur after the adoption of an IHL treaty. When adopting IHL treaties, States included norms that anticipate the development of new means and methods of warfare and presumed that IHL will apply to them. Already in 1868, the St Petersburg Declaration intended that the principles it established should be maintained with respect to "future improvements which science may effect in the armament of troops".[44] An important and more recent IHL rule in this respect is found in Article 36 of the 1977 Additional Protocol I (AP I),[45] which states:

> In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

Undoubtedly, this obligation is based on the assumption that IHL applies to such new weapons, means and methods – otherwise, it would not be necessary to review their lawfulness under existing law. This includes weapons, means and methods of warfare that rely on cyber technology.

The conclusion that IHL applies to cyber operations during armed conflict finds further support in the views expressed by the ICJ. In its Advisory Opinion on the legality of the threat or use of nuclear weapons, the Court recalled that the established principles and rules of humanitarian law applicable in armed conflict apply "to all forms of warfare and to all kinds of weapons", including "those of

43 See ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, Geneva, 2011 (ICRC Challenges Report 2011), pp. 36–39, available at: www.icrc.org/en/doc/assets/files/red-cross-crescent-movement/31st-international-conference/31-int-conference-ihl-challenges-report-11-5-1-2-en.pdf; K. Dörmann, above note 12.

44 Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight, St Petersburg, 29 November/11 December 1868.

45 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).

the future".[46] Again, this includes cyber operations. This view is also widely recognized among experts.[47]

There is increasing recognition among States that international law applies in cyberspace, and in particular that IHL applies to, and therefore restricts, cyber operations during armed conflicts. As mentioned above, in the 2013 and 2015 reports of the UN GGE, experts concluded that "international law, and in particular the Charter of the United Nations, is applicable" in the information and communication technologies environment,[48] a conclusion that has been first welcomed[49] and then confirmed[50] by the UN General Assembly. The 2015 report also cited "the established international legal principles, including, where applicable, the principles of humanity, necessity, proportionality and distinction".[51] While this list of principles does not mention IHL explicitly, commentators have pointed out that these are "IHL's core principles".[52]

In line with this conclusion, an increasing number of States and international organizations have publicly asserted that IHL applies to cyber operations during armed conflict. This includes, for example, the EU[53] and NATO.[54] Moreover, the Paris Call for Trust and Security in Cyberspace (supported by seventy-eight States as of April 2020) has reaffirmed the applicability of IHL to cyber operations during armed conflict;[55] the heads of government of the 54 Commonwealth States have "[c]ommit[ted] to move forward discussions on how … applicable international humanitarian law, applies in cyberspace in all its aspects";[56] and States' responses to a study conducted by

46 ICJ, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, 8 July 1996, para. 86.

47 See Tallinn Manual 2.0, above note 13, Rule 80; Oxford Statement, above note 32, point 5. Also see the article by Zhixiong Huang and Yaohui Ying in this issue of the *Review*; and see Ma Xinmin, at the time deputy director-general of the Department of Treaty and Law, Ministry of Foreign Affairs of the People's Republic of China, writing in a personal capacity: "[T]he scope of applicability of the rules of IHL has been expanded. … [I]t has also been broadened to cyberspace. The UN GGE on Developments in the Field of Information and Telecommunications in the Context of International Security confirmed in its 2013 and 2015 reports that international law, particularly the UN Charter, is applicable in cyberspace. IHL should, therefore, in principle be applicable to cyber attacks, but how to apply it is still open to discussion" (unofficial and informal translation). Ma Xinmin, "International Humanitarian Law in Flux: Development and New Agendas – In Commemoration of the 40th Anniversary of the 1977 Adoption Protocols to the Geneva Conventions", *Chinese Review of International Law*, Vol. 30, No. 4, 2017, p. 8.

48 UN General Assembly, "Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security: Note by the Secretary-General", UN Doc. A/68/98, 24 June 2013, para. 19, and UN Doc. A/70/174, 22 July 2015, para. 24.

49 UNGA Res. 70/237, "Developments in the Field of Information and Telecommunications in the Context of International Security", UN Doc. A/RES/70/237, 30 December 2015, preambular para. 16.

50 UNGA Res. 73/27, above note 15, preambular para. 17; UNGA Res. 73/266, above note 15, preambular para. 12.

51 UN Doc. A/70/174, above note 48, para. 28(d).

52 Michael N. Schmitt, "France Speaks Out on IHL and Cyber Operations: Part I", *EJIL: Talk!*, 30 September 2019, available at: www.ejiltalk.org/france-speaks-out-on-ihl-and-cyber-operations-part-i/.

53 EU Council Conclusions, above note 20.

54 Wales Summit Declaration, above note 21, para. 72.

55 See "Cybersecurity: Paris Call of 12 November 2018 for Trust and Security in Cyberspace", *France Diplomacy*, available at: www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/france-and-cyber-security/article/cybersecurity-paris-call-of-12-november-2018-for-trust-and-security-in.

56 Commonwealth Cyber Declaration, above note 19, p. 4, para. 4.

the OAS Juridical Committee have "reveal[ed] support for the applicability of IHL" in cyberspace.[57]

At the same time, in the context of discussions on the applicability of IHL to cyber operations during armed conflict, a number of States have expressed opposition to the militarization of cyberspace, or a cyber arms race. States have also expressed concerns regarding a possible legitimization of the use of military cyber operations,[58] called for prudence in the discussion of the applicability of IHL,[59] and noted that IHL "must apply taking note of the peculiarities of cyber warfare".[60] These are important considerations, but they should not be understood as being incompatible with the application of IHL to cyber operations during armed conflict.

In our view, however, asserting that IHL applies to cyber operations during armed conflict is not an encouragement to militarize cyberspace and should not, in

57 See OAS, above note 22, para. 43 (mentioning Bolivia, Chile, Guyana, Peru and the United States); Ecuador's response may appear to have implied such support (see also paras 19–21, 25). Other member States of the OAS expressed this position in the context of the OEWG. See comments by Brazil, Colombia and Uruguay on the initial pre-draft of the OEWG report, available at: www.un.org/disarmament/open-ended-working-group/. See, however, the views of Cuba, Nicaragua and Venezuela, who note, among other things, that there is not yet consensus on the applicability of IHL in cyberspace and that direct reference to IHL in the report may validate or legitimize the militarization of cyberspace.

58 See, most recently, the submissions of China, Cuba, Iran, Nicaragua, Russia and others on the initial pre-draft of the OEWG report, available at: www.un.org/disarmament/open-ended-working-group/. See also, for example, People's Republic of China, *Position Paper of the People's Republic of China for the 73rd Session of the United Nations General Assembly*, 2018, p. 10, available at: https://tinyurl.com/y4qquywp; "Declaration by Miguel Rodríguez, Representative of Cuba, at the Final Session of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security", 23 June 2017, p. 2; Ministry of Foreign Affairs of the Russian Federation, "Response of the Special Representative of the President of the Russian Federation for International Cooperation on Information Security Andrey Krutskikh to TASS' Question Concerning the State of International Dialogue in This Sphere", 29 June 2017.

59 "The applicability of the law of armed conflicts and *jus ad bellum* needs to be handled with prudence. The lawfulness of cyber war should not be recognized under any circumstance. States should not turn cyberspace into a new battlefield": "China's Submissions to the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security", September 2019, p. 6, available at: https://s3.amazonaws.com/unoda-web/wp-content/uploads/2019/09/china-submissions-oewg-en.pdf. "We should be extremely cautious against any attempt to introduce use of force in any form into cyberspace, have sober assessment on possible conflicts and confrontations resulted from the indiscriminate application of the law of armed conflicts in cyberspace, and refrain from sending wrong messages to the world": "China's Contribution to the Initial Pre-Draft of OEWG Report", April 2020, p. 5, available at: https://front.un-arm.org/wp-content/uploads/2020/04/china-contribution-to-oewg-pre-draft-report-final.pdf. "[W]ithout state practice, we should be very prudent on the discussion of application of humanitarian law in so called 'cyber wars.' The reason is very simple but fundamental: firstly, no cyber wars shall be permitted; and secondly, cyber war will be a totally new form of high-tech war": China statement at AALCO 58th Annual Session, in AALCO, *Verbatim Record of Discussions: Fifty-Eighth Annual Session*, Doc No. AALCO/58/DAR ES SALAAM/2019/VR, 2019, p. 176, available at: www.aalco.int/Verbatim%20(FINAL)%2020200311.pdf.

60 China has stated at a meeting of the AALCO Working Group on International Law in Cyberspace that "the regimes of *jus ad bellum* and *jus in bello* must apply taking note of the peculiarities of cyber warfare". AALCO, *Summary Report of the Fourth Meeting of the Open-Ended Working Group on International Law in Cyberspace*, 3 September 2019, available at: www.aalco.int/Summary%20Report%20as%20Adopted.pdf.

any way, be understood as legitimizing cyber warfare.[61] Any resort to force by States, whether cyber or kinetic in nature, always remains governed by the UN Charter and customary international law, in particular the prohibition against the use of force.[62] International disputes must be settled by peaceful means. This principle applies in cyberspace as in all other domains. In addition to – and independent of – the requirements of the UN Charter, IHL provides limits on the conduct of hostilities if and when States or non-State parties chose to resort to cyber operations during armed conflict. In particular, IHL protects civilians and civilian objects from the effects of hostilities by restricting the belligerents' choice of means and methods of warfare, independent of whether or not the use of force was lawful. This means that instead of legitimizing cyber operations (or any other military operation) during armed conflict, IHL – the *jus in bello* – provides limits in addition to those found in the UN Charter and customary international law – the *jus ad bellum*. Furthermore, IHL actually imposes some limits on the militarization of cyberspace. For example, it prohibits the development of cyber capabilities that would qualify as weapons and would be indiscriminate by nature or would be of a nature to cause superfluous injury or unnecessary suffering.[63]

If it is accepted that IHL applies to cyber operations during armed conflicts generally, the subsequent question is whether all or only some rules of IHL apply. In this respect, the scope of application of IHL rules regulating means and methods of warfare may be broadly divided into rules that apply to all weapons, means and methods of warfare, wherever they are used (such as the principles of distinction, proportionality and precaution), and rules that are specific to certain weapons (such as weapons treaties) or certain domains (such as rules specifically regulating naval warfare). All the main customary principles and rules regulating the conduct of hostilities belong to the first category and do apply to cyber operations during armed conflict.[64] In contrast, a more detailed analysis will be required regarding the applicability of IHL rules that are specific to certain weapons or certain domains.

The fact that IHL applies does not prevent States from developing international law further, agreeing on voluntary norms, or working towards common interpretations of existing rules. For instance, when it established a UN OEWG in 2018, a majority of States in the UN General Assembly "welcome[d]" a set of "international rules, norms and principles of responsible behaviour of States" that build on the norms that were developed over the years by the UN GGEs.[65] Another example of possible new rules in the field of information

61 This view has also been expressed in, among others, the submissions of Australia, Brazil, Chile, Denmark and the United Kingdom on the initial pre-draft of the OEWG report, available at: www.un.org/disarmament/open-ended-working-group/.

62 UN Charter, Art. 2(4).

63 See Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, Vol. 1: *Rules*, Cambridge University Press, Cambridge, 2005 (ICRC Customary Law Study), Rules 70, 71, available at: https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul.

64 The principles and rules regulating the conduct of hostilities are highlighted further below, under the section entitled "The Limits that IHL Imposes on the Use of Cyber Capabilities during Armed Conflicts".

65 UNGA Res. 73/27, above note 15.

security is included in the International Code of Conduct for Information Security, submitted in 2011 to the UN by the member States of the SCO. Under the Code, States would pledge, *inter alia*, "not to proliferate information weapons and related technologies".[66] There are also academic suggestions, including with regard to further legal or policy restrictions on cyber operations during armed conflicts.[67]

To sum up, while cogent legal reasons and increasing international support exist for the conclusion that IHL applies to cyber operations during armed conflict, the issue does not enjoy universal agreement yet. As has been shown in this section, however, a careful examination of the various arguments raised in multilateral discussions shows that affirming IHL applicability does not legitimize either the militarization of cyberspace or the use of malicious cyber operations. Moreover, it does not preclude the development of possible new rules but rather provides a fundamental legal framework that possible new rules could – and should – build on.

Can cyber operations alone cross "the threshold"? Clarifying the difference between the relevant thresholds under IHL and the UN Charter

In light of the manifold cyber operations that are reported on a daily basis, it is important to recall that IHL only applies to cyber operations that form part of an armed conflict otherwise waged with traditional weapons, or, less likely, cyber operations that alone amount to an armed conflict in the absence of kinetic operations. As underlined in the previous section, the question of whether IHL applies to cyber operations during armed conflicts must be analyzed separately from that of whether there has been a violation of the rules governing the use of force under the UN Charter. In the context of the application of IHL and of the UN Charter, a key issue is the attribution of cyber operations to States. These

66 The proposed International Code of Conduct for Information Security is available at: http://nz.chineseembassy.org/eng/zgyw/t858978.htm. It was submitted by China, Russia, Tajikistan and Uzbekistan in 2011, and co-sponsored by Kazakhstan and Kyrgyzstan in 2013 (see UN Doc. A/68/98, above note 48, p. 8, para. 18). Similarly, in 2011 the Ministry of Foreign Affairs of the Russian Federation presented a draft Convention on International Information Security (22 September 2011, available at: www.mid.ru/en/foreign_policy/official_documents/-/asset_publisher/CptICkB6BZ29/content/id/191666) which lists among the "Main Measures for Averting Military Conflict in the Information Space" that States shall "take action aimed at limiting the proliferation of 'information weapons' and the technology for their creation" (Art. 6(10)). Its Article 7(2) also foresees that "[i]n any international conflict, the right of the States Parties that are involved in the conflict to choose the means of 'information warfare' is limited by applicable norms of international humanitarian law".

67 Among many others, Pascucci, for instance, has suggested that the negotiation of an Additional Protocol IV could enable some of the issues raised by the application of the principle of distinction and proportionality in cyberspace to be addressed: Peter Pascucci, "Distinction and Proportionality in Cyberwar: Virtual Problems with a Real Solution", *Minnesota Journal of International Law*, Vol. 26, No. 2, 2017. Schmitt, meanwhile, has put forward proposals in terms of policies that States could adopt: Michael N. Schmitt, "Wired Warfare 3.0: Protecting the Civilian Population during Cyber Operations", *International Review of the Red Cross*, Vol. 101, No. 910, 2019, pp 333–355.

three points – which cyber operations are governed by IHL,[68] the relationship between IHL and the UN Charter, and questions of attribution – are addressed in this section.

Cyber operations that are governed by IHL

When cyber operations are conducted in the context of – and have a nexus to – an existing international or non-international armed conflict carried out through kinetic means, relevant IHL rules apply to, and regulate the conduct of, all parties to the conflict.[69] Cyber operations alongside, and in support of, kinetic operations during armed conflicts are the only type of operations that States have acknowledged and have considered to be governed by IHL.[70]

A separate question is whether cyber operations alone – absent kinetic operations – may be regulated by IHL. In other words, can a cyber operation be the first, and possibly only, shot in an armed conflict as defined by IHL? This is to be assessed according to Articles 2 and 3 common to the four Geneva Conventions of 1949[71] for international and non-international armed conflicts respectively.[72] These two types of armed conflict differ in the nature of the parties they involve, the intensity of violence that triggers the applicability of IHL, and some of the IHL rules that apply.

With regard to international armed conflicts, common Article 2 states that "the present Convention shall apply to all cases of declared war or of any other armed conflict which may arise between two or more of the High Contracting Parties, even if the state of war is not recognized by one of them". It is today agreed that "an armed conflict exists whenever there is 'a resort to armed force between States'".[73] With regard to the question of whether there is a threshold of intensity with regard to international armed conflicts, there is some State practice,

68 For an illustration of these debates, see "Scenario 13: Cyber Operations as a Trigger of the Law of Armed Conflict", in Kubo Mačák, Tomáš Minárik and Taťána Jančárková (eds), *Cyber Law Toolkit*, available at: https://cyberlaw.ccdcoe.org/.

69 See ICRC, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field*, 2nd ed., Geneva, 2016 (ICRC Commentary on GC I), para. 254; Tallinn Manual 2.0, above note 13, Rule 80.

70 See references in note 2 above.

71 Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of 12 August 1949, 75 UNTS 31 (entered into force 21 October 1950) (GC I); Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea of 12 August 1949, 75 UNTS 85 (entered into force 21 October 1950) (GC II); Geneva Convention (III) relative to the Treatment of Prisoners of War of 12 August 1949, 75 UNTS 135 (entered into force 21 October 1950) (GC III); Geneva Convention (IV) relative to the Protection of Civilian Persons in Time of War of 12 August 1949, 75 UNTS 287 (entered into force 21 October 1950) (GC IV).

72 Common Article 2(1): "[T]he present Convention shall apply to all cases of declared war or of any other armed conflict which may arise between two or more of the High Contracting Parties, even if the state of war is not recognized by one of them." Common Article 3(1): "In the case of armed conflict not of an international character occurring in the territory of one of the High Contracting Parties …"

73 International Criminal Tribunal for the former Yugoslavia (ICTY), *The Prosecutor v. Duško Tadić*, Case No. IT-94-1, Decision on the Defence Motion for Interlocutory Appeal on Jurisdiction, 2 October 1995, para. 70; ICRC Commentary on GC I, above note 69, para. 218.

and some strong humanitarian and conceptual arguments, that IHL applies as soon as armed force is used between States, irrespective of the intensity of the violence. IHL is primarily concerned with the protection of persons affected by armed conflict. Thus, as soon as they use armed force, States must direct their attacks at military objectives and not at civilians or civilian objects, and must take constant care to spare the latter. It cannot matter whether there is one or many civilians needing protection against attack.[74] At least where the use of cyber operations between States leads to effects akin to those of more traditional means and methods of warfare, IHL applies.

Experts generally agree that cyber operations, on their own, have the potential to cross the threshold of an international armed conflict under IHL.[75] The ICRC shares this view.[76] In a rare expression of a State's position on the issue, France has stated that "[c]yberoperations that constitute hostilities between two or more States may characterise the existence of international armed conflict".[77]

The question of exactly where this threshold lies remains unsettled.[78] In the ICRC's view, there is no reason to treat one or more cyber operations resulting in the destruction of civilian or military assets, or in the death or injury of soldiers or civilians, differently from equivalent attacks conducted through more traditional means and methods of warfare. Cyber operations might, however, also disable objects without physically damaging them. It remains to be seen if and under what conditions States might consider such operations to amount to a resort to armed force as understood in IHL, and therefore to be governed by this body of law.[79]

With regard to non-international armed conflicts, situations of internal violence may amount to a non-international armed conflict if there is "protracted armed violence between governmental authorities and organized armed groups or between such groups within a State".[80] The two criteria that derive from this definition – the organization of the parties to the conflict and the intensity of the violence – pose various questions regarding cyber operations. First, while State armed forces satisfy the organization criterion, determining the degree of

74 Similarly, if the resort to armed force leads, for example, to injuries or the capture of a member of another State's armed forces, IHL rules on the protection of the wounded and sick or the status and treatment of prisoners of war are relevant whether there is one or many prisoners, one or many wounded to be cared for. See ICRC Commentary on GC I, above note 69, paras 236–244.

75 Tallinn Manual 2.0, above note 13, Rule 82, para. 16.

76 ICRC Commentary on GC I, above note 69, paras 253–256.

77 French Ministry of the Armies, *International Law Applied to Operations in Cyberspace*, 2019, p. 12, available at: www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf. This document specifies that "[w]hile an armed conflict consisting exclusively of digital activities cannot be ruled out in principle, it is based on the capacity of autonomous cyberoperations to reach the threshold of violence required to be categorised as such".

78 Tallinn Manual 2.0, above note 13, Rule 82, paras 11–16; as can be seen from paras 12–13, the question is not fully settled for kinetic operations either, and this uncertainty will permeate the debate on whether cyber operations alone can cross the threshold of an international armed conflict beyond the cyber-specific issues.

79 ICRC Commentary on GC I, above note 69, para. 255; Tallinn Manual 2.0, above note 13, Rule 82, para. 11.

80 ICTY, *Tadić*, above note 73, para. 70.

organization of an armed group is a more complicated and fact-specific assessment; it becomes all the more challenging – yet arguably not impossible – when that group is only organized online.[81] Second, unlike IHL applicable to international armed conflicts, which governs any resort to armed force between States regardless of its intensity,[82] a non-international armed conflict will only exist if violence between two or more organized parties is sufficiently intense. Again, while arguably not impossible in exceptional circumstances, it will be unlikely that cyber operations alone would meet the intensity requirement for a non-international armed conflict.[83] While expressing the view that prolonged cyber operations may in principle and depending on the circumstances constitute a non-international armed conflict, France held that the state of technology seems to rule out this possibility for the time being.[84]

It has rightly been emphasized that the law of armed conflict "does not regulate cyber operations that fall outside of an armed conflict situation".[85] Diverging views exist, however, on whether some or all of its principles should be applied, as a matter of policy, to cyber operations at all times.

The United States has recently stated that "even if the law of war does not technically apply because the proposed military cyber operation would not take place in the context of armed conflict, [the Department of Defense] nonetheless applies law-of-war principles"[86] as it does more generally with regard to all its operations.[87] In contrast, Russia has cautioned against "potentially dangerous … attempts to impose the principle of full and automatic applicability of IHL to the ICT environment in peacetime".[88]

81 ICRC Commentary on GC I, above note 69, para. 437; Tallinn Manual 2.0, above note 13, Rule 83, paras 13–15. For an in-depth analysis of the issue, see Tilman Rodenhäuser, *Organizing Rebellion: Non-State Armed Groups under International Humanitarian Law, Human Rights Law, and International Criminal Law*, Oxford University Press, Oxford, 2018, pp. 104–108.

82 ICRC Commentary on GC I, above note 69, paras 236–244.

83 *Ibid.*, para. 437. For further discussion, see Tallinn Manual 2.0, above note 13, Rule 83, paras 7–10; Cordula Droege, "Get Off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians", *International Review of the Red Cross*, Vol. 94, No. 886, 2012, p. 551; Michael N. Schmitt, "Classification of Cyber Conflict", *Journal of Conflict and Security Law*, Vol. 17, No. 2, 2012, p. 260.

84 French Ministry of the Armies, above note 77, p. 12.

85 New Zealand Defence Force, *Manual of Armed Forces Law*, Vol. 4: *Law of Armed Conflict*, 2nd ed., DM 69, 2017 (New Zealand Military Manual), para. 5.2.23, available at: www.nzdf.mil.nz/assets/Publications/DM-69-2ed-vol4.pdf.

86 Paul C. Ney Jr., US Department of Defence General Counsel, Remarks at US Cyber Command Legal Conference, 2 March 2020, available at: www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/.

87 See US Department of Defense (DoD), Directive 2311.01E, "DoD Law of War Program", 2006 (amended 2011), paras 4–4.1: "It is DoD policy that … [m]embers of the DoD Components comply with the law of war during all armed conflicts, however such conflicts are characterized, *and in all other military operations*" (emphasis added). See also US Department of Defense (DoD), *Law of War Manual*, 2015 (DoD Law of War Manual), para. 3.1.1.2, available at: https://tinyurl.com/y6f7chxo.

88 Russia, "Commentary of the Russian Federation on the Initial 'Pre-draft' of the Final Report of the United Nations Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security", April 2020, available at: https://front.un-arm.org/wp-content/uploads/2020/04/russian-commentary-on-oweg-zero-draft-report-eng.pdf.

While policy debates on this question are likely to continue, from a legal point of view it is undisputed that IHL does not apply outside the context of an armed conflict. It is true that some rules of IHL, such as the protection of persons not or no longer taking part in hostilities enshrined in common Article 3, or the strong protection of health-care facilities or objects indispensable to the survival of the civilian population, could have positive effects if applied at all times. In contrast, it may be more problematic to apply other IHL rules outside armed conflict, notably those derived from the principles of distinction and proportionality. These rules are based on the premise that attacks against military objectives are lawful under IHL during armed conflict. However, outside armed conflict, the notion of "military objectives" that may lawfully be attacked does not exist – even attacks against another State's military are prohibited. While the principle of proportionality also exists outside of armed conflict, it has a distinct meaning under other bodies of law and therefore operates differently during and outside armed conflicts.[89] Outside armed conflict, disputes among States and the use of force are solely regulated by other fields of international law, such as the UN Charter and human rights law, as applicable.

The relationship between IHL and the UN Charter

A State that considers carrying out a cyber operation against another State must analyze the lawfulness of such an operation under the *jus ad bellum* framework (as found in the UN Charter and customary international law) and the *jus in bello* framework (IHL). The UN Charter and IHL are complementary when it comes to the protection of humans from war and its effects, even though they are distinct fields of international law. Their objectives are complementary: while the preamble of the UN Charter states that it aims to "save succeeding generations from the scourge of war", the preamble of AP I states that the objective of IHL is "protecting the victims of armed conflict". Concretely, the UN Charter prohibits the use of force other than in self-defence or when authorized by the Security Council. The applicability of IHL does not replace or set aside the essential rules of the UN Charter, but if an armed conflict breaks out, IHL defines protections for those who do not (civilians) or no longer (for example, wounded soldiers or detainees) participate in hostilities and limits the belligerents' choice in the means and methods of warfare. Thus, while the UN Charter sets out – subject to narrow exceptions – a prohibition against the use of force, IHL imposes limits on how hostilities may be conducted once a conflict breaks out.

At the same time, IHL and the UN Charter are different fields of international law, each having its own concepts and terminology. As both are concerned with regulating the use of force, some of the terminology they use is similar and at times confusing. This is the case, for example, as regards the notion of "resort to armed force between States" to classify a conflict under IHL, and the prohibition against "the threat or use of force" and the right to self-

89 For a brief assessment, see ICRC Challenges Report 2019, above note 36, pp. 18–22.

defence against an "armed attack" under the UN Charter. While international law treaties do not define these notions – neither in general nor as regards cyberspace – certain basic elements can be distilled from jurisprudence and commentary.

As discussed above, IHL applies as soon as there is a resort to armed force between States, irrespective of the intensity of the violence.

The UN Charter does not define the term "use of force" under Article 2(4), and the question of what type of force may qualify remains subject to debate. Following the provision's drafting history and subsequent State practice, it may be concluded that the use of political or economic coercion is not included in this notion.[90] Instead, it has been argued that the prohibition against the use of force under the UN Charter is "limited to armed force".[91] Importantly with regard to cyber operations, the ICJ has stated that Article 2(4) prohibits "any use of force, regardless of the weapons employed".[92] Based on this finding, some States have emphasized that "crossing the threshold of the use of force depends not on the digital means employed but on the effects of the cyberoperation", and have concluded accordingly that a "cyberoperation carried out by one State against another State violates the prohibition of the use of force if its effects are similar to those that result from the use of conventional weapons".[93] A number of examples given by States of the use of force in cyberspace seem to reflect this understanding, such as cyber operations causing injury or death of persons or damage to or destruction of property;[94] triggering a nuclear plant meltdown; opening a dam above a populated area, causing destruction; disabling air traffic control services, resulting in airplane crashes; and crippling a military's logistics systems.[95] Some States seem to interpret the prohibition against the use of force even more broadly, stating that it cannot be ruled out that "a cyberoperation without physical effects may also be characterised as a use of force",[96] or that "a cyber operation with a very serious financial or economic impact may qualify as the use of force".[97]

90 See Oliver Dörr and Albrecht Randelzhofer, "Article 2(4)", in Bruno Simma *et al.* (eds), *The Charter of the United Nations: A Commentary*, Vol. 1, Oxford University Press, Oxford, 2016, paras 17–20 of the commentary on Art. 2(4). Accordingly, experts have concluded that "neither non-destructive cyber psychological operations intended solely to undermine confidence in a government, nor a State's prohibition of e-commerce with another State designed to cause negative economic consequences, qualify as uses of force": Tallinn Manual 2.0, above note 13, para. 3 of the commentary on Rule 69.

91 O. Dörr and A. Randelzhofer, above note 90, p. 208, para. 16.

92 ICJ, above note 46, para. 39.

93 French Ministry of the Armies, above note 77, p. 7. See also Tallinn Manual 2.0, above note 13, para. 1 of the commentary on Rule 69.

94 See Estonia, "President of the Republic at the Opening of CyCon 2019", 29 May 2019, available at: www.president.ee/en/official-duties/speeches/15241-president-of-the-republic-at-the-opening-of-cycon-2019/index.html. Australian Department of Foreign Affairs and Trade, "Australia's International Cyber Engagement Strategy", 2019, available at: www.dfat.gov.au/international-relations/themes/cyber-affairs/Pages/australias-international-cyber-engagement-strategy.

95 DoD Law of War Manual, above note 87, para. 16.3.1.

96 French Ministry of the Armies, above note 77, p. 7. Examples that France provides of actions that could "be deemed uses of force" are "penetrating military systems in order to compromise French defence capabilities, or financing or even training individuals to carry out cyberattacks against France".

97 Dutch Ministry of Foreign Affairs, "Letter to the Parliament on the International Legal Order in Cyberspace", 5 July 2019, p. 4, available at: www.government.nl/ministries/ministry-of-foreign-affairs/

Turning to the right to self-defence under the UN Charter and customary international law, this right may only be exercised against an "armed attack". Following the ICJ's finding that only "the most grave forms of the use of force" may qualify as armed attacks and that such attacks must reach certain "scale and effects",[98] it may be concluded that a use of force must reach a certain intensity to qualify as an "armed attack".[99] Again, experts have held that "some cyber operations may be sufficiently grave to warrant classifying them as an 'armed attack' within the meaning of the Charter",[100] notably those whose effects are comparable to more traditional armed attacks. This view is also reflected in the public positions of some States.[101]

The questions of how the thresholds of a resort to armed force to which IHL applies, the prohibition of the use of force under the UN Charter, and the notion of "armed attacks" giving rise to the inherent right to self-defence are interpreted in cyberspace are evolving. While certain signposts may be identified based on the jurisprudence of the ICJ, the case law of international criminal tribunals and courts, State practice, and expert views, many issues remain blurred for the moment.

Nonetheless, it is important to emphasize that these three notions and concepts stem from different bodies of international law and have different meanings. As noted above, in the ICRC's view, a cyber operation that amounts to a resort to armed force between States under IHL is governed by that body of law even in the absence of a pre-existing armed conflict. In practice, such an operation may also amount to a prohibited use of force under the UN Charter. However, the two conclusions require a separate legal analysis: concluding that the threshold has been reached under one body of law does not necessarily preclude reaching a different conclusion under the other body of law. This is especially important when differentiating the applicability of IHL from the right to self-defence under the UN Charter. In view of the position that only the gravest forms of the use of force – meaning those that reach a certain scale and effects – may qualify as armed attacks, it is clear that not every resort to armed force to which IHL applies amounts to an armed attack under the UN Charter triggering the right to self-defence.[102] These differences have significant legal and practical consequences. Therefore, any analysis of a situation in which a State uses cyber operations against another State needs to distinguish the various notions and not merge them into one unspecified "threshold".

documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace; French Ministry of the Armies, above note 77, p. 7. For a recent overview of States' positions, see Przemysław Roguski, *Application of International Law to Cyber Operations: A Comparative Analysis of States' Views*, Policy Brief, Hague Program for Cyber Norms, 2020. For an illustration of these debates, see, for example, Kenneth Kraszewski, "Scenario 14: Ransomware Campaign", in K. Mačák, T. Minárik and T. Jančárková (eds), above note 68, paras L5–L13.

98 ICJ, *Case Concerning Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Judgment, 27 June 1986, paras 191, 195.

99 This view is not, however, accepted by all States. For instance, the United States considers that any use of force is an armed attack.

100 Tallinn Manual 2.0, above note 13, para. 4 of the commentary on Rule 71.

101 Dutch Ministry of Foreign Affairs, above note 97, p. 4; French Ministry of the Armies, above note 77, p. 7.

102 H. Durham, above note 25.

The issue of attribution

In warfare generally – and in cyberspace in particular – States will at times use non-State actors, such as non-State armed groups or private military and security companies, to carry out certain acts, including cyber operations. The specific characteristics of cyberspace, such as the variety of possibilities for actors to hide or falsify their identity, complicate the attribution of conduct to specific individuals, and to parties to armed conflicts.[103] This raises important challenges when determining the applicability of IHL in a particular situation. If the perpetrator of a given operation – and thus the link between the operation and an armed conflict – cannot be identified, it is extremely difficult to determine whether IHL is even applicable to the operation.[104] First, as discussed above, different thresholds of violence are relevant to qualify State or non-State cyber attacks as an armed conflict. Thus, if the State or non-State origin of a cyber operation outside an ongoing armed conflict is not known, it is unclear which threshold applies. Second, even when an armed conflict is taking place, cyber attacks that have no nexus to the conflict (such as criminal acts unrelated to the conflict) are not regulated by IHL, and the inability to identify the author of a cyber operation might hamper the determination of whether such a nexus to the conflict exists. These examples show that determining who the author of a cyber operation is, and whether the operation can be attributed to a State or non-State party to the conflict, has important legal consequences.

Attribution of cyber operations is also important to ensure that actors who violate international law, including IHL, can be held accountable. The perception that it will be easier to deny responsibility for unlawful attacks may also weaken the taboo against such uses – and may make actors less scrupulous about conducting operations in violation of international law.[105]

This being said, attribution is not a problem from the perspective of the actors who conduct, direct or control cyber operations: they have all the facts at hand to determine under which international legal framework they are operating and which obligations they must respect.[106]

Under international law, a State is responsible for conduct attributable to it, including possible violations of IHL. This includes:

(a) violations committed by its organs, including its armed forces;
(b) violations committed by persons or entities it empowered to exercise elements of governmental authority;

103 For an examination of the technical challenges for attributing cyber attacks to specific actors, see Vitaly Kamluk, "Know Your Enemy and Know Yourself: Attribution in the Cyber Domain", *Humanitarian Law and Policy Blog*, 3 June 2019, available at: https://blogs.icrc.org/law-and-policy/2019/06/03/know-your-enemy-know-yourself-cyber-domain-attribution/.
104 ICRC Challenges Report 2011, above note 43, p. 36.
105 ICRC, above note 1, p. 9.
106 *Ibid.*

(c) violations committed by persons or groups acting in fact on its instructions, or under its direction or control; and
(d) violations committed by private persons or groups which it acknowledges and adopts as its own conduct.[107]

These principles apply whether the violation of IHL has been committed by cyber means or by any other means.[108]

The limits that IHL imposes on the use of cyber capabilities during armed conflicts

Recognizing that IHL applies to cyber operations having a nexus to an armed conflict is only a first step. The specific characteristics of this new technology raise several challenges for the interpretation of IHL rules, including those on the conduct of hostilities.

The partly non-physical (i.e., digital) nature of cyberspace and the interconnectedness of military and civilian networks pose practical and legal challenges in applying the general IHL rules protecting civilians and civilian objects against cyber operations, in particular those amounting to attacks under IHL. It is even suggested that it may be impossible, at times, to apply basic IHL principles in cyberspace. As will be shown below, this challenge might be overstated. Nonetheless, key issues arise with regard to protecting essential civilian cyber infrastructure against military attack. As many of the IHL rules governing the conduct of hostilities apply only to military operations that amount to "attacks" as defined in IHL, this section first examines various issues relating to cyber operations that qualify as attacks, including the salient question of *which* operations qualify as attacks under IHL. Second, it explores obligations of parties to armed conflicts in military operations other than those amounting to "attacks". Third, the section analyzes certain challenges regarding the legal review of cyber capabilities.

Cyber operations that amount to an attack under IHL

IHL sets out essential rules restricting cyber operations that amount to "attacks" as defined in IHL. This section looks at those rules and principles that have been subject to the most intense debates. It examines first whether, from a technological perspective, cyber attacks are capable of being directed at specific

107 See ICRC Customary Law Study, above note 63, Rule 149. See also International Law Commission, *Responsibility of States for Internationally Wrongful Acts*, 2001, in particular Arts 4–11.

108 ICRC, above note 1, p. 9; Tallinn Manual 2.0, above note 13, Rules 15–17. For a different view, see the Chinese submission on the initial pre-draft of the report of the OEWG, which states that with regard to "state responsibility, which, unlike the law of armed conflicts or human rights, has not yet gained international consensus, there is no legal basis at all for any discussion on its application in cyberspace". Comments by China on the initial pre-draft of the OEWG report, available at: www.un.org/disarmament/open-ended-working-group/.

military objectives as required by the principle of distinction. The second part analyzes how the notion of attack under IHL should be interpreted in cyberspace. The third part discusses the closely related debate on whether civilian data must be granted the same protection as civilian "objects" for the purposes of IHL, and the final part addresses the ongoing debate on how IHL rules on the conduct of hostilities apply to objects used simultaneously for civilian and military purposes (often called dual-use objects), which are particularly prevalent in cyberspace.

From a technical perspective, cyber attacks can be directed at specific military objectives

Implementing the principles of distinction and proportionality, and the prohibition against indiscriminate attacks, requires that an attack can be and is directed at a military objective and will not cause excessive incidental harm to civilians or civilian objects. Contrary to the assumption that these principles might become meaningless in cyberspace because of the interconnectivity that characterizes it, careful examination of cyber operations shows that such operations are not inherently indiscriminate. For instance, if a cyber operation is carried out by operators who enter a target and carry out an operation, the operators will know where they are and what they are doing. Similarly, analyses of cyber tools show that they are not necessarily indiscriminate. However, programming malware that discriminates between civilian objects and military objectives, and conducting cyber operations without causing excessive incidental damage, requires sophisticated capabilities and testing.

Those who develop malware or plan cyber attacks can design their tools without self-propagation functions. In that case, malware cannot spread without additional human intervention. Even if self-propagating, attacks over the years have shown that malware can be designed to only attack specific hardware or software. This means that even if malware is programmed to spread widely, it can be designed to only cause damage to a specific target or specific sets of targets. Especially cyber attacks that aim to cause physical damage to industrial control systems may require cyber tools that are designed for that specific target and purpose. In many cases, the need for such custom-made tools would effectively hamper – from a technical perspective – the ability to carry out a cyber attack on a large scale or in an indiscriminate manner. The fact that cyber attacks can technically be targeted precisely does not mean they are necessarily lawful if carried out in a conflict. However, the characteristics seen in a number of cyber operations show that they can be very precisely tailored to create effects on specific targets only, and that such operations are therefore capable of complying with IHL principles and rules.

The fact that some of the known cyber tools were designed to self-propagate and caused harmful effects on widely used civilian computer systems does not support an argument that the interconnectedness of cyberspace makes it challenging, if not impossible, to implement basic IHL rules. On the contrary, during armed conflicts, the use of such cyber tools would be prohibited by

IHL.[109] IHL prohibits attacks that employ means and methods of warfare, including cyber means and methods, that cannot be directed at a specific military objective or may be expected to escape the control of the user,[110] or – while being targeted at a military objective – may be expected to cause incidental civilian damage that is excessive in relation to the concrete and direct military advantage anticipated.[111]

The notion of "attack" under IHL and its application to cyber operations

The question of whether or not an operation amounts to an "attack" as defined in IHL is essential for the application of many of the rules deriving from the principles of distinction, proportionality and precaution, which afford important protection to civilians and civilian objects. Concretely, rules such as the prohibition on *attacks* against civilians and civilian objects,[112] the prohibition on indiscriminate[113] and disproportionate *attacks*,[114] and the obligation to take all feasible precautions to avoid or at least reduce incidental harm to civilians and damage to civilian objects when carrying out an *attack*[115] apply to those operations that qualify as "attacks" as defined in IHL. The question of how widely or narrowly the notion of "attack" is interpreted with regard to cyber operations is therefore essential for the applicability of key rules – and the protection they afford to civilians and civilian infrastructure – to cyber operations.

Article 49 of AP I defines attacks as "acts of violence against the adversary, whether in offence or in defence". It is well established that the notion of violence in this definition can refer to either the means of warfare or their effects, meaning that an operation causing violent effects can be an attack even if the means used to cause those effects are not violent as such.[116] On the basis of this understanding, the Tallinn Manual 2.0 proposes the following definition of a cyber attack: "A cyber attack is a cyber operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to objects."[117]

It is widely accepted by States that took a position on the issue, by the ICRC and by experts that at least those cyber operations which cause death, injury or

109 Similarly, the DoD Law of War Manual, above note 87, para. 16.6, concludes: "For example, a destructive computer virus that was programmed to spread and destroy uncontrollably within civilian internet systems would be prohibited as an inherently indiscriminate weapon."

110 Yves Sandoz, Christophe Swinarski and Bruno Zimmerman (eds.), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949*, ICRC, Geneva, 1987 (ICRC Commentary on the APs), para. 1963.

111 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) (AP I), Art. 51(4)–(5); ICRC Customary Law Study, above note 63, Rules 11, 14.

112 See AP I, Art. 52; ICRC Customary Law Study, above note 63, Rules 7–10.

113 See AP I, Art. 54(c); ICRC Customary Law Study, above note 63, Rule 11.

114 See AP I, Art. 51(5)(b); ICRC Customary Law Study, above note 63, Rule 14.

115 See AP I, Art. 57(1); ICRC Customary Law Study, above note 63, Rule 15.

116 See C. Droege, above note 83, p. 557; William H. Boothby, *The Law of Targeting*, Oxford University Press, Oxford, 2012, p. 384. As Droege points out, "it is uncontroversial that the use of biological, chemical, or radiological agents would constitute an attack, even though the attack does not involve physical force".

117 Tallinn Manual 2.0, above note 13, Rule 92.

physical damage constitute attacks under IHL.[118] Some States expressly include harm due to the foreseeable indirect (or reverberating) effects of attacks,[119] a view that is also taken by the ICRC.[120] This could be the case, for example, if patients in an intensive care unit are killed as a result of a cyber operation against an electricity network that causes the hospital's electricity supply to be cut off.

Beyond this basic consensus, different views exist on whether a cyber operation that disables an object without physically damaging it amounts to an attack under IHL.[121] There were extensive discussions on this issue in the process of drafting the Tallinn Manual. A majority of the experts held that a cyber operation amounts to an attack if it is expected to interfere with functionality and if restoration of functionality requires replacement of physical components. For some experts, a cyber operation will also amount to an attack if the restoration of functionality requires the reinstallation of the operating system or of particular data.

The ICRC has taken the position that an operation designed to disable a computer or a computer network during an armed conflict constitutes an attack as defined in IHL whether or not the object is disabled through destruction or in any other way.[122]

Two main reasons underpin the ICRC's position. The first one results from an interpretation of the notion of attack in its context.[123] Given that the definition of military objectives in Article 52(2) of AP I refers not only to destruction or capture but also to "neutralization" as a possible result of an attack, the notion of "attack" under Article 49 of AP I should be understood to encompass operations designed to impair the functioning of objects (i.e. neutralize them) without causing physical damage or destruction. Indeed, it has been submitted that the explicit mentioning of neutralization under Article 52(2) would be superfluous otherwise.[124] Second,

118 See ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, Geneva, 2015 (ICRC Challenges Report 2015), pp. 41–42, available at: www.icrc.org/en/download/file/15061/32ic-report-on-ihl-and-challenges-of-armed-conflicts.pdf; Tallinn Manual 2.0, above note 13, Rule 92. For States that have taken a view on how the notion of attack under IHL applies to cyber operations, see, in particular, Australian Department of Foreign Affairs and Trade, above note 94, Annex A; Danish Ministry of Defence, *Military Manual on International Law Relevant to Danish Armed Forces in International Operations*, 2016 (Danish Military Manual), pp. 290–291, available at: www2.forsvaret.dk/omos/publikationer/Documents/Military%20Manual%20updated%202020.pdf; French Ministry of the Armies, above note 77, p. 13; Norway, *Manual i krigens folkerett*, 2013 (Norwegian Military Manual), para. 9.54, available at: https://fhs.brage.unit.no/fhs-xmlui/bitstream/handle/11250/194213/manual_krigens_folkerett.pdf?sequence=1&isAllowed=y; New Zealand Military Manual, above note 85, para 8.10.17; DoD Law of War Manual, above note 87, para. 16.5.1.

119 Danish Military Manual, above note 118, p. 677 (when discussing computer network attacks); New Zealand Military Manual, above note 85, para 8.10.22; Norwegian Military Manual, above note 118, para. 9.54.

120 ICRC, above note 1, p. 7.

121 See, for instance, Tallinn Manual 2.0, above note 13, commentary on Rule 92, paras 10–12.

122 See ICRC Challenges Report 2015, above note 119, pp. 41–42. See also Tallinn Manual 2.0, above note 13, para. 12 of the commentary on Rule 92.

123 Vienna Convention on the Law of Treaties, Art. 31(1).

124 Knut Dörmann, "Applicability of the Additional Protocols to Computer Network Attacks", 2004, p. 4, available at: www.icrc.org/en/doc/assets/files/other/applicabilityofihltocna.pdf; C. Droege, above note 83, p. 559. For a different view, see Michael N. Schmitt, "Cyber Operations and the *Jus in Bello*: Key Issues", *International Law Studies*, Vol. 87, 2011, pp. 95–96; Heather Harrison Dinniss, *Cyber Warfare and the Laws of War*, Cambridge University Press, Cambridge, 2012, p. 198.

an overly restrictive understanding of the notion of attack would be difficult to reconcile with the object and purpose of the rules on the conduct of hostilities, which are to ensure the protection of the civilian population and civilian objects against the effects of hostilities. Indeed, under an overly restrictive understanding, a cyber operation that is directed at making a civilian network (electricity, banking, communications or other network) dysfunctional, or risks causing this incidentally, might not be covered by essential IHL rules protecting the civilian population and objects.[125]

In a similar manner, expert commentators suggest that it is important "to interpret the provision [Article 49 of AP I] taking into account the recent technological developments and to expand the concept of 'violence' to include not only material damage to objects, but also incapacitation of infrastructures without destruction".[126]

Because cyber operations can significantly disrupt essential services without necessarily causing physical damage, this constitutes one of the most critical debates for the protection of civilians against the effects of cyber operations. It is therefore crucial for States to express their views on the issue and to work towards a common understanding. For the moment, opinions vary among the States that have taken public positions.

The definitions of the notion of "attack" adopted in the military manuals of Norway and New Zealand mirror the definition adopted by the Tallinn Manual 2.0. It is, however, unclear whether these manuals were meant to express a position on this debate, because the commentary on Rule 92 of the Tallinn Manual 2.0 notes different views of how "damage" should be understood in the cyber context. Australia has stated that cyber operations qualify as attacks if they rise "to the same threshold as that of a kinetic 'attack under IHL'",[127] but it is unclear whether this was intended as a position in this debate.

A few States focus on physical damage to qualify a cyber operation as an attack. According to one OAS study, Peru has opined that in order for an operation to qualify as an attack, people or objects must be "physically harmed".[128] The Danish Military Manual specifies for the term "attack" that "[a]s far as damage to objects is concerned, the term covers any physical damage. However, the term does not cover temporary inoperability and other neutralization which does not involve physical damage (e.g., a digital 'freeze' of a

125 In the same sense, see also M. N. Schmitt, above note 67, p. 339.

126 Marco Roscini, *Cyber Operations and the Use of Force in International Law*, Oxford University Press, Oxford, 2014, p. 181. See also Dieter Fleck, "Searching for International Rules Applicable to Cyber Warfare – A Critical First Assessment of the New *Tallinn Manual*", *Journal of Conflict and Security Law*, Vol. 18, No. 2, 2013, p. 341: "It would, indeed, be less than convincing to insist that the term 'attacks' should be limited to acts directly causing injury or physical destruction, when the same action can, eg lead to disrupt [*sic*] essential supplies for hospitals or other important civilian infrastructure."

127 Australian Department of Foreign Affairs and Trade, above note 94, Annex A.

128 OAS, above note 22, para. 43.

communication control system)."[129] In its 2014 submission to the UN GGE, the United States noted:

> When determining whether a cyber activity constitutes an "attack" for *jus in bello* purposes, States should consider, *inter alia*, whether a cyber activity results in kinetic and irreversible effects on civilians, civilian objects, or civilian cyber infrastructure, or non-kinetic and reversible effects on the same.[130]

Along the same lines, the US Department of Defense (DoD) Law of War Manual provides the example of a "cyber attack that would destroy enemy computer systems", and notes that "[f]actors that would suggest that a cyber operation is not an 'attack' include whether the operation causes only reversible effects or only temporary effects".[131] Unfortunately, these documents do not clarify what they mean by "reversible" or "temporary" effects, or what the difference is – if any – between the two notions.[132] They do not discuss whether – and if so, after how long – an effect may no longer be considered temporary, or how to consider repeated operations that would each cause a temporary – but deliberately accumulating – effect. They do not discuss either whether "reversible" refers solely to operations in which the author may reverse the effects of the attack,[133] or also to operations where the target needs to take action to restore the functionality of the targeted system or otherwise end or reverse the effects of the attack. In this respect, it should be remembered that the possibility of repairing physical damage caused by a military operation (whether cyber or kinetic) is not generally understood as a criterion disqualifying an operation as an attack under IHL.[134] This is the case even if the repair reverses the direct effect of that operation and restores the functionality of the object in question.[135]

France has expressed a clearer and broader understanding of the notion of cyber attack. It considers that

129 Danish Military Manual, above note 118, p. 290. The Manual specifies with regard to computer network attacks and operations that "[t]his means, for instance, that network-based operations must be regarded as attacks under IHL if the consequence is that they cause physical damage". *Ibid.*, p. 291.

130 United States Submission to the UN Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, 2014–15, p. 5.

131 See also DoD Law of War Manual, above note 87, paras 16.5.1, 16.5.2.

132 Gary Brown and Kurt Sanger, "Cyberspace and the Law of War", *Cyber Defense Review*, 6 November 2015, available at: https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/1136032/cyberspace-and-the-law-of-war/.

133 For example, a distributed denial-of-service (DDoS) attack where the targeted network or system would automatically get back to operating normally when the attacker ends the DDoS attack and where no other indirect effect would have been caused during the time that the network or system was affected.

134 Laurent Gisel, "The Use of Cyber Technology in Warfare: Which Protection Does IHL afford and Is It Sufficient?", in G. Venturini and G. L. Beruto (eds), above note 7.

135 For example, Michael Lewis discusses the practice of conducting bridge attacks longitudinally during the 1991 Gulf War, and, *inter alia*, notes that "damage to the bridge would be nearer midspan and therefore more easily repaired", without claiming that this quality would prevent the operation to qualify as an attack. See Michael Lewis, "The Law of Aerial Bombardment in the 1991 Gulf War", *American Journal of International Law*, Vol. 97, No. 3, 2003, p. 501.

> a cyberoperation is an attack where the targeted equipment or systems no longer provide the service for which they were implemented, whether temporarily or permanently, reversibly or not. If the effects are temporary and/or reversible, the attack is characterised where action by the adversary is necessary to restore the infrastructure or system (repair of equipment, replacement of a part, reinstallation of a network, etc.).[136]

Commenting this position, Schmitt has noted that "[t]his view is highly defensible as a matter of law, for the plain meaning of damage reasonably extends to systems that do not operate as intended and require some form of repair to regain functionality".[137] In a similar manner, according to the OAS study mentioned above, Chile suggests that for an operation to qualify as an attack, its result must require the affected State to "take action to repair or restore the affected infrastructure or computer systems, since in those cases the consequences of the attack are similar to those described above, in particular physical damage to property".[138] The study also indicated that Guatemala expressed the position that a cyber operation which "only produce[s] a loss of functionality" would amount to an attack, a position also held by Ecuador.[139] Bolivia, Ecuador and Guyana further specify that such cyber operations may constitute an attack under IHL in particular when they disable critical infrastructure or the provision of basic services to the population.[140]

In any case, not all cyber operations during armed conflicts would constitute "attacks" as understood in IHL. First, the concept of attack in IHL does not include espionage. Second, the rules on the conduct of hostilities do not prohibit all operations that interfere with civilian communication systems: the jamming of radio communications or television broadcasts has traditionally not been considered an attack as defined in IHL. However, the distinction between attacks and interferences with communications that do not amount to an attack is probably less clear in cyber operations than in more traditional kinetic or electromagnetic operations.[141] Third, the notion of "military operations" under IHL, including those carried out by cyber means, is broader than the notion of "attacks", as will be discussed below.

136 French Ministry of the Armies, above note 77, p. 13.
137 M. N. Schmitt, above note 52. In the same sense, see W. H. Boothby, above note 116, p. 386.
138 OAS, above note 22, para. 43.
139 Ecuador specified that "[a] cyber operation can qualify as an attack if it renders inoperable a state's critical infrastructure or others that endanger the security of the state". *Ibid.*, para. 44.
140 Bolivia suggested that a cyber operation "could be considered an attack when its objective is to disable a state's basic services (water, electricity, telecommunications, or the financial system"); Guyana suggested that "cyber operations that undermine the functioning of computer systems and infrastructure needed for the provision of services and resources to the civilian population constitute an attack", among which it included "nuclear plants, hospitals, banks, and air traffic control systems". *Ibid.*, paras 44–45.
141 ICRC Challenges Report 2015, above note 118, pp. 41–42; C. Droege, above note 83, p. 560.

The protection of data as a "civilian object"

In addition to the foundational question of which cyber operations amount to "attacks" under IHL, the question of whether civilian data enjoys the same protection as civilian objects has been subject to significant debate and remains unsettled. The protection of civilian data against malicious cyber operations during armed conflict is becoming increasingly important because data is an essential component of the digital domain and a cornerstone of life in many societies: individual medical data, social security data, tax records, bank accounts, companies' client files, and election lists and records are key to the functioning of most aspects of civilian life. As this trend is expected to continue, if not accelerate, in years to come, there is increasing concern about safeguarding such essential civilian data.

With regard to data belonging to certain categories of objects that enjoy specific protection under IHL, the protective rules are comprehensive. As discussed below, the obligations to respect and protect medical facilities and humanitarian relief operations must be understood as extending to medical data belonging to those facilities and data of humanitarian organizations that are essential for their operations.[142] Similarly, deleting or otherwise tampering with data in a manner that renders useless objects indispensable to the survival of the civilian population, such as drinking water installations and irrigation systems, is prohibited.[143]

Still, it is important to clarify the extent to which civilian data are protected by the existing general rules on the conduct of hostilities. In particular, debate has arisen on whether data constitute objects as understood under IHL, in which case cyber operations against data (such as deleting them) would be governed by the principles of distinction, proportionality and precaution and the protection they afford to civilian objects.[144]

This question is closely related to discussions on the notion of "attack" above. To start with, if data are deleted or manipulated in a manner that is designed or expected to cause, directly or indirectly, death or injury to a person, or damage to (including – in our view – by disabling) a physical object, the operation is an attack regardless of whether data themselves constitute objects for the purpose of IHL. This is the case because the consequences of an operation against data can qualify that operation as an attack under IHL and therefore subject to pertinent IHL. For these attacks, it is not important whether or not data qualifies as an object under IHL.

142 See discussion in the above section entitled "IHL Rules Protecting Objects Indispensable to the Survival of the Civilian Population".

143 AP I, Art. 54; Protocol Additional (II) to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts, 1125 UNTS 609, 8 June 1977 (entered into force 7 December 1978) (AP II), Art. 14; ICRC Customary Law Study, above note 63, Rule 54.

144 See Tallinn Manual 2.0, above note 13, paras 6–7 of the commentary on Rule 100. For academic discussion, see *Israel Law Review*, Vol. 48, No. 1, pp. 39–132; M. N. Schmitt, above note 67.

The question of whether data are objects for the purpose of IHL is, however, critical for operations that are not designed or expected to cause such consequences. Broadly speaking, two general approaches can be considered. Under the first approach, which considers data as objects under IHL, an operation designed or expected to delete or manipulate data would be an attack governed by all the relevant IHL rules because it would amount to destroying or damaging an object (the data). This would also be the case if such deletion or manipulation were not expected to cause death or injury to a person or to damage or disable a physical object. Even under this view, however, an operation designed solely to access (possibly confidential) data without deleting or manipulating them – such as spying – would not be an attack.

Conversely, if data are not considered to be objects under IHL, an operation designed to delete or manipulate them without causing death or injury to a person or damage to an object would not be governed by the rules on attacks, or by some of the more general rules affording protection to civilian objects (such as the obligation to take constant care to spare civilians and civilian objects, as discussed below in the section on "Rules Governing Military Operations Other than Attacks"). The operations could, however, be governed by other specific protection regimes under IHL, which will be analyzed below in the section on "IHL Rules Protecting Objects Indispensable to the Survival of the Civilian Population, Medical Services, and Humanitarian Relief Operations". Still, there would be a gap in protection for essential civilian data that do not benefit from a specific protection regime, and this would raise concern.

Experts hold different views on whether data qualify as objects for the purposes of the IHL rules on the conduct of hostilities.[145] One view, held by the majority of experts involved in the Tallinn Manual process, is that the ordinary meaning of the term "object", as discussed in the 1987 ICRC Commentary on AP I, cannot be interpreted as including data because objects are material, visible and tangible.[146] The relevant explanation in the ICRC Commentary, however, aims at distinguishing objects from concepts such as "aim" or "purpose", not at differentiating between tangible and intangible goods, and therefore cannot be seen as determinative for the debate on data.[147] In contrast, others have argued that either all or some types of data should be considered as objects under IHL. One view is that the "modern meaning" of the notion of objects in today's society, and an

145 For an illustration of this debate, see "Scenario 12: Cyber Operations against Computer Data", in K. Mačák, T. Minárik and T. Jančárková (eds), above note 68.

146 The Oxford Dictionary defines an object as "a material thing that can be seen and touched". Recalling the ordinary meaning of the word object, the 1987 ICRC Commentary on the Additional Protocols describes an object as "something that is visible and tangible". ICRC Commentary on the APs, above note 110, para. 2008. See also Tallinn Manual 2.0, above note 13, para. 6 of the commentary on Rule 100. It is interesting to note here that today, the Oxford Dictionary includes a specific definition of objects for computing: "A data construct that provides a description of anything known to a computer (such as a processor or a piece of code) and defines its method of operation."

147 See also International Law Association (ILA) Study Group on the Conduct of Hostilities in the 21st Century, "The Conduct of Hostilities and International Humanitarian Law: Challenges of 21st Century Warfare", *International Law Studies*, Vol. 93, 2017 (ILA Report), pp. 338–339.

interpretation of the term in light of its object and purpose, must lead to the conclusion that "data is an 'object' for the purposes of the IHL rules on targeting".[148] This interpretation is supported by the traditional understanding of the notion of "object" under IHL, which is broader than the ordinary meaning of the word and encompasses also locations and animals. Another proposal is to differentiate between "operational-level data", or "code", and "content-level data".[149] In this model, it has been argued that, notably, operational-level data may qualify as a military objective, which implies that this type of data could also qualify as a civilian object.[150] While considering operational data as objects would align with the view discussed above that disabling objects constitutes an attack, it does not appear to provide additional protection. In this debate, it has been argued that none of the proposed conclusions is entirely satisfactory, each being either under- or over-inclusive.[151]

For its part, the ICRC has stressed the need to safeguard essential civilian data, emphasizing that in cyberspace, deleting or tampering with data could quickly bring government services and private businesses to a complete standstill and could thereby cause more harm to civilians than the destruction of physical objects. Thus, in the ICRC's view the conclusion that this type of operation would not be prohibited by IHL in today's ever more cyber-reliant world seems difficult to reconcile with the object and purpose of this body of norms.[152] Logically, the replacement of paper files and documents with digital data should not decrease the protection that IHL affords to them.[153] As the ICRC has emphasized, "[e]xcluding essential civilian data from the protection afforded by IHL to civilian objects would result in an important protection gap".[154]

So far, few States have expressed views on whether the notion of "object" should be understood to encompass data for the rules governing the conduct of hostilities. For example, the Danish Military Manual considers that "(digital) data do not in general constitute an object".[155] Conversely, the Norwegian Military

148 Kubo Mačák, "Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law", *Israel Law Review*, Vol. 48, No. 1, 2015, p. 80; Robert McLaughlin, "Data as a Military Objective", Australian Institute of International Affairs, 20 September 2018, available at: www.internationalaffairs.org.au/australianoutlook/data-as-a-military-objective/.

149 Under the proposed distinction, content-level data would include data "such as the text of this article, or the contents of medical databases, library catalogues and the like", whereas operational-level data would describe "essentially the 'soul of the machine'", meaning the "type of data that gives hardware its functionality and ability to perform the tasks we require". Heather Harrison Dinniss, "The Nature of Objects: Targeting Networks and the Challenge of Defining Cyber Military Objectives", *Israel Law Review*, Vol. 48, No. 1, 2015, p. 41.

150 *Ibid.*, p. 54.

151 Schmitt therefore argues that as a matter of policy, States should "accord special protection to certain 'essential civilian functions or services' by committing to refrain from conducting cyber operations against civilian infrastructure or data that interfere with them". M. N. Schmitt, above note 67, p. 342.

152 ICRC Challenges Report 2015, above note 118, p. 43.

153 ICRC Challenges Report 2019, above note 36, p. 21.

154 ICRC, above note 1, p. 8. See also P. Pascucci, above note 67, who notes that the position adopted by the majority of the experts in the Tallinn Manual with regard to data creates a "seemingly expansive gap in what constitutes an object", and later argues that "[i]t is unrealistic in an information age for data to fall outside the scope of constituting an object, thus failing to receive IHL protection associated with the principles of distinction and proportionality".

155 Danish Military Manual, above note 118, p. 292.

Manual holds that data shall be regarded as objects and may only be attacked directly if they qualify as a lawful target.[156] France has expressed what could be seen as a middle-ground view, stating that "[g]iven the current state of digital dependence, content data (such as civilian, bank or medical data, etc.) are protected under the principle of distinction".[157] The description of Peru's position in the OAS's *Improving Transparency* report appears to reflect a similar position: while not expressly taking a position on whether data are objects, Peru's position is explained as assessing operations against data under the notion of "military objective", suggesting that some data systems may not be subjected to attacks because such attacks would "not create a legitimate military advantage".[158] As the definition of "military objectives" under Article 52(2) of AP I applies "in so far as objects are concerned", this reasoning appears to imply that data constitute objects. Chile proposes to look at the effects of an attack against data, concluding that "[t]he principle of distinction must therefore be taken into consideration in the context of cyber operations, whereby a state should refrain from attacking data in case it could affect the civilian population". Reportedly, Chile has further emphasized that "an attack directed exclusively at computer data could well produce adverse consequences affecting the civilian population".[159]

In an increasingly data-reliant world, the question of how States interpret and apply IHL rules to safeguard essential data against destruction, deletion or manipulation will be a litmus test for the adequacy of existing humanitarian law rules.

The protection of cyber infrastructure serving simultaneously military and civilian purposes

In order to protect critical civilian infrastructure that relies on cyberspace, it is also crucial to protect the infrastructure of cyberspace itself. The challenge lies, however, in the interconnectedness of civilian and military networks. Most military networks rely on civilian cyber infrastructure, such as undersea fibre-optic cables, satellites, routers or nodes. Civilian vehicles, shipping and air traffic control are increasingly equipped with navigation equipment that relies on global navigation satellite system (GNSS) satellites such as BeiDou, GLONASS, GPS and Galileo, which may also be used by the military. Civilian logistical supply chains (for food and medical supplies) and other businesses use the same web and communication networks through which some military communication passes. Except for certain networks that are specifically dedicated to military use, it is to a large extent impossible to differentiate between purely civilian and purely military cyber infrastructures.

156 Norwegian Military Manual, above note 118, para. 9.58.
157 French Ministry of the Armies, above note 77, p. 14.
158 OAS, above note 22, para. 49, fn. 115.
159 *Ibid.*, para. 48.

Under IHL, attacks must be strictly limited to military objectives. Insofar as objects are concerned, military objectives are limited to those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage. All objects which are not military objectives under this definition are civilian objects under IHL and must not be made the object of an attack or of reprisals. In case of doubt as to whether an object that is normally dedicated to civilian purposes is being used to make an effective contribution to military action, it must be presumed to remain protected as a civilian object.[160]

It is traditionally understood that an object may become a military objective when its use for military purposes is such that it fulfils the definition of military objective even if it is simultaneously used for civilian purposes. A wide interpretation of this rule could lead to the conclusion that many objects forming part of cyberspace infrastructure would constitute military objectives and would therefore not be protected against attack, whether cyber or kinetic. This would be a matter of serious concern because of the ever-increasing civilian reliance on cyberspace.

Such a conclusion would, however, be incomplete. First, the analysis of when a civilian object becomes a military objective cannot be done for cyberspace or the Internet in general. Instead, belligerents must identify which computer, nodes, routers or networks might have become a military objective. In this respect, parts of the network, specific computers, or other hardware that can be separated from a network or system as a whole need to be analyzed individually. The means and methods used must enable directing the attack at the specific military objective(s) that may have been identified, and all feasible precautions must be taken to avoid or at least minimize incidentally affecting the remaining civilian objects or parts of the network.[161] It has also been argued that it is prohibited to treat as a single target a number of clearly discrete cyber military objectives in cyber infrastructure primarily used for civilian purposes if to do so would harm protected persons or objects.[162] Second, cyberspace is designed with a high level of redundancy, meaning that one of its characteristics is the ability to immediately re-route data traffic. This inbuilt resilience needs to be considered when assessing whether the target's destruction or neutralization would offer a definite military advantage as required by the definition of a military objective. If this is not the case, the object would remain civilian and cannot be attacked. And third, any attack is governed by the prohibition against indiscriminate attacks and the rules of proportionality and precautions in attack. Stopping or impairing the civilian use of an object in violation of one of these rules would

160 See AP I, Art. 52. ICRC Customary Law Study, above note 63, Rules 7–10.
161 See AP I, Arts 51(4), 57(2)(a)(ii); ICRC Customary Law Study, above note 63, Rules 12–17.
162 See Tallinn Manual 2.0, above note 13, Rule 112, which derives from the prohibition on area bombardment found in Article 51(5)(a) of AP I and customary IHL (see ICRC Customary Law Study, above note 63, Rule 13).

render the attack unlawful despite the fact that the object had become a military objective.[163]

Compared to kinetic military operations, the use of cyber operations may, depending on the circumstances, enable achieving a particular effect while causing less destruction (on the target or incidentally on other objects or systems) or causing damage that may be more easily repaired or restored. This consideration is particularly relevant with regard to dual-use objects, as illustrated by the scenario of a belligerent trying to neutralize an enemy underground command bunker by cutting its electricity supply, which is simultaneously providing power to civilian infrastructure. A cyber operation may allow the operator to remotely choose which parts of the network to switch off.[164] This could enable the attacker to achieve the desired effect while avoiding, or at least minimizing, harmful effects to the delivery of electricity to civilians. In such a case, and provided that choosing to use a cyber operation instead of a kinetic one is *feasible*, conducting the cyber operation would become required by the principle of precaution. Indeed, the obligation to take all feasible precautions in the choice of means and methods of warfare to avoid or at least minimize incidental civilian harm[165] is technologically neutral: it also applies to means and methods relying on new technologies, and may even require their use.[166] Whether this is feasible in a specific instance depends on the circumstances ruling at the time, including humanitarian and military considerations.[167]

Limitations on cyber operations other than those amounting to "attacks", including the specific protection of certain persons and objects

While many of the general rules on the conduct of hostilities are limited to acts amounting to attacks as defined in IHL, some IHL rules governing the conduct of hostilities apply to a broader set of operations: first, a few rules apply to all "military operations", and second, the specific protection afforded to certain categories of persons and objects goes beyond the protection against attacks.

163 While acknowledging that the other view also exists, the ILA Study Group on the Conduct of hostilities deemed this "the better view" based on State practice, official documents and doctrine: see ILA Report, above note 147, pp. 336–337. See also ICRC, *International Expert Meeting Report: The Principle of Proportionality in the Rules Governing the Conduct of Hostilities under International Humanitarian Law*, Geneva, 2018, p. 39, available at www.icrc.org/en/document/international-expert-meeting-report-principle-proportionality; Helen Durham, Keynote Address, in Edoardo Greppi (ed.), *Conduct of Hostilities: The Practice, the Law and the Future*, 37th Round Table on Current Issues of International Humanitarian Law, International Institute of Humanitarian Law, Sanremo, 2015, p. 31.

164 This was reportedly done in the 2015 cyber operations against the electricity grid in Ukraine. See Kim Zetter, "Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid", *Wired*, 3 March 2016, available at: www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/.

165 AP I, Art. 57(2)(a)(ii); ICRC Customary Law Study, above note 63, Rule 17.

166 See ILA Report, above note 147, p. 384.

167 While military considerations might include the "fragility" of cyber means and methods, this is not the only relevant factor determining feasibility. It is not possible to rule out that it is feasible, and therefore required, to use cyber operations to avoid or minimize incidental civilian harm on the sole basis that the cyber means or methods used are "fragile", without looking at the entirety of the situation, including all relevant humanitarian considerations.

Rules governing military operations other than attacks

Identifying, and possibly clarifying, the rules that offer general protection to the civilian population and civilian objects against the effects of cyber operations that do not amount to attacks is an issue requiring more attention. This is all the more critical if the view is taken that only those operations causing physical damage are considered as attacks: in that case, there would be a rather broad category of cyber operations to which only a limited set of IHL rules applies. Such a conclusion would cause real concern for the protection of civilians and civilian infrastructure.

The notion of "military operation" appears in a number of articles of the 1949 Geneva Conventions and their 1977 Additional Protocols.[168] Of most interest here are the rules that regulate the conduct of military operations, including those carried out by cyber means. They include the basic rule that "parties to the conflict … shall direct their operations only against military objectives" (AP I, Article 48), the principle that "the civilian population and individual civilians shall enjoy general protection against dangers arising from military operations" (AP I, Article 51(1)),[169] and the obligation that "constant care shall be taken to spare the civilian population, civilians and civilian objects" in the conduct of military operations (AP I, Article 57(1)).[170]

The ordinary meaning of the term "military operation" and a systematic interpretation of these articles lead to the conclusion that this notion is different from the notion of "attack" as defined in Article 49 of AP I.[171] While the ICRC Commentary on Article 48 of AP I notes that the notion refers to military operations during which violence is used, and not to ideological, political or religious campaigns, it clarifies that it is a broader notion than "attacks". The Commentary defines "military operations" for the purpose of these articles as "any movements, manoeuvres and other activities whatsoever carried out by the armed forces with a view to combat" or "related to hostilities" – an understanding that is widely accepted.[172]

168 See GC III, Art. 23; GC IV, Art. 28; AP I, Arts 3, 39, 44, 51, 56–60; AP II, Art. 13.

169 See also AP I, Art. 58; AP II, Art. 13(1).

170 See also ICRC Customary Law Study, above note 63, Rule 15; Tallinn Manual 2.0, above note 13, Rule 114.

171 An interpretation that assimilates the notions of "operation" and "attack" would deprive the rules applying to "operations" of meaningful content and render them essentially superfluous. See C. Droege, above note 83, p. 556.

172 ICRC Commentary on the APs, above note 110, paras 2191, 1936, 1875. In the same vein, see Michael Bothe, Karl Josef Partsch and Waldemar A. Solf, *New Rules for Victims of Armed Conflict: Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949*, Martinus Nijhoff, Leiden, 2013, para. 2.2.3 on Art. 48, para. 2.8.2 on Art. 57; UK Ministry of Defence, *The Joint Service Manual of the Law of Armed Conflict*, Joint Service Publication 383, 2004 (UK Military Manual), para 5.32, fn. 187; ILA Report, above note 147, p. 380. The *HPCR Manual on International Law Applicable to Air and Missile Warfare* (Program on Humanitarian Policy and Conflict Research, Harvard University, 2009) applies the constant care obligation to "air or missile combat operations" (Rule 34), a notion broader than "attack" that includes, *inter alia*, refuelling, jamming of enemy radars, use of airborne warning systems and dropping an airborne force (commentary on Rule 1(c), para. 3). See also Noam Neuman, "A Precautionary Tale: The Theory and Practice of Precautions in Attack", *Israel Yearbook on Human Rights*, Vol. 48, 2018, p. 28; Jean-François Quéguiner, "Precautions under

The notion is mostly discussed in relation to the treaty and customary obligation to take constant care to spare the civilian population, civilians and civilian objects in the conduct of military operations. France has explicitly stated that this obligation also applies in cyberspace.[173] This obligation requires all those involved in military operations to continuously bear in mind the effects of military operations on the civilian population, civilians and civilian objects, to take steps to reduce such effects as much as possible, and to seek to avoid any unnecessary effects.[174] It has been described as a positive and continuous obligation aimed at risk mitigation and harm prevention that imposes requirements which increase commensurably with the risk to civilians.[175] The Tallinn Manual explains in this respect that

> [t]he law admits of no situation in which, or time when, individuals involved in the planning and execution process may ignore the effects of their operations on civilians or civilian objects. In the cyber context, this requires situational awareness at all times, not merely during the preparatory stage of an operation.[176]

A more challenging issue is the application of the principle of distinction to military operations other than attacks. As noted above, Article 48 of AP I requires that military operations be directed only against military objectives. While the commentaries by the ICRC and by Bothe, Partsch and Solf[177] underline the fundamental character of this article, they do not shed much light on the precise meaning and scope of this obligation, which remain subject to debate.

Article 48 is sometimes understood as an overarching principle that is implemented through the application of the various rules of the section of the Protocol that it opens. Some commentators therefore argue that the specific rules stemming from the principle of distinction apply only to attacks and not to military operations other than attacks.[178] Accordingly, some military manuals

the Law Governing the Conduct of Hostilities", *International Review of the Red Cross*, Vol. 88, No. 864, 2006, p. 797; Chris Jenks and Rain Liivoja, "Machine Autonomy and the Constant Care Obligation", *Humanitarian Law and Policy*, 11 December 2018, available at: https://blogs.icrc.org/law-and-policy/2018/12/11/machine-autonomy-constant-care-obligation/. Specifically with regard to cyber operations, see Tallinn Manual 2.0, above note 13, para. 2 of the commentary on Rule 114 (noting that the notion of hostilities, to which it applies the constant care obligation, is broader than the notion of attacks); H. Harrison Dinniss, above note 124, p. 199. For a different view at least with regard to the principle of distinction, see M. Roscini, above note 127, p. 178.

173 French Ministry of the Armies, above note 77, p. 15.

174 UK Military Manual, above note 172, para. 5.32.1; Tallinn Manual 2.0, above note 13, para. 4 of the commentary on Rule 114; Dieter Fleck, *The Handbook of International Humanitarian Law*, 3rd ed., Oxford University Press, Oxford, 2013, p. 199; N. Neuman, above note 172, pp. 28–29.

175 ILA Report, above note 147, p. 381.

176 Tallinn Manual 2.0, above note 13, para. 4 of the commentary on Rule 114.

177 M. Bothe, K. J. Partsch and W. A. Solf, above note 172.

178 M. Roscini, above note 126, p. 178. See also, though expressed under customary law, Tallinn Manual 2.0, above note 13, para. 5 of the commentary on Rule 93; Michael N. Schmitt, "'Attack' as a Term of Art in International Law: The Cyber Operations Context", in Christian Czosseck, Rain Ottis and Katharina Ziolkowski (eds), *4th International Conference on Cyber Conflict: Proceedings*, NATO CCD COE Publications, Tallinn, 2012, pp. 283–293, 289–290.

expressly state that cyber operations other than attacks may be directed at civilians or civilian objects.[179] This assertion would seem difficult to reconcile with Article 48 for States party to the Protocol, or at least would need to be carefully articulated. Indeed, experts have pointed out that "[w]hile … there is a distinction between military operations and attacks, it does not follow that non-violent computer network attacks may be therefore conducted against civilian objects".[180] This conclusion stems from the rules of treaty interpretation, which require that provisions are interpreted to "have a meaningful content and are not superfluous".[181]

As noted above, "military operations" are understood as any movements, manoeuvres or other activities whatsoever carried out by the armed forces with a view to combat or related to hostilities. Manoeuvring is also an integral part of cyber operations.[182] For instance, establishing remote access to one system or device might be a step towards reaching or attacking another system or device.[183] Assuming that the former system or device is civilian in character and the latter is a military objective, the question may arise as to whether establishing access to the civilian system or device would be a prohibited military operation. In the view of the present authors, provided that the civilian system or device is not damaged or disabled in the process, such a scenario does not appear to contravene Article 48 because the operation is, ultimately, directed at a military objective.[184] Such cyber operations may indeed be assessed in the same way as traditional military operations – for example, when a commando moves through a civilian house to attack a military objective which is located behind it. Still, other obligations would remain relevant, such as the obligation to take constant care to spare civilian objects.

In the view of the present authors, Article 48, alone or in combination with Articles 51(1) and 57(1) of AP I, should be interpreted as prohibiting cyber operations designed solely at disrupting internet services for the civilian population even if such cyber operations do not disable objects or otherwise have consequences that qualify them as attacks. The civilian use of the Internet is today so all-pervading that any other interpretation would leave an important gap in the protection that IHL affords to civilians against the effects of hostilities carried out by cyber means.[185]

179 Norwegian Military Manual, above note 118, para. 9.57. See also DoD Law of War Manual, above note 87, para. 16.5.2.

180 H. Harrison Dinniss, above note 124, p. 199.

181 See also C. Droege, above note 83, p. 556.

182 See, for example, US DoD, *Cyberspace Operations*, Joint Publication 3-12, 8 June 2018, p. xii: "Movement and Maneuver. Cyberspace operations enable force projection without the need to establish a physical presence in foreign territory. Maneuver in the DODIN [Department of Defense Information Network] or other blue [friendly] cyberspace includes positioning of forces, sensors, and defenses to best secure areas of cyberspace or engage in defensive actions as required. Maneuver in gray [neutral] and red [enemy] cyberspace is a cyberspace exploitation action and includes such activities as gaining access to adversary, enemy, or intermediary links and nodes and shaping this cyberspace to support future actions."

183 L. Gisel and L. Olejnik (eds), above note 11, p. 57.

184 Compare with H. Harrison Dinniss, above note 124, p. 201.

185 The experts who drafted the Tallinn Manual discussed whether disrupting all email communications throughout a country during an armed conflict would amount to an attack – a narrower notion than

As discussed above, some hold the view that not all cyber operations which disable objects, or which delete or manipulate data, constitute attacks. As a consequence of such interpretations, a significantly broader range of cyber operations would not be governed by the rules on attacks, including operations that would create a significant risk of harm. It is therefore all the more important for the protection of the civilian population that those who interpret narrowly the notions of "attack" and "objects" clarify whether they consider that cyber operations which merely disable objects or delete data amount to "military operations", and what this means for the application of the principle of distinction to such operations – in particular, what the requirement of Article 48 of AP I that parties to armed conflicts "shall direct their operations only against military objectives" entails. For instance, at least a certain level of protection would be retained if those who interpret the notion of "attack" narrowly accept that a cyber operation that merely disables objects is a "military operation" which as a consequence must be directed only against military objectives.

Even cyber operations that do not fall within the notion of "military operation" as understood in AP I might be regulated by some IHL rules stemming from the principle of distinction. For example, it has been noted that "directing" psychological operations or other types of propaganda at civilians would not violate Article 48 of AP I because these operations would not fall within the meaning of "military operations" as understood in Article 48.[186] Yet, psychological operations are not beyond the protective reach of other IHL norms. For instance, they must not amount to prohibited acts or threats of violence the primary purpose of which is to spread terror among the civilian population or encourage IHL violations.[187]

Limitations on cyber operations other than attacks may also stem from the principle of military necessity. Relying on the customary rule going back to the 1907 Hague Regulation, the US DoD Law of War Manual states that "[a] cyber operation that would not constitute an attack, but would nonetheless seize or destroy enemy property, would have to be imperatively demanded by the necessities of war".[188] The Manual also refers to military necessity in a more generic manner, specifying that cyber operations which do not amount to attack "must not be directed against enemy civilians or civilian objects unless the operations are militarily necessary".[189] In a similar manner, Australia notes that "[a]pplicable IHL rules will also apply to cyber operations in an armed conflict that do not constitute or rise to the level of an 'attack', including the principle of military necessity".[190] While these references to military necessity as a restraining principle are

military operations. While a minority held the view that the international community would generally regard such an operation as an attack, the majority held the view that IHL did not presently extend this far, but nevertheless considered that there was logic in characterizing these operations as attacks. Tallinn Manual 2.0, above note 13, para. 13 of the commentary on Rule 92.

186 C. Droege, above note 83, p. 556.

187 ICRC Challenges Report 2019, above note 36, pp. 28–29.

188 DoD Law of War Manual, above note 87, para. 16.5.1.

189 *Ibid.*, para. 16.5.2.

190 Australian Department of Foreign Affairs and Trade, above note 94, p. 4.

welcome, more clarity is needed on exactly what the principle of military necessity prescribes when conducting cyber operations.

This brief discussion shows that cyber operations other than attacks are not unregulated. The legal regime governing military operations remains, however, less complete, precise and demanding than the legal regime governing operations that amount to attacks under IHL. To address this protection gap at least to some extent, Schmitt has put forward a suggestion that States should apply – as a matter of policy – an adapted proportionality assessment to cyber operations that do not amount to attacks.[191]

The specific protection that IHL provides to certain persons and objects further restricts the scope of permissible military operations.

IHL rules protecting objects indispensable to the survival of the civilian population, medical services, and humanitarian relief operations

In addition to the general rules on the conduct of hostilities, IHL sets out specific regimes for certain objects and services that afford additional and stronger protection than the protection granted to all civilians and civilian objects.

For instance, IHL specifically makes it illegal "to attack, destroy, remove or render useless objects indispensable to the survival of the civilian population".[192] This rule protects, for example, "food-stuffs", "agricultural areas for the production of food-stuffs", and "drinking water installations and supplies and irrigation works".[193] While the experts who drafted the Tallinn Manual held that the Internet as such cannot be considered as an object indispensable to the survival of the civilian population, they noted that "cyber infrastructure indispensable to the functioning of electrical generators, irrigation works and installations, drinking water installations, and food production facilities could, depending on the circumstances, qualify".[194] The explicit mention of "rendering useless" must be understood as covering a broader range of operations that may impact these goods, beyond attacks or destruction. As noted in the ICRC Commentary on Article 54(2) of AP I, the intent of the drafters was "to cover all possibilities" of how objects for the subsistence of the civilian population can be rendered useless.[195] Today, cyber operations that are designed, or can be expected, to render objects indispensable for the civilian population useless are prohibited, regardless of whether they amount to an attack. The debate on whether military operations against these objects amount to an attack (as discussed above) is therefore moot for these objects.

191 M. N. Schmitt, above note 67, p. 347: "States would commit, as a matter of policy, to refraining from conducting cyber operations to which the IHL rules governing attacks do not apply when the expected concrete negative effects on individual civilians or the civilian population are excessive relative to the concrete benefit related to the conflict that is anticipated to be gained through the operation."

192 See AP I, Art. 54(2); AP II, Art. 14; ICRC Customary Law Study, above note 63, Rule 54.

193 AP I, Art. 54(2).

194 Tallinn Manual 2.0, above note 13, para. 5 of the commentary on Rule 141.

195 ICRC Commentary on the APs, above note 110, paras 2101, 2103.

IHL also provides specific protection for medical services. Given the fundamental importance of health care for anyone affected by armed conflict, belligerents must respect and protect medical facilities and personnel at all times.[196] The obligation to "respect" medical facilities and personnel is understood as not only protecting them against operations that amount to attacks – it is prohibited to "harm them in any way. This also means that there should be no interference with their work (for example, by preventing supplies from getting through) or preventing the possibility of continuing to give treatment to the wounded and sick who are in their care."[197] The special protection of medical facilities includes medical communication: while jamming enemy communication is generally considered permissible, "an intentional disruption of [medical] units' ability to communicate for medical purposes" may not be permissible, even if medical units communicate with the armed forces.[198] Moreover, the obligation to respect and protect medical facilities encompasses a prohibition against deleting, altering or otherwise negatively affecting medical data.[199] It may also provide protection against cyber operations directed at the confidentiality of medical data, which at least in some circumstances would be hard to reconcile with the obligation to protect and respect medical facilities.[200] Relevant data in the medical context include "data necessary for the proper use of medical equipment and for tracking the inventory of medical supplies" as well as

196 See, for instance, GC I, Art. 19; GC II, Art. 12; GC IV, Art. 18; AP I, Art. 12; AP II, Art. 11; ICRC Customary Law Study, above note 63, Rules 25, 28, 29; Tallinn Manual 2.0, above note 13, Rules 131–132. Protection of medical facilities and personnel ceases only if they commit, or are used to commit, outside their humanitarian duties, acts harmful to the enemy. Protection may, however, cease only after a due warning has been given, naming, in all appropriate cases, a reasonable time limit, and after such warning has remained unheeded. See GC I, Art. 21; GC II, Art. 34; GC IV, Art. 19; AP I, Art. 13; AP II, Art. 11(2); ICRC Customary Law Study, above note 63, Rules 25, 28, 29; Tallinn Manual 2.0, above note 13, Rule 134.

197 ICRC Commentary on the APs, above note 110, para. 517. See also ICRC Commentary on GC I, above note 69, para. 1799; Oxford Statement, above note 32, point 5 ("During armed conflict, international humanitarian law requires that medical units, transport and personnel must be respected and protected at all times. Accordingly, parties to armed conflicts: must not disrupt the functioning of health-care facilities through cyber operations; must take all feasible precautions to avoid incidental harm caused by cyber operations, and; must take all feasible measures to facilitate the functioning of health-care facilities and to prevent their being harmed, including by cyber operations"); Tallinn Manual 2.0, above note 13, para. 5 of the commentary on Rule 131 ("For instance, this Rule [Rule 131, which states that "[m]edical and religious personnel, medical units, and medical transports must be respected and protected and, in particular, may not be made the object of cyber attack"] would prohibit altering data in the Global Positioning System of a medical helicopter in order to misdirect it, even though the operation would not qualify as an attack on a medical transport").

198 ICRC Commentary on the APs, above note 110, para. 1804.

199 See ICRC Challenges Report 2015, above note 118, p. 43.

200 See L. Gisel and L. Olejnik (eds), above note 11, p. 36, discussing the hypothetical of hacking into the medical or administrative records of a medical facility in order to gain knowledge of an enemy commander's medical appointment so as to locate him in order to capture or kill him on the way to or back from the medical facility. This could indeed unduly impede the facility's medical functioning and hinder the ability of health-care professionals to uphold their ethical duty of preserving medical confidentiality. The Tallinn Manual 2.0, above note 13, para. 2 of the commentary on Rule 132, proposes the following as an example of an operation that would not violate IHL: "non-damaging cyber reconnaissance to determine whether the medical facility or transports (or associated computers, computer networks, and data) in question are being misused for militarily harmful acts".

"personal medical data required for the treatment of patients".[201] The obligation to "protect" medical facilities, including their data, entails positive obligations. Parties to the conflict must actively take measures to protect medical facilities from harm to the extent feasible, including harm resulting from cyber operations.[202]

IHL also prescribes that humanitarian personnel and relief consignments must be respected and protected.[203] This obligation certainly prohibits any "attacks" against humanitarian operations. In the same way as for the obligation to respect and protect medical personnel and facilities, relevant rules should also be understood as prohibiting "other forms of harmful conduct outside the conduct of hostilities" against humanitarians or undue interference with their work.[204] Moreover, parties to armed conflicts are required to agree, allow and facilitate humanitarian relief operations.[205] Accordingly, Rule 145 of the Tallinn Manual 2.0 states that "cyber operations shall not be designed or conducted to interfere unduly with impartial efforts to provide humanitarian assistance", and specifies that such operations are prohibited "even if they do not rise to the level of an 'attack'".[206] The obligation to respect and to protect relief personnel and operations should also be understood as protecting relevant data.[207] At least for States party to AP I, the protection of humanitarian data should encompass data of the ICRC that the organization needs "to carry out the humanitarian functions assigned to it by the [Geneva] Conventions and this Protocol in order to ensure protection and assistance to the victims of conflicts".[208]

These special protections show that IHL provides more stringent rules for military operations against certain goods or services that are essential for the survival, health and well-being of the civilian population.

The importance of legal reviews of cyber means and methods of warfare to ensure respect for IHL

In light of the particular challenges that the characteristics of cyberspace pose to the interpretation and application of some IHL principles in the conduct of hostilities, parties to armed conflicts who choose to develop, acquire or adopt weapons, means or methods of warfare relying on cyber technology need to exercise care in doing so. In this respect, States party to AP I that develop or acquire cyber warfare

201 Tallinn Manual 2.0, above note 13, para. 3 of the commentary on Rule 132.
202 ICRC Commentary on GC I, above note 69, paras 1805–1808; Tallinn Manual 2.0, above note 13, para. 6 of the commentary on Rule 131.
203 AP I, Arts 70(4), 71(2); ICRC Customary Law Study, above note 63, Rules 31, 32.
204 ICRC Commentary on GC I, above note 69, paras 1358, 1799.
205 See, for instance, GC IV, Art. 59; AP I, Arts 69–70; ICRC Customary Law Study, above note 63, Rule 55.
206 Tallinn Manual 2.0, above note 13, para. 4 of the commentary on Rule 80.
207 For further discussion, see Tilman Rodenhäuser, "Hacking Humanitarians? IHL and the Protection of Humanitarian Organizations against Cyber Operations", *EJIL: Talk!*, 16 March 2020, available at: www.ejiltalk.org/hacking-humanitarians-ihl-and-the-protection-of-humanitarian-organizations-against-cyber-operations/.
208 AP I, Art. 81. Such data include, for example, those needed to establish tracing agencies to collect information on persons reported missing in the context of an armed conflict, or those collected by the ICRC when visiting and interviewing detainees without witnesses.

capacities – whether for offensive or defensive purposes – have an obligation to assess whether the employment of the cyber weapon, means or method of warfare would be prohibited by international law in some or all circumstances.[209] More broadly, legal reviews are critical for all States to ensure respect for IHL by their armed forces,[210] so that they only use weapons, means or methods of warfare, including those relying on cyber technology, that comply with the State's obligations under IHL.[211] Such reviews should involve a multi-disciplinary team including legal, military and technical experts as relevant.[212] These legal reviews need to be conducted earlier and in more depth than the analysis of the legality of the actual use of a tool in the specific circumstance of an attack.

In view of the novelty of the technology, it is critical that the legal review of cyber weapons, means and methods of warfare is given particular attention. The prohibition of weapons that are by nature indiscriminate may be particularly relevant considering the ability of certain cyber tools to self-propagate autonomously.[213] The legal review of cyber weapons, means and methods of warfare may, however, present a number of challenges. In the following, we illustrate some issues, without being exhaustive.

First, a State conducting a legal review needs to determine against which legal standards it reviews a cyber tool. In other words, the State will need to have answers to some of the questions discussed above, such as whether the use of a tool will qualify as an attack and is therefore subject to a broad range of IHL rules. For matters where the law is unclear or unsettled, a cautious approach may be warranted to avoid the subsequent appearance that the employment of a cyber tool was, or should have been deemed, unlawful.

Second, a State needs to determine what needs to be reviewed. This may not necessarily be evident with regard to cyber tools or cyber capabilities, as shown by the widespread use of these terms instead of notions such as cyber weapons. Commentators have discussed whether and which cyber tools or capabilities are weapons, means or methods of warfare, and what the implications are with respect to their legal review.[214] In any case, as noted above, States party to AP I must review all cyber tools or capabilities that qualify as weapons, means or methods of warfare. For States that are not party to AP I, the obligation to respect and ensure respect for IHL by their armed forces and the newness of the use of cyber technologies as weapon, means or method of warfare would make it

209 AP I, Art. 36.

210 See common Article 1; ICRC Customary Law Study, above note 63, Rule 139.

211 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*, Geneva, 2006, p. 1.

212 *Ibid.*, pp. 22–23.

213 See ICRC Customary Law Study, above note 63, Rule 71. For an illustration of some of the issues raised by the legal review of cyber weapons, see "Scenario 10: Cyber Weapons Review", in K. Mačák, T. Minárik and T. Jančárková (eds), above note 68.

214 Jeffrey T. Biller and Michael N. Schmitt, "Classification of Cyber Capabilities and Operations as Weapons, Means, or Methods of Warfare", *International Law Studies*, Vol. 95, 2019, p. 219.

prudent to cast as large a net as possible in terms of the capabilities being reviewed.[215]

Third, a weapon or means of warfare should not be assessed in isolation from the way in which it will be used, meaning that the normal or expected use of the weapon or means of warfare must be considered in the legal review. Military cyber capabilities might, however, be less standardized than kinetic weapons, especially if designed for a specific operation. This would mean that the review needs to be done in view of the specific cyber environment in which the weapon will likely be used.

Fourth, and relatedly, a State should conduct a legal review not only for a weapon, means or method of warfare it intends to acquire or adopts for the first time, but also when it modifies a weapon, means or method that has already passed a legal review. This may pose a challenge with regard to cyber tools that are likely to be subject to frequent adaptation, including to respond to the software security upgrades that a potential target undergoes. While the question of the type and extent of change that would require a new legal review might need to be further clarified, a new legal review must be conducted, notably, when the weapon, means or method of warfare is modified in a way that alters its function or when the modification could otherwise have an impact on whether the employment of the weapon, means or method would comply with the law.[216] It has been noted in this respect with regard to cyber weapons that "the assessment of whether a change will affect a program's operation must be qualitative rather than quantitative in nature".[217] For legal reviews to be effective, States that study, develop, acquire or adopt new weapons, means or methods relying on new technologies need to navigate these and other complexities. In other words, testing regimes must adapt to the unique characteristics of cyber technology. In light of the above-mentioned complexities, a good practice to ensure respect for IHL by all States would be to share information about a State's legal review mechanisms and, to the extent feasible, about the substantive results of legal reviews.[218] This would be especially important where problems of compatibility of a weapon with IHL arise, in order to avoid other States encountering the same problems and to notify other States of the testing State's conclusions that such tools are prohibited by IHL. Exchange of information on legal reviews of weapons, means or methods relying on new technologies can also help build expertise and facilitate the identification of good practices, which may

215 For example, while the US DOD had the policy of carrying out legal review of weapons, including weapons that employ cyber capabilities (DoD Law of War Manual, above note 87, para 16.6), the relevant US Air Force instruction mandates the review of weapons and cyber capabilities: US Department of the Air Force, *Legal Reviews of Weapons and Cyber Capabilities*, Air Force Instruction 51-402, 27 July 2011.

216 ICRC, above note 211, p. 10.

217 Gary D. Brown and Andrew O. Metcalf, "Easier Said than Done: Legal Reviews of Cyber Weapons", *Journal of National Security Law and Policy*, Vol. 7, 2014, p. 133.

218 This has been proposed in the introductory remarks delivered by Helen Durham, Director of International Law and Policy of the ICRC, during the 22 January 2019 public hearing conducted by the Global Commission on the Stability of Cyberspace (statement on file with the ICRC).

assist States that wish to establish or strengthen their own legal review mechanisms.[219]

Conclusion

For the protection of the civilian population and civilian infrastructure in armed conflict, it is fundamentally important to recognize that cyber operations conducted during armed conflicts do not occur in a legal void but are regulated by international law, most notably IHL. As this article shows, recognizing IHL applicability is, however, not the end of the conversation. More discussion – in particular among States – is needed on how IHL is to be interpreted in cyberspace. Any such discussion should be informed by an in-depth understanding of the development of military cyber capabilities, the potential human cost they may cause, and the protection afforded by existing law. This article is meant to provide a basis for such discussions. While the use of cyber operations during armed conflicts, their potential human cost and States' legal positions on the subject are evolving, the analysis in this article presents a number of conclusions.

First, cyber operations during armed conflicts are a reality in today's armed conflicts and their use is likely to increase in the future. They can cause significant harm to the civilian population, especially if affecting critical civilian infrastructure such as medical facilities, electricity, water or sanitation. While the risk of causing human harm does not appear extremely high based on current observations, especially considering the destruction and suffering that conflicts always cause, the evolution of cyber operations requires close attention due to existing uncertainties and the rapid pace of change.

Second, in the ICRC's view, there is no question that cyber operations during armed conflicts are regulated by IHL – just as is any weapon, means or method of warfare used by a belligerent in a conflict, whether new or old. While the issue does not (yet) enjoy universal agreement, a careful examination of the various arguments raised in multilateral discussions shows that affirming IHL applicability legitimizes neither the militarization of cyberspace nor the use of malicious cyber operations. A State that considers carrying out a cyber operation against another State must analyze the lawfulness of this operation under the UN Charter and IHL. These two frameworks are complementary when it comes to the protection of humans from war and its effects. While some of the terminology they use is similar, the two frameworks are legally separate and require distinct analyses, as similar terminology has (at times) distinct meaning. For instance, concluding that a cyber operation triggers the applicability of IHL does not necessarily mean that it amounts to an armed attack giving rise to the right of self-defence.

219 ICRC Challenges Report 2019, above note 36, p. 35.

Third, the partly non-physical – i.e., digital – nature of cyberspace and the interconnectedness of military and civilian networks pose practical and legal challenges in applying the general IHL principles and rules protecting civilians and civilian objects. This is particularly the case with regard to the notion of "attack" under IHL, the question of whether civilian data enjoys similar protection as "civilian objects", and the protection of "dual-use" cyber infrastructure.

The question of whether or not an operation amounts to an "attack" as defined in IHL is essential for the application of many of the rules deriving from the principles of distinction, proportionality and precaution, which afford critical protection to civilians and civilian objects. For many years, the ICRC has taken the position that an operation designed to disable a computer or a computer network during an armed conflict constitutes an attack as defined in IHL whether the object is disabled through destruction or in any other way. This view is also reflected in the positions of a number of States.

While many of the general rules on the conduct of hostilities are limited to acts amounting to attacks as defined in IHL, some IHL rules governing the conduct of hostilities apply to a broader set of operations. IHL includes a few rules that apply to all "military operations", such as the obligation to take constant care to spare civilians and civilian objects. Moreover, IHL defines specific rules protecting certain categories of persons and objects, such as objects indispensable to the survival of the civilian population, medical services, and humanitarian relief operations. The protection they afford goes beyond the general protection afforded to civilians and civilian objects.

The protection of data against malicious cyber operations during armed conflict is becoming increasingly important because data is an essential component of the digital domain and a cornerstone of life in many societies. In the ICRC's view, the conclusion that cyber operations designed or expected to delete or tamper with essential civilian data would not be prohibited by IHL in today's ever more cyber-reliant world seems difficult to reconcile with the object and purpose of this body of norms, and raises significant concern.

In order to protect critical civilian infrastructure that relies on cyberspace, it is also crucial to protect the infrastructure of cyberspace itself. It is traditionally understood that a civilian object may become a military objective when its use for military purposes is such that it fulfils the definition of military objective even if it is simultaneously used for civilian purposes. However, a party to a conflict that considers carrying out an attack against cyberspace infrastructure must analyze which distinct parts of the infrastructure make an effective contribution to military action, and whether their destruction or neutralization would, in the circumstances ruling at the time, offer a definite military advantage. Furthermore, this party must take all feasible precaution to avoid or at last minimize incidental civilian harm, including harm caused by indirect or reverberating effects, and must refrain from carrying out the attack if such harm may be expected to be excessive.

Fourth, in light of the particular challenges that the characteristics of cyberspace pose to the interpretation and application of some IHL principles in the conduct of hostilities, parties to armed conflicts who choose to develop, acquire or adopt weapons, means or methods of warfare relying on cyber technology need to exercise care in doing so. While legal reviews of new weapons, means and methods of warfare are mandatory for States party to AP I, legal reviews are critical for all States to ensure that their armed forces only use weapons, means or methods of warfare that comply with the State's obligations under IHL.

To conclude, recognizing that IHL applies in cyberspace and engaging in discussions on how it addresses the various challenges posed by the specific characteristics of the cyber domain and whether existing law is adequate and sufficient does not exclude that new rules might be useful or even needed. In our view, the answer to this question depends notably on how States interpret existing IHL obligations. If narrow interpretations are adopted, significant gaps in the protection of civilian populations and infrastructure may arise, and the existing legal framework might need strengthening. If new rules are developed, however, in our view it is critical that they build on and strengthen the legal framework that already exists – in particular IHL.

International Review of the Red Cross (2020), **102** (913), 335–365.
Digital technologies and war
doi:10.1017/S1816383121000023

IRRC

The application of the principle of distinction in the cyber context: A Chinese perspective

Zhixiong Huang and Yaohui Ying*
Zhixiong Huang is a Changjiang Outstanding Young Scholar Professor at the Wuhan University Institute of International Law/Institute for Cyber Governance, and a Research Fellow at the East China University of Political Science's Shanghai Key Innovation Team – Research of Legal Safeguard Mechanisms for the Belt and Road Construction. Email: fxyhzx@whu.edu.cn.

Yaohui Ying is a PhD candidate at Wuhan University Law School, China. Email: yingyaohui@whu.edu.cn.

Abstract

Up to now, the Chinese government has only made very general comments on the application of international humanitarian law to cyberspace. There are indeed Chinese academic papers concerning this issue, but the discussion of the principle of distinction is limited both in length and in academic depth. Compared with the West, research by Chinese scholars on this topic is still in a relatively preliminary stage. At present, there is no specific deconstruction or clarification of the application of the principle of distinction in cyberspace in Chinese academia. As

* This research is supported by the Major Projects of National Social Science Fund of China (Grant No. 20&ZD204). The authors are grateful to all the editors and anonymous referees for their useful suggestions and to Eric Jensen, Kubo Mačák, Ignacio de la Rasilla del Moral, Jinyuan Su, Nicole Hogg and Nicholas Tsagourias for their helpful comments on earlier drafts of this article. An earlier draft of this article was submitted to the workshop "Law in Today's Hybrid Armed Conflicts" held by Brigham Young University in February 2019, and all the feedback and comments from participants are most sincerely appreciated.

the first paper written by Chinese scholars specifically devoted to this question, this piece provides a different perspective by injecting the positions of Chinese officials and the views of Chinese scholars. The authors aim to clarify whether the existing rules are still completely applicable in the cyber context, and if needed, to find out what kind of improvements and clarifications can be made. Weighing in on these debates, we argue that despite the potential technical challenges and uncertainties, the principle of distinction should be applied to cyberspace. It should also be carefully re-examined and clarified from the standpoint of preventing over-militarization and maximizing the protection of the interests of civilians. For human targets, the elements of combatant status identified in customary international law and relevant treaties are not well suited to the digital battlefield. Nevertheless, cyber combatants are still obligated to distinguish themselves from civilians. In applying the principle of distinction, we argue that it makes more sense to focus on substantive elements over formal elements such as carrying arms openly or having a fixed distinctive sign recognizable at a distance. In interpreting "direct participation in hostilities", the threshold of harm requires an objective likelihood instead of mere subjective intention; the belligerent nexus should be confirmed, and the causal link should be proximate. Applying the "cyber kill chain" model by analogy helps us to grasp the whole process of direct participation in hostilities during cyber warfare. For non-human targets, all military objectives must cumulatively fulfil both the "effective contribution" and "definite military advantage" criteria, which are equally indispensable. The same requirements apply to dual-use objects. Furthermore, certain data should fall within the ambit of civilian objects.

Keywords: China, principle of distinction, cyberspace, cyber combatant, military objective, data.

Introduction

Up to now, the Chinese government has not been clear about the application of international humanitarian law (IHL)[1] to cyberspace. There have been some

1 In order to avoid confusion, a note is made at this point to clarify two terminologies, "law of armed conflict" (LOAC) and "international humanitarian law". There are some concerns about the inaccurate use of these two terms. Some think they have essentially the same meaning and can be used interchangeably, e.g. "the law of armed conflict, also known as international humanitarian law, includes principles such as distinction between military and civilian targets" (International Committee of the Red Cross (ICRC), *The Law of Armed Conflict: Basic Knowledge*, Geneva, June 2002, p. 2, available at: www.icrc.org/eng/assets/files/other/law1_final.pdf), while others render "international humanitarian law" as a potentially narrower concept that relates only to the laws in armed conflict that are designed to regulate the treatment of persons – civilian or military, wounded or active – in armed conflicts (Mary O'Connell, "Historical Development and Legal Basis", in Dieter Fleck (ed.), *The Handbook of International Humanitarian Law*, 3rd ed., Oxford University Press, Oxford, 2013, p. 11). There are also some critiques regarding the melding of battlefield laws and humanitarian goals, e.g. "a

preliminary debates concerning IHL in cyberspace among Chinese scholars,[2] especially those with a military background,[3] but the discussion of the principle of distinction in cyberspace is limited both in length and in academic depth. Compared with the West, research by Chinese scholars on this issue is still in a relatively preliminary stage, and some doctoral theses on the application of IHL to cyberspace are under way. At present, there is no specific deconstruction or clarification of the application of the principle of distinction in cyberspace in Chinese academia.

As the first paper written by Chinese scholars specifically devoted to the application of the principle of distinction in cyber warfare, this piece provides a

possible disadvantage of the term [IHL] is that it could be thought to exclude some parts of the laws of war (such as the law on neutrality) whose primary purpose is not humanitarian" (Jean Pictet, *Humanitarian Law and the Protection of War Victims*, A. W. Sijthoff, Leiden, 1975, p. 11). The International Law Commission distinguishes between LOAC and IHL, with the former governing the conduct and consequences of armed conflict while the latter forms part of the former and constitutes the *lex specialis* governing the conduct of hostilities (para. 4 of the Commentary to Art. 2 of the Draft Articles on the Effects of Armed Conflicts on Treaties, *ILC Yearbook*, Vol. 2, Part 2, 2011). For more detailed discussion on the terminology, see Gary D. Solis, *The Law of Armed Conflict: International Humanitarian Law in War*, Cambridge University Press, Cambridge and New York, 2010, pp. 22–26. Chinese textbooks and papers generally hold the view that the term IHL has evolved from the law of war or LOAC, and thus treat them as synonymous; see, for example, 朱文奇, 何谓国际人道法, 武大国际法评论, 2003, 1 (Wenqi Zhu, "What Is International Humanitarian Law?", *Wuhan University International Law Review*, Vol. 1, 2003, only available in Chinese). For the purpose of this paper, the term "IHL" will be used generally, while the term "law of armed conflict" is used when the cited sources use that particular term.

2 See, for example, Li Zhang, "A Chinese Perspective on Cyber War", *International Review of the Red Cross*, Vol. 94, No. 886, 2012, p. 804, available at: https://international-review.icrc.org/sites/default/files/irrc-886-zhang.pdf (all internet references were accessed in January 2021); Longdi Xu, "The Applicability of the Laws of War to Cyberspace: Exploration and Contention", 2014, p. 7, available at: www.gov.uk/government/publications/the-applicability-of-the-laws-of-war-to-cyberspace-exploration-and-contention; Chris Wu, "An Overview of the Research and Development of Information Warfare in China", in Edward Halpin, Philippa Trevorrow, David Webb and Steve Wright (eds), *Cyberwar, Netwar and the Revolution in Military Affairs*, Palgrave Macmillan, London, 2006; 朱　欣, 信息网络战的国际法问题研究, 河北法学, 2009, 27(01) (Lixin Zhu, "Research on the International Law of Information Network Operations", *Hebei Law Science*, Vol. 27, No. 1, 2009, only available in Chinese); 姜世波, 网络攻击与战争法的适用, 武大国际法评论, 2013, 16(02) (Shibo Jiang, "War by Internet Cyber Attack and the Application of the Law of War", *Wuhan University International Law Review*, Vol. 16, No. 2, 2013, only available in Chinese); 李伯军, 论网络战及战争法的适用问题, 法学评论, 2013, 31(04) (Bojun Li, "On Cyber Warfare and the Application of the Law of War", *Law Review*, Vol. 31, No. 4, 2013, only available in Chinese); 朱　欣, 平战结合与网络空间国际规则制定, 信息安全与通信保密, 2018(07) (Lixin Zhu, "Competition for International Rules in Cyberspace under the Combination of Peacetime and Wartime", *Information Security and Communications Privacy*, No. 7, 2018).

3 王海平, 武装冲突法研究进展及需要关注的问题, 当代法学, 2012, 26(05) (Haiping Wang, "The Research Progress of the Law of Armed Conflict and the Issues Needing Attention", *Contemporary Law Review*, Vol. 26, No. 5, 2012, only available in Chinese); 李　, 鲁笑英.浅析信息化战争条件下武装冲突法所面临的问题, 西安政治学院学报, 2012, 25(01) (Li Li and Xiaoying Lu, "A Brief Analysis of the Problems Faced by the Law of Armed Conflict under the Condition of Information-Based Warfare", *Journal of Xi'an Politics Institute of PLA*, Vol. 25, No. 1, 2012, only available in Chinese); 朱雁新, 计算机网络攻击之国际法问题研究, 中国政法大学, 2011 (Yanxin Zhu, "The Research on the International Issues of Computer Network Attack", doctoral diss., China University of Political Science and Law, 2011, only available in Chinese); 张天舒, 从 "塔林手册" 看网络战争对国际法的挑战, 西安政治学院学报, 2014, 27(01) (Tianshu Zhang, "The Challenges of Cyber Warfare to International Law: From the Perspective of The Tallinn Manual on the International Law Applicable to Cyber Warfare", *Journal of Xi'an Politics Institute of PLA*, Vol. 27, No. 1, 2014, only available in Chinese).

different perspective by injecting the positions of Chinese officials and the views of Chinese scholars into the discussion. The authors hold the view that although States have vastly differing interpretations of exactly how IHL applies to cyberspace, the core principle of distinction is definitely applicable in cyberspace. This paper aims to clarify whether the existing rules are still completely applicable in cyber warfare, and if needed, to find out what kind of improvements and clarifications can be made. Given this, the first part introduces the *status quo* of the application of IHL to cyberspace and illuminates the Chinese official attitude alongside Chinese academic opinions on this issue. Subsequently, the second part reviews the concept of the principle of distinction and points out the contentious challenges of its application in the cyber context. Applying the persons–objects dichotomy, the third and fourth parts examine the substantive legal challenges involved and inject the relevant Chinese views. From the perspective of human targets, the third part analyzes the application of traditional criteria for defining who can be attacked in the cyber battlefield, identifies the relevant obstacles and makes corresponding suggestions. The fourth part focuses on non-human targets and discusses what can be attacked in cyber warfare – namely, what constitutes a military objective. It further addresses the Chinese scholarship on whether digital data *per se* is an object. The final part offers some preliminary concluding observations.

It is beyond doubt that the peaceful use of the cyberspace domain is of great importance to the common well-being of mankind. Fortunately, to date the world has remained free of any catastrophic mass-casualty cyber attacks, or equivalent catalysts for war such as a "cyber Pearl Harbor"[4] situation. However, the increasingly disturbing occurrence of belligerent cyber incidents, such as the inclusion of cyber means and methods in armed conflicts, are forcing us to pay close attention to the application of IHL in cyberspace.

Cyber warfare,[5] despite having the potential to allow for some level of anonymity on an *ad hoc* basis and a sense of interconnectedness, is still a kind of

4 James J. Wirtz, "The Cyber Pearl Harbor", *Intelligence and National Security*, Vol. 32, No. 6, 2017; James J. Wirtz, "The Cyber Pearl Harbor Redux: Helpful Analogy or Cyber Hype?", *Intelligence and National Security*, Vol. 33, No. 5, 2018; US Department of Defense (DoD), "Remarks by Secretary Panetta on Cybersecurity to the Business Executives for National Security, New York City", 12 October 2012, available at: https://content.govdelivery.com/accounts/USDOD/bulletins/571813.

5 In this article, the term "cyber warfare" is understood as "means and methods of warfare that rely on information technology and are used in the context of an armed conflict". See Jakob Kellenberger, "International Humanitarian Law and New Weapon Technologies, 34th Round Table on Current Issues of International Humanitarian Law, Sanremo, Italy, 8–10 September 2011: Keynote Address by Dr Jakob Kellenberger", *International Review of the Red Cross*, Vol. 94, No. 886, 2012, available at: https://international-review.icrc.org/sites/default/files/irrc-886-kellenberger-spoerri.pdf. For some Chinese scholars, cyber warfare is a special form of information warfare and is a new means or method of warfare. Information warfare refers to a series of hostile activities carried out by belligerent parties in order to maintain their right to acquire, control and use information. Its connotation and extension are broader than cyber warfare and can include cyber warfare, intelligence warfare, electronic warfare, psychological warfare, etc. Cyber warfare refers to the process of disrupting, destroying or threatening the other belligerent parties' information and network systems while ensuring the security of one's own information and network systems through computer networks. See, for example, B. Li, above note 2. Some argue that the main question expressed by the concept of cyber warfare is whether

warfare. As such, multilateral discussions have been ongoing for over a decade now concerning whether IHL – as "the set of rules that seeks to limit the effects of armed conflicts"[6] – applies to the cyberspace domain. No consensus has been reached yet. There seemed to have been a glimmer of hope in the report of the 2014/15 United Nations Group of Governmental Experts (UN GGE) on Developments in the Field of Information and Telecommunications in the Context of International Security, since it had already mentioned the applicability of the principles of distinction and proportionality in cyberspace:[7] the wording "international legal principles, including the principle of distinction"[8] is seen as a compromise because some States (presumably including China) do not wish to refer directly to the term IHL.[9] However, the subsequent 2016/17 UN GGE failed to arrive at a consensus, and one of the controversial issues concerned the application of IHL in cyberspace.[10] With the adoption of two separate (some may say competing) resolutions by the First Committee of the General Assembly in 2018,[11] the future for States' consensus on IHL in cyberspace seems more and more uncertain and confusing.

In an ideal world, it seems that once a situation has reached the threshold of an armed conflict, the application of *jus in bello* rules to cyberspace should be nothing more than putting old wine into a new bottle. If cyber warfare is merely a new means or methods of warfare, then the existing *jus in bello* rules would automatically apply, and there is nothing mysterious or inscrutable about it. However, the reality often runs counter to the ideal. Due to the huge difference between cyber and traditional battlefields, many existing rules appear to be rather confusing in cyber warfare, and must be re-conceptualized. This is especially true in the case of the principle of distinction. For instance, an important issue relating to this principle is that of distinguishing between cyber combatants and civilians. Combatants are obligated to carry arms openly and to have a fixed

cyber attackers "armed" with keyboards, computer viruses and malware can become (or have become) a new means or method of warfare. See 黄志雄主编, 网络空间国际规则新动向: "塔林手册 2.0 版"研究文集, 社会科学文献出版社, 2019: 301 (Zhixiong Huang (ed.), *New Trends in International Rules for Cyberspace: Collection of Papers on Tallinn Manual 2.0*, Social Sciences Academic Press, China, 2019, p. 301); 黄志雄, 国际法视角下的 "网络战" 及中国的对策——以诉诸武力权为中心, 现代法学, 2015, 37(05) (Zhixiong Huang, "International Legal Issues concerning 'Cyber Warfare' and Strategies for China: Focusing on the Field of *Jus ad Bellum*", *Modern Law Science*, Vol. 37, No. 5, 2015).

6 See ICRC, "War and Law", available at: www.icrc.org/en/war-and-law.

7 See UN GGE, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, UN Doc. A/70/174, 22 July 2015, para. 28, available at: www.un.org/ga/search/view_doc.asp?symbol=A/70/174.

8 *Ibid.*

9 Michael N. Schmitt and Liis Vihul, "International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms", *Just Security*, 30 June 2017, available at: www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms/.

10 See, for example, *ibid.*; Arun Mohan Sukumar, "The UN GGE Failed. Is International Law in Cyberspace Doomed as Well?", *Lawfare*, 4 July 2017, available at: https://lawfareblog.com/un-gge-failed-international-law-cyberspace-doomed-well.

11 See "The United Nations Doubles Its Workload on Cyber Norms, and Not Everyone Is Pleased", *Council on Foreign Relations Blog*, 15 November 2018, available at: www.cfr.org/blog/united-nations-doubles-its-workload-cyber-norms-and-not-everyone-pleased. The two resolutions are sponsored by Russia (UN Doc. A/C.1/73/L.27/Rev.1) and the United States (UN Doc. A/C.1/73/L.37) respectively.

distinctive sign recognizable at a distance.[12] This is apparently not practical in the cyber context, where anonymity is often the norm and it is impossible to tell who is sitting in front of the computer that is implementing an attack. The rules were drafted in an era when warfare involved a certain amount of physical proximity between opposing forces; for the most part, combatants could see one another and hence distinguish between combatants and non-combatants, friends and foes.[13] When it comes to civilians who directly participate in hostilities,[14] the question becomes even more confusing. It is highly possible for unorganized individuals to launch cyber attacks against an adversary; the typical example would be a group of hacktivists performing a distributed denial-of-service (DDoS) attack for patriotic or ideological reasons. For instance, the anonymous cyber attack against Estonian essential infrastructures, telecommunications, DNS servers, websites and email servers in 2007 seemed to have followed a political row over the relocation of a Soviet "Monument to the Liberators of Estonia", which represents the USSR's victory over Nazism, from the centre of Tallinn to a military cemetery on the outskirts of the city.[15] Is the person who inputs the malicious code, or the person who writes (but does not execute) the code, or the person who gives the order for the code to be written in the first place, the one directly taking part in hostilities?

As the country with the largest number of netizens and one which suffers from frequent cyber attacks,[16] China has been very active in promoting the rule of law in cyberspace. Yet, while it has been a State party to the Geneva Conventions[17]

12 Geneva Convention (III) relative to the Treatment of Prisoners of War of 12 August 1949, 75 UNTS 135 (entered into force 21 October 1950) (GC III), Art. 4(A)(2); 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) (AP I), Art. 44(3); Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, Vol. 1: *Rules*, Cambridge University Press, Cambridge, 2005 (ICRC Customary Law Study), pp. 14–17, available at: https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1.

13 Heather Harrison Dinniss, "Participants in Conflict – Cyber Warriors, Patriotic Hackers and the Laws of War", in Dan Saxon (ed.), *International Humanitarian Law and the Changing Technology of War*, Martinus Nijhoff, Boston, MA, and Leiden, 2013, p. 256; Heather Harrison Dinniss, *Cyber Warfare and the Laws of War*, Cambridge University Press, Cambridge, 2012, p. 145.

14 The ICRC Customary Law Study, above note 12, Rule 6, stipulates that civilians are protected against attack unless and for such time as they take a direct part in hostilities. For substantive discussion about "direct participation in hostilities", see Nils Melzer, *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law*, ICRC, Geneva, 2009 (Interpretive Guidance).

15 For a detailed description of the cyber attack against Estonia in 2007, see "Cyber Attacks against Estonia (2007)", *International Cyber Law in Practice: Interactive Toolkit*, NATO Cooperative Cyber Defence Centre of Excellence (CCD COE), available at: https://cyberlaw.ccdcoe.org/wiki/Cyber_attacks_against_Estonia_(2007); Eneken Tikk, Kadri Kaska and Liis Vihul, *International Cyber Incidents: Legal Considerations*, CCD COE, Tallinn, 2010, pp. 15–16, 31.

16 Chinese Academy of Cyberspace Studies (ed.), *China Internet Development Report 2017*, Springer, Berlin, 2019, p. 107; 国家互联网应急中心，2020 年上半年我国互联网网络安全监测数据分析报告，2020 (National Computer Network Emergency Response Technical Team/Coordination Centre of China, *Analysis Report of China's Internet Network Security Monitoring Data in the First Half of 2020*, 2020, only available in Chinese), available at: https://tinyurl.com/y2lpzdh4; Ministry of Foreign Affairs of the People' Republic of China, "Foreign Ministry Spokesperson Wang Wenbin's Regular Press Conference on September 29, 2020", available at: https://tinyurl.com/y4xolw3g.

and Additional Protocols I and II to the Geneva Conventions (AP I and AP II)[18] for many years, China has not had much enthusiasm on the issue of IHL in cyberspace and has always avoided addressing the issue of cyber warfare and the law applicable to it.[19]

China's reluctance to discuss the IHL issue in depth has been evidenced on many occasions. For instance, in its recent submission to the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security, China stated that "the applicability of the law of armed conflicts and *jus ad bellum* needs to be handled with prudence";[20] this suggests that China, for some (maybe political) reason, does not want to discuss the details of IHL in cyberspace and therefore delays any clarification of the issue. Instead of specifying its position and rationale, China has only repeatedly affirmed that "the lawfulness of cyber warfare should not be recognized under any circumstance".[21] This resistant attitude is prominent in the speech given by the Chinese delegate at the 2019 Annual Session of the Asian–African Legal Consultative Organization (AALCO):

> China sticks to the principle of peaceful use of cyberspace and firmly opposes … cyber warfare or [the] cyber arms race. … Without state practice, we should be very prudent on the discussion of application of humanitarian law in so called "cyber wars". The reason is very simple but fundamental: firstly, no cyber wars shall be permitted; and secondly, cyber war will be a totally new form of high-tech war. Given the "digital gap" between developing and … developed countries, developing countries in general will be in a disadvantaged position in the discussion and development of such rules, [and] it will be difficult to ensure the rules are fair and equitable.[22]

China attaches great importance to the peaceful use of cyberspace, and asserts that too much discussion of the application of IHL would have potential negative impacts on international peace and security, aggravating an arms race and the militarization of cyberspace. For instance, China has expressed its criticism by saying that "this military paradigm"[23] disregards the principle of non-use of

17 China's date of ratification/accession to the Geneva Conventions is 28 December 1956. See the ICRC Treaty Database, available at: https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/vwTreatiesByCountrySelected.xsp?xp_countrySelected=CN.

18 China's date of ratification/accession to AP I and AP II is 14 September 1983. See *ibid.*

19 Binxin Zhang, "Cyberspace and International Humanitarian Law: The Chinese Approach", in Suzannah Linton, Tim McCormack and Sandesh Sivakumaran (eds), *Asia-Pacific Perspectives on International Humanitarian Law*, Cambridge University Press, Cambridge, 2019, p. 323.

20 See "China's Submissions to the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security", p. 6, available at: www.un.org/disarmament/wp-content/uploads/2019/09/china-submissions-oewg-en.pdf.

21 *Ibid.*

22 AALCO, *Verbatim Record of Discussions: Fifty-Eighth Annual Session*, AALCO/58/DAR ES SALAAM/2019/VR, Dar es Salaam, 21–25 October 2019, available at: www.aalco.int/Final%20Verbatim%202019.pdf.

23 AALCO, *Verbatim Record of Discussions: Fifty-Fourth Annual Session*, AALCO/54/BEIJING/2015/VR, Beijing, 13–17 April 2015.

force[24] and may affect strategic trust between countries and increase the risk of inter-State misperception and conflict.[25] In this context, it is not surprising that the government of China has not been clear about the application of the principle of distinction in cyberspace. China's conservative attitude is understandable to some extent. Firstly, there is no widely recognized national practice that constitutes a cyber attack; secondly, due to the hysteretic nature of law, IHL in cyberspace should not be determined too early.[26] The existing negative attitude of the Chinese government on this issue may also be a delaying tactic in the process when China has not come up with a self-explanatory plan. From the authors' point of view, there is no legal obstacle to the application of IHL in cyberspace, especially the principle of distinction. It is undeniable that cyber warfare has already taken place and will continue to do so. Whether China likes it or not, it will probably have to express its stance on IHL in cyberspace.

The principle of distinction and the challenge of applying it to cyberspace

Having introduced the *status quo* of the application of IHL in cyberspace, China's official attitude, and some Chinese scholars' views on this point as the starting point of our analysis, it is now time to review the principle of distinction *per se* and summarize the contentious challenges of its application in the cyber context. The principle of distinction, according to the International Court of Justice (ICJ) in its *Legality of the Threat or Use of Nuclear Weapons* advisory opinion, is a cardinal principle of the law of armed conflict and has achieved the status of customary international law.[27] Article 48 of AP I stipulates that parties to a conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives, and shall accordingly direct their operations only against military objectives.[28]

Generally speaking, the principle of distinction takes a two-pronged approach to the regulation of hostilities. It prohibits indiscriminate means and methods of warfare, and it also regulates the use of those means and methods that are lawful – meaning that a distinction shall be made between military objectives and combatants, on the one hand, and other persons and objects that should be respected and protected, on the other. Indiscriminate attacks are prohibited.[29]

24 Xinmin Ma, "What Kind of Internet Order Do We Need?", *Chinese Journal of International Law*, Vol. 14, No. 2, 2015. Xinmin Ma served as deputy director of the Department of Treaty and Law of the Ministry of Foreign Affairs of China from 2014 to 2019.

25 AALCO, *Verbatim Record of Discussions: Fifty-Fifth Annual Session*, AALCO/55/NEW DELHI (HEADQUARTERS)/2016/VR, New Delhi, 17–20 May 2016.

26 For more explanation on China's attitude towards IHL, see B. Zhang, above note 19.

27 ICJ, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, 8 July 1996, *ICJ Reports 1996*, p. 266.

28 AP I, Art. 48; ICRC Customary Law Study, above note 12, Rules 1, 7, pp. 3, 25.

29 AP I, Art. 51(4); ICRC Customary Law Study, above note 12, Rule 11, p. 37.

An "attack" triggers a wide array of legal protections concerning distinction, especially those contained in Articles 49–58 of AP I. Therefore, in order to clarify exactly how the principle of distinction can be applied to cyberspace, a proper definition of "cyber attack" is a prerequisite. There have been some in-depth and meaningful academic discussions on what constitutes a cyber attack.[30] The most widely accepted definition takes a consequence-based approach. For example, the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Tallinn Manual 2.0) defines a cyber attack as "a cyber operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to objects".[31] We take this definition in this article.[32] No apparent legal provision explicitly bans or addresses the use of cyber warfare, as distinct from other forms of warfare. IHL is currently silent on distinction matters in cyber warfare, and some scholars therefore argue that the existing treaty-based framework is ill-suited to cope with it; this aspect of virtual war negatively affects the application of the principle of distinction.[33] One reason for this, as some scholars contend,[34] is that civilian and military infrastructures are not only closely interrelated and interconnected but are, in fact, one and the same thing. This assertion can lead to conclusions that pose significant obstacles to the application of the principle of distinction. If most components of cyberspace – such as fibre-optic cables, satellites, routers and nodes – are dual-use objects, simultaneously serving both military and civilian purposes, the classification of these objects can be problematic, leading to tricky

30 See Marco Rossini, *Cyber Operations and the Use of Force in International Law*, Oxford University Press, Oxford, 2014, pp. 178–182; William H. Boothby, "Where Do Cyber Hostilities Fit in the International Law Maze?", in Hitoshi Nasu and Robert McLaughlin (eds), *New Technologies and the Law of Armed Conflict*, Springer, Berlin, 2014, pp. 60–62; Knut Dörmann, "Applicability of the Additional Protocols to Computer Network Attacks", paper presented at the International Expert Conference on Computer Network Attacks and the Applicability of International Humanitarian Law, Stockholm, 17–19 November 2004; Cordula Droege, "Get Off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians", *International Review of the Red Cross*, Vol. 94, No. 886, 2012; Michael N. Schmitt, "The Law of Cyber Warfare: Quo Vadis?", *Stanford Law and Policy Review*, Vol. 25, No. 2, 2014.

31 Michael N. Schmitt (ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, Cambridge University Press, Cambridge, 2017 (Tallinn Manual 2.0), Rule 92, p. 415.

32 The consequence-based approach is very useful as it switches the focus from the means and nature of an act to the effect and consequence of an act, thus fulfilling the requirement of "violence" and keeping the provision dynamic and evolutive. However, the present authors still have two concerns. The first is that from a practical perspective, the assessment of the damage turns out to be extremely tricky, especially when the consequences are mostly indirect. The second concern is that the consequence-based approach limits the notion of attack so as to exclude those operations that result in severe and disruptive non-physical harm. Similar concerns can be found in ICRC, *International Humanitarian Law and Cyber Operations during Armed Conflicts*, Geneva, November 2019 (ICRC Cyber Operations Paper), pp. 7–8. The ICRC has also mentioned that an overly restrictive understanding of the notion of attack would be difficult to reconcile with the object and purpose of the rules on the conduct of hostilities, which is to ensure the protection of the civilian population and civilian objects against the effects of hostilities. See ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, 32IC/15/11, October 2015 (ICRC Challenges Report 2015), p. 41.

33 See Jeffrey Kelsey, "Hacking into International Humanitarian Law: The Principles of Distinction and Neutrality in the Age of Cyber Warfare", *Michigan Law Review*, Vol. 106, No. 7, 2008, pp. 1429–1430.

34 Robin Geiss and Henning Lahmann, "Cyber Warfare: Applying the Principle of Distinction in an Interconnected Space", *Israel Law Review*, Vol. 45, No. 3, 2012, pp. 381, 383.

issues concerning the principle of proportionality.[35] At the same time, the classification of individuals as combatants or civilians is not always clear given that the mounting phenomenon of civilianization of war,[36] characterized by the increasing use of sophisticated cyber technologies, has blurred the contours. Militaries and civilian enterprises are communicating, cooperating and integrating at an unprecedented depth.[37] For instance, China has twice included the strategy of civil–military integration in its white papers.[38] Also, the attribution of responsibility presents difficulties;[39] while it is mostly easy to see where a missile was launched from, the deployment of cyber operations doesn't create smoke plumes.

Several scholars have made a rigorous effort to research how the principle of distinction applies in cyber warfare,[40] and several states, such as the United States[41] and Denmark,[42] have added the principle of distinction's application in cyber warfare into their respective Military Manuals. It is generally agreed, for instance, that an attack does not have to be kinetic for IHL rules to apply to it; that indiscriminate attacks[43] are prohibited; and that if an attack does not specifically target any particular military persons or objects, it shall never be permitted. This could be the case with a computer virus, if it can spread uncontrollably from military systems to connected civilian systems. While there is a consensus that a distinction must be made between military objectives/combatants and civilian objects/civilians, when it comes to the more practical level of exactly what constitutes a military objective and who is a combatant in a cyber armed conflict, the question becomes extremely controversial. Moreover, as raised by one Chinese scholar, the non-lethal underlying feature of cyber means and methods makes traditionally protected objects and individuals more

35 The principle of distinction provides that only military objectives may be directly targeted in armed conflict. However, an attack on a legitimate military objective may sometimes cause incidental damage to civilian persons or objects. These harmful side effects are regulated by the principle of proportionality, which prohibits attacks that may be expected to cause injury to civilian life or property that is excessive in relation to the anticipated military advantage. A clear statement of the principle of proportionality can be found in AP I, Art. 51(5)(b). See also Jonathan Crowe and Kylie Weston-Scheuber, *Principles of International Humanitarian Law*, Edward Elgar, Cheltenham, 2013, pp. 55–57.

36 "Civilians play an increasingly important and complex role in armed conflicts, both as victims and perpetrators." This overall trend is called "civilianization" in Andreas Wenger and Simon J. A. Mason, "The Civilianization of Armed Conflict: Trends and Implications", *International Review of Red Cross*, Vol. 90, No. 872, 2008.

37 L. Zhu, "Competition for International Rules in Cyberspace", above note 2, p. 40.

38 State Council Information Office of the People's Republic of China (SCIO), *China's National Defense in the New Era*, Beijing, July 2019, available at: www.scio.gov.cn/zfbps/32832/Document/1660325/1660325.htm; SCIO, *China's Military Strategy*, Beijing, May 2015, available at: www.scio.gov.cn/zfbps/ndhf/2015/Document/1435159/1435159.htm.

39 See ICRC Cyber Operations Paper, above note 32, pp. 8–9.

40 See, for example, J. Kelsey, above note 33, p. 1427; Yoram Dinstein, "The Principle of Distinction and Cyber War in International Armed Conflicts", *Journal of Conflict and Security Law*, Vol. 17, No. 2, 2012, p. 261; Michael N. Schmitt, "Wired Warfare: Computer Network Attack and *Jus in Bello*", *International Review of the Red Cross*, Vol. 84, No. 846, 2002, p. 365.

41 DoD, *Law of War Manual*, Washington, DC, 12 June 2015, pp. 985–999.

42 Danish Ministry of Defence, Defence Command Denmark, *Military Manual on International Law Relevant to Danish Armed Forces in International Operations*, Copenhagen, September 2016.

43 AP I, Art. 51(4).

vulnerable in cyber warfare than in conventional warfare. This will lead to confusion in evaluating the legitimacy of cyber operations and make the principle of distinction more frequently violated in cyber military operations.[44] Given the significance of the principle of distinction on the cyber battlefield, it is necessary to clarify whether the existing rules are still completely applicable in cyber warfare, and to find out what kind of improvements and clarifications can be made.

The principle of distinction concerning human targets in cyber warfare

The principle of distinction follows a persons–objects dichotomy to define the nature of the target. No matter how cyber technology evolves, the perpetrator of a hostile act is still a person, and even when planting viruses or attacking firewalls in ways that look like mere keystrokes and mouse clicks, the persons–objects dichotomy, which defines "who" and "what" can be attacked, still applies. This part of the article will deal with the issue of who can be lawfully attacked in the cyber context. The foundational principle is that civilians shall not be the object of attack.[45] The principle of distinction assumes that belligerents can clearly distinguish between civilians and combatants; the anonymity of cyberspace, however, makes this assumption hard to maintain.

Every combatant is a former civilian, and any civilian may convert himself into a combatant,[46] either by being conscripted or volunteering to join the armed forces of a belligerent party, or by taking a direct part in hostilities (this leads to the loss of protected status while doing so),[47] or by becoming part of a *levée en masse*, a concept that allows the transition from civilians to lawful combatants.[48] The authors will not address *levée en masse* here, because this concept requires the physical invasion of national territory and the involvement of a large segment of population,[49] which is almost impossible by cyber means.[50]

Due to the advantages of easy denial of State responsibility and low cost, "the majority of cyber operations are outsourced to civilian cyber experts".[51] In

44 陈鹏飞, 论当代武装冲突法面临的挑战, 西安政治学院学报, 2014, 27(05) (Pengfei Chen, "Analysis of the Challenges to Contemporary Armed Conflict Law", *Journal of Xi'an Politics Institute of PLA*, Vol. 27, No. 5, 2014, only available in Chinese).
45 AP I, Art. 51(2); ICRC Customary Law Study, above note 12, Rule 6, pp. 19–24.
46 Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict*, Cambridge University Press, Cambridge, 2016, p. 174.
47 AP I, Art. 51(3); ICRC Customary Law Study, above note 12, Rule 6, pp. 20–21; Interpretive Guidance, above note 14, pp. 41–68.
48 GC III, Art. 4A(6); ICRC Customary Law Study, above note 12, Rule 106, pp. 386–387, and in particular Rule 5, which explains that members of a *levée en masse* are an exception to the definition of civilians in that although they are not members of the armed forces, they qualify as combatants.
49 GC III, Art. 4A(6).
50 Tallinn Manual 2.0, above note 31, Rule 88, p. 409.
51 Elizabeth Mavropoulou, "Targeting in the Cyber Domain: Legal Challenges Arising from the Application of the Principle of Distinction to Cyber Attacks", *Journal of Law and Cyber Warfare*, Vol. 4, No. 2, 2015, p. 78.

light of this trend, there is a high probability that, with the exception of cyber units incorporated into the regular armed forces, "many of the personnel substantively involved in cyber operations may actually be civilians".[52] Could a patriotic hacker or computer scientist thus become the object of an attack? The answer depends on the interpretation of "direct participation in hostilities" in the context of cyber operations.

Who is a cyber combatant?

Civilians who directly participate in hostilities lose their protected status and are not entitled to combatant immunity; some scholars even argue that they are "unlawful"[53] combatants. IHL encourages a clear and reliable division between combatants and non-combatants, and this reflects the fundamental role played by the principle of distinction in this body of law. Combatants have the right to participate directly in hostilities[54] and are subsequently immune from prosecution for acts which are carried out in accordance with IHL;[55] thus, they are targetable. Cyber warfare is no exception to this. Since the definition of civilians is a purely negative one (civilians are persons who are not combatants[56]), the question of who is a cyber combatant becomes a critical issue.[57]

It has been seen that some States have established special sections within their armed forces responsible for cyber operations. For instance, the United States has established US Cyber Command (USCYBERCOM), which was elevated from a sub-unit of the US Strategic Command to the status of a Unified Combatant Command,[58] while Colombia has created an Armed Forces Joint Cyber Command, tasked with preventing and countering cyber threats or attacks affecting national values and interests.[59] The definition of cyber combatant is worthy of discussion because it not only involves the issue of who is a legitimate target, but also has an impact on who is entitled to prisoner of war (PoW) status if captured.

52 David Turns, "Cyber Warfare and the Notion of Direct Participation in Hostilities", *Journal of Conflict and Security Law*, Vol. 17 No. 2, 2012, p. 292; see also Michael N. Schmitt, "'Direct Participation in Hostilities' and 21st Century Armed Conflict", in Horst Fischer and Dieter Fleck (eds), Crisis *Management and Humanitarian Protection: Festschrift for Dieter Fleck*, BWV, Berlin, 2004, p. 527.

53 Y. Dinstein, above note 46, p. 44.

54 AP I, Art. 43(2).

55 H. Harrison Dinniss, "Participants in Conflict", above note 13, p. 254.

56 AP I, Art. 50(1); ICRC Customary Law Study, above note 12, Rule 5, pp. 17–19.

57 Vijay M. Padmanabhan, "Cyber Warriors in the Jus in Bello", *International Law Studies*, Vol. 89, 2013; Maurizio D'Urso, "The Cyber Combatant: A New Status for a New Warrior", *Philosophy and Technology*, Vol. 28, No. 3, 2015; Jake B. Sher, "Anonymous Armies: Modern 'Cyber-Combatants' and Their Prospective Rights under International Humanitarian Law", *Pace International Law Review*, Vol. 28, No. 1, 2016; Sean Watts, "The Notion of Combatancy in Cyber Warfare", paper presented at the 4th International Conference on Cyber Conflict, Tallinn, 5–8 June 2012.

58 Donald Trump, "Statement by President Donald J. Trump on the Elevation of Cyber Command", 18 August 2017, available at: www.whitehouse.gov/briefings-statements/statement-president-donald-j-trump-elevation-cyber-command/.

59 UN, *Developments in the Field of Information and Telecommunications in the Context of International Security: Report of the Secretary-General*, UN Doc. A/67/167, 23 July 2012, p. 5.

Combatants are basically members of the armed forces of a belligerent party – whether these forces are regular or irregular, and irrespective of belonging to the standing army or to reservist units – including paramilitary militias incorporated *de facto* into the armed forces. The specific task assigned to an individual within the military apparatus is irrelevant.[60]

The Geneva Conventions have enumerated five conditions which must be satisfied for lawful combatant status.[61] The first four are cumulative conditions set out by the Hague Regulations and Geneva Conventions for the applicability of PoW and lawful combatant status: (i) being under the command of a person responsible for his or her subordinates (organization); (ii) having a fixed distinctive sign recognizable at a distance; (iii) carrying arms openly; and (iv) conducting operations in accordance with the laws and customs of war (compliance).[62] These four conditions apply to members of other militias and members of other volunteer corps, but they are also implicit requirements for members of the armed forces of a party to the conflict. An additional condition may be implied from the Geneva Conventions, which is (v) belonging to a party to the conflict.[63]

The authors believe that elements (i), (iv), and (v) are substantive elements, while elements (ii) and (iii) are formal ones. Considering the fact that anonymity is the normal status in cyber warfare, it makes more sense to focus on the substantive elements instead of the formal ones.

The first element, that of organization, is essential in cyber warfare. This is more of a factual issue than a legal one, and this requirement reflects the presence of a responsible command and a hierarchical relationship.[64] If a cyber group does not have sufficient organization, typically a superior–subordinate structure, division of duties and accountability, and certain elements of discipline and supervision, its members cannot be lawful combatants and certainly would not be entitled to combatant immunity. Given that members of most cyber groups have the same intention but lack common discipline, the chances that an armed group which exists exclusively online will be sufficiently organized are slim.[65] For instance, if no consequence will occur when members of a group suddenly decide to stop or not to participate in cyber hostilities (it may be the case that cyber group members do not know each other at all), or the members of a group do not feel compelled to follow the orders of a commander, it is not reasonable to submit

60 Y. Dinstein, above note 46, p. 41.
61 H. Harrison Dinniss, *Cyber Warfare*, above note 13, p. 144.
62 Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of 12 August 1949, 75 UNTS 31 (entered into force 21 October 1950), Art. 13(2); Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea of 12 August 1949, 75 UNTS 85 (entered into force 21 October 1950), Art. 13 (2); GC III, Art. 4(A)(2); Geneva Convention (IV) relative to the Protection of Civilian Persons in Time of War of 12 August 1949, 75 UNTS 287 (entered into force 21 October 1950), Art. 4(2); H. Harrison Dinniss, *Cyber Warfare*, above note 13, p. 145.
63 GC III, Art. 4A(6); H. Harrison Dinniss, *Cyber Warfare*, above note 13, p. 145.
64 Y. Dinstein, above note 46, p. 39; International Criminal Tribunal for Rwanda (ICTR), *The Prosecutor v. Jean-Paul Akayesu*, Case No. ICTR-96-4-T, Judgment (Trial Chamber), 2 September 1998, para. 626.
65 Marco Roscini, *Cyber Operations and the Use of Force in International Law*, Oxford University Press, Oxford, 2014, p. 195.

that such a loosely organized group fulfils the element of organization. This is particularly true in the case of patriotic cyber groups.[66]

The fourth element, that of compliance with IHL, remains indispensable and has not changed markedly with the advent of computer network technology.[67] If combatants are themselves unwilling to respect IHL, they are prevented from relying on that body of law when desirous of reaping its benefits.[68]

The last element is that of belonging to a party to the conflict, which aims at proving a certain relationship between a group launching cyber attacks and a belligerent state.[69] While computer network attacks enable the use of "cyber militia" and offer the attractiveness of "plausible deniability" for a State, unless a relationship can be established between the group and the State, the participants will not be considered as lawful combatants.[70] The regular armed forces of the State would have no need to prove such a connection, but when it comes to organized online groups, it is not clear what degree of control over them is required.[71]

The most puzzling issue concerns the second and third elements, which require combatants to have a fixed distinctive sign recognizable at a distance and to carry arms openly. These two conditions are closely linked to the principle of distinction between combatants and civilians. Given that the two conditions are intended to eliminate confusion in this regard and to preclude any attempt at deception,[72] there is an inherent difficulty in transplanting them into an online environment, where it is impossible to tell who is sitting at any given computer due to the anonymity of cyberspace. Some scholars have proposed that given the impossibility of computer users being marked with distinctive signs, the requirement of displaying signs should be applied to computers or systems, just as military automobiles, aircraft and ships need to be marked with distinctive signs. This proposal is untenable since marking a military computer is tantamount to making a lawful target of any system to which it is connected.[73]

66 Tilman Rodenhäuser, *Organizing Rebellion: Non-State Armed Groups under International Humanitarian Law, Human Rights Law, and International Criminal Law*, Oxford University Press, Oxford, 2018, pp. 104–108.

67 H. Harrison Dinniss, *Cyber Warfare*, above note 13, p. 149.

68 Y. Dinstein, above note 46, p. 54.

69 See Denise Bindschedler-Robert, "A Reconsideration of the Law of Armed Conflicts", in *The Law of Armed Conflicts: Report of the Conference on Contemporary Problems of the Law of Armed Conflict*, 1971, p. 40; Katherine Del Mar, "The Requirement of 'Belonging' under International Humanitarian Law", *European Journal of International Law*, Vol. 21, No. 1, 2010.

70 H. Harrison Dinniss, "Participants in Conflict", above note 13, p. 262.

71 The effective control standard elaborated by the ICJ in Nicaragua appears inappropriate to define what "belonging to a party to the conflict" means, as, unlike the overall control and complete dependency standards, it expresses control over the act and not over the actor and thus focuses on specific activities. Marko Milanović, "State Responsibility for Acts of Non-State Actors: A Comment on Griebel and Plücken", *Leiden Journal of International Law*, Vol. 22, No. 2, 2009, p. 317. On the meaning of the effective control, overall control and complete dependency standards, see Antonio Cassese, "The Nicaragua and Tadić Tests Revisited in Light of the ICJ Judgment on Genocide in Bosnia", *European Journal of International Law*, Vol. 18 No. 4, 2007.

72 Y. Dinstein, above note 46, p. 37.

73 The Internet is constantly searched by millions of software bots intent on finding connected computers; a bot searching for military-designated IP addresses would be able to find them in a matter of minutes. Once

One may argue that armed forces could still wear uniforms in order to comply with the obligation of having a fixed distinctive sign recognizable at a distance;[74] for example, requiring members of USCYBERCOM to wear military uniforms when conducting cyber operations. This opinion apparently has merit – it would be ideal if regular forces could wear uniforms or otherwise distinguish themselves from civilians – but in practice such a requirement would probably mean little, since the warring parties remain anonymous. The object and purpose of this provision is that the aim of wearing a uniform is to eliminate the possibility of confusion in distinguishing between civilians and combatants. In traditional armed conflicts, by wearing uniforms, in most instances it is clear who is a combatant and who is not.[75] But when cyber combatants are sitting in front of their computers, sometimes a great distance from the view of those they are attacking, whether they wear uniforms or not makes no difference to the other belligerent State. In any event, even if we insist that formal military forces should wear uniforms, this requirement is absurd when dealing with cyber militias, volunteer corps or other organized cyber groups. What is more, it seems that cyberspace leaves no room for the requirement of carrying arms openly. Defining cyber weapons is already difficult enough, and to carry them openly is just impractical.[76] Certainly, it should not be ignored that there is a possibility of a kinetic attack on cyber combatants. In conclusion, we argue that in cyber warfare, the second and third elements would not be deleted outright, but there would be little need for much discussion about them.

Some may deem that on the digital battlefield, there is no real need for such distinctions; in the context of a cyber attack against military assets, the one who committed the attack is either a combatant or a civilian directly participating in hostilities. In either case, this specific person has lost his or her protected status. Nevertheless, some questions remain, particularly as to whether he or she would enjoy PoW status once captured.[77] Moreover, a civilian attacker might fail to meet the requirement of "threshold of harm" and "belligerent nexus",[78] and thus he or she would not lose the protected status at all.

identified, the only way to effectively move the computer or system out of range is to disconnect it, a solution which is likely to disrupt its normal running and/or usefulness; thus, any system remaining connected to the network in any way would be solely reliant on its electronic defences to prevent intrusions and defend against them. So, while initially the idea of displaying signs on computers or systems appears a useful solution, in practice it creates an imbalance between the purpose of the requirement of displaying signs and the ability of the military to conduct operations. See H. Harrison Dinniss, "Participants in Conflict", above note 13, p. 257; H. Harrison Dinniss, *Cyber Warfare*, above note 13, pp. 145–149.

74 Tallinn Manual 2.0, above note 31, Rule 87, p. 405.

75 This is not always the case; for example, civilians who directly participate in hostilities can be attacked, but they are hardly likely to wear military uniforms.

76 See Prashant Mali, "Defining Cyber Weapon in Context of Technology and Law", in Information Management Association, *Multigenerational Online Behavior and Media Use: Concepts, Methodologies, Tools, and Applications*, IGI Global, Hershey, PA, 2019; Jeffrey T. Biller and Michael N. Schmitt, "Classification of Cyber Capabilities and Operations as Weapons, Means, or Methods of Warfare", *International Law Studies*, Vol. 95, 2019; H. Harrison Dinniss, *Cyber Warfare*, above note 13, pp. 250–278.

77 H. Harrison Dinniss, *Cyber Warfare*, above note 13, p. 148.

78 Interpretive Guidance, above note 14, p. 46.

In conclusion, defining who is a cyber combatant is not only a legal intricacy, but also an extremely difficult technical issue for most States. The reality is that there is currently no way to clearly identify cyber combatants, and the existing rules are therefore applicable only to a limited extent. In comparison to a traditional armed conflict, civilians are more likely to be involved in a cyber armed conflict.[79] As Michael Schmitt has noted, the reasons for heavy civilian representation are multiple. From a cost-benefit perspective, training military personnel with cyber attack and defence expertise is extremely expensive and time-consuming for most countries, and what is more, the results are not guaranteed. In addition, cyber technology, by its nature, cannot be standardized and quantified. Not only is the technology always being developed and upgraded, it is also too limited and specialized.[80]

Elements (ii) and (iii) identified above – having a fixed distinctive sign recognizable at a distance and carrying arms openly – are ill-suited to the cyber context and thus probably need not be considered in cyber warfare. However, a person still has to at least satisfy elements (i), (iv) and (v) – the presence of a responsible command and a hierarchical relationship, conducting operations in accordance with the laws and customs of war, and belonging to a party to the conflict – to become a lawful combatant. Otherwise, they either remain protected from attack or will be considered as taking direct part in hostilities. Under these circumstances, the priority should be preventing over-militarization and minimizing unnecessary harm to civilians. Meanwhile, we should bear in mind that in case of doubt as to whether a person is a civilian, that person shall be considered to be a civilian.[81] Thus, it would be both unethical and unlawful to interpret the definition of cyber combatants in too broad a way.

Civilians taking direct part in cyber hostilities

Unlike combatants, civilians are not entitled to directly participate in hostilities; those who do so lose their general protection against the dangers of military operations and may be attacked for such time as they do so.[82] In addition, they may be prosecuted in domestic courts for their actions, even if the acts committed were lawful under IHL.[83] In the cyber context, the concept of civilians who directly participate in hostilities may be even more important, given the contemporary tendency in armed forces to outsource specialist work which requires cyber expertise to civilians.[84]

79 L. Zhu, "Competition for International Rules in Cyberspace", above note 2, p. 40.
80 M. N. Schmitt, above note 52, p. 527.
81 AP I, Art. 50(1); ICRC Customary Law Study, above note 12, Rule 6, pp. 23–24.
82 AP I, Art. 51(3); Protocol Additional (II) to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts, 1125 UNTS 609, 8 June 1977 (entered into force 7 December 1978), Art. 13(3); ICRC Customary Law Study, above note 12, Rule 6, pp. 19–24.
83 H. Harrison Dinniss, "Participants in Conflict", above note 13, p. 258.
84 D. Turns, above note 52, p. 279.

As discussed above, the term "direct participation in hostilities" refers to the notion that, as a general rule, civilians are not to be made the targets of attacks, unless and for such time as they directly participate in hostilities.[85] This is also known as the rule on non-combatant immunity.[86] When debating Article 51 of AP I, States did not settle on a precise definition of what was meant by the phrase "direct part in hostilities".[87] Both the *Targeted Killings* case[88] and the International Committee of the Red Cross's (ICRC) *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law* (Interpretive Guidance)[89] have made an important contribution to the interpretation of the notion of direct participation in hostilities. The Interpretive Guidance has generated considerable debate and some controversy.[90] While uncertainties remain and it is not crystal-clear how the guidance might be applied in practice on the physical battlefield, this is *a fortiori* the case when it comes to the virtual battlefield.[91]

Determining direct participation in hostilities is complex enough; determining direct participation in cyber hostilities seems even harder. As noted in the *Targeted Killings* case, it is possible to take part in hostilities without using weapons at all.[92] Thus, while the means of warfare today may be profoundly different from those of the last century, the effects of such means of warfare are essentially similar. A military communication system is rendered equally inoperative whether it is disabled by a computer virus or a bombing raid.

To further deconstruct this issue and provide guidance for practitioners, the Interpretive Guidance posits three cumulative elements which together constitute the act of direct participation in hostilities. First, the act must be likely to adversely affect the military operations of a party to an armed conflict or, alternatively, to inflict death, injury or destruction on persons or objects protected against direct attack (threshold of harm). Second, there must be a direct causal link between the act and the harm likely to result either from that act or from a coordinated military operation of which that act constitutes an integral part (direct causation). And third, the act must be specifically designed to directly cause the required threshold of harm in support of a party to the conflict and to the detriment of another (belligerent nexus).[93] Computer network attacks and computer network exploitation are also discussed, leading to the assessment that

85 AP I, Art. 51(3).
86 Judith G. Gardam, *Non-Combatant Immunity as a Norm of International Law*, Martinus Nijhoff, Dordrecht, 1993.
87 Michael Bothe, Karl Josef Partsch and Waldemar A. Solf, *New Rules for Victims of Armed Conflicts: Commentary to the Two 1977 Protocols Additional to the Geneva Conventions of 1949*, Martinus Nijhoff, Dordrecht, 1982, pp. 301–304.
88 Israel High Court of Justice, *Public Committee against Torture in Israel v. Israel et al.*, Case No. HCJ 769/02, Judgment, 11 December 2005 (*Targeted Killings*).
89 Interpretive Guidance, above note 14, p. 46.
90 "Forum: Direct Participation in Hostilities: Perspectives on the ICRC Interpretive Guidance", *New York University Journal of International Law and Politics*, Vol. 42, No. 3, 2010.
91 D. Turns, above note 52, p. 285.
92 Israel High Court of Justice, *Targeted Killings*, above note 88, para. 33.
93 Interpretive Guidance, above note 14, p. 46.

"electronic interference with military computer networks could suffice as direct participation in hostilities, whether through computer network attacks or computer network exploitation, as well as wiretapping the adversary's high command or transmitting tactical targeting information for attack".[94] This three-part conjunctive test, focusing on the threshold of harm, direct causation and the belligerent nexus, provides a useful starting point for assessing whether and to what extent a civilian is conducting cyber combatant activities should thus lose their protected status.[95] It remains an open question whether these criteria are interpreted in the same way in the cyber context.

The first element, that of threshold of harm, relates to the objective likelihood of causing death or injury to humans or destruction to property. If, for example, both the 2007 Estonia incident[96] and the 2010 Stuxnet incident[97] had been perpetrated by civilians in an international armed conflict, we could conclude that the cyber attacks in the Estonia incident would have failed to reach the threshold of harm, while in the Stuxnet scenario, the attacks would have reached such a threshold. The cyber attacks against Estonian cyber infrastructure caused large-scale inconvenience since Estonia is one of the most "wired" States in the world, but no one died or was injured, nor was any property destroyed or damaged, and the causing of mere inconvenience, however unpleasant, does not reach the threshold of harm.[98] However, what is covered by "inconvenience" is not defined, and this terminology is not used in IHL.[99]

On the other hand, the cyber attack against the Iranian nuclear centrifuges, used for enriching uranium, caused physical damage to those centrifuges.[100] In this respect, the Tallinn Manual 2.0 dictates that "the act must have the intended or actual effect of negatively affecting the adversary's military operations or capabilities, or inflicting death, physical harm, or material destruction on persons or objects protected against direct attack".[101] Thus, as set out in the Manual, the threshold of harm element is met even if the acts merely have the intended effect. This interpretation expands the threshold of harm element from objective likelihood to either subjective intention or objective likelihood, and further leaves a lot of room for discretion on this point.

94 *Ibid.*, p. 48.
95 This three-part test has also been adopted for application to cyber warfare in the Tallinn Manual 2.0, above note 31, pp. 429–430.
96 "Cyber Attacks against Estonia (2007)", above note 15; E. Tikk, K. Kaska and L. Vihul, above note 15, pp. 14–33.
97 "Stuxnet (2010)", *International Cyber Law in Practice: Interactive Toolkit*, CCD COE, available at: https://cyberlaw.ccdcoe.org/wiki/Stuxnet_(2010); E. Tikk, K. Kaska and L. Vihul, above note 15, pp. 66–89.
98 D. Turns, above note 52, p. 286.
99 ICRC Challenges Report 2015, above note 32, p. 42.
100 A report shows that between the end of 2009 and early 2010, about 1,000 centrifuges at a fuel enrichment plant facility in Natanz, Iran, had to be replaced, implying that those centrifuges were broken. David Albright, Paul Brannan and Christina Walrond, *Did Stuxnet Take Out 1,000 Centrifuges at the Natanz Enrichment Plant?*, Institute for Science and International Security, 22 December 2010; "Stuxnet (2010)", above note 97.
101 Tallinn Manual 2.0, above note 31, p. 429.

The second element, that of a direct causal link, should be interpreted broadly. According to the Interpretive Guidance, the harm in question must be brought about in "one causal step".[102] Such a strict interpretation of the causal proximity will be particularly problematic for cyber operations where the secondary or knock-on effect of a particular act may in fact be the purpose of the attack. We believe that "proximate causality", which contains both the subjective and objective perspective, is more suitable in the cyber context – that is to say, objectively, the damage caused by the cyber act is the normal and natural consequence, and such damage is subjectively foreseeable.[103]

Some hypothetical scenarios could help us better understand the proximate causality test in the cyber context. Civilians hired to perform general computer and IT services would not be deemed to be directly participating in hostilities if they were simply performing service contracts, such as running web pages and managing email log-in terminals,[104] because the causality is not proximate, any damage caused is not the normal and natural consequence of the actions involved, and any negative consequences may not be foreseeable by those carrying out the services. On the other hand, any employee or contractor who is specifically employed to conduct hostile cyber attacks would, in theory, satisfy the proximate causality test once he or she has done so.

It is also worth attempting to apply the "cyber kill chain" model,[105] which has been developed by Lockheed Martin to test whether there is proximate causality in specific conditions. The cyber kill chain model is an ordered list of the seven steps of a cyber attack, namely reconnaissance, weaponization, delivery, exploitation, installation, command and control, and action on objectives.[106] It gives a bird's-eye view of how a hacker can strike a target, and although not every attack may adhere to all of these steps, it still provides a good starting point. The first phase is reconnaissance, which includes the research, identification and selection of targets; this followed by the weaponization phase, which couples malware and exploits into a deliverable payload. The next step, delivery, involves transmitting the weapon to the target (e.g., via USB drives or email attachments); subsequently, the weapon will try to exploit a vulnerability in order to gain access to the victim. Until the end of the fourth phase, it is still hard to say whether the acts have a direct causal link with the consequence, since what will happen next is not necessarily foreseeable for the perpetrators. However, when it comes to the installation, command and control, and action on objectives phases, there is a high chance that the perpetrator will be able to foresee what will happen, and the damage caused is the natural or normal consequence of the acts in question.

102 Interpretive Guidance, above note 14, p. 53.
103 Bin Cheng, *General Principles of Law as Applied by International Courts and Tribunals*, Cambridge University Press, Cambridge, 1987, p. 181.
104 See Emily Crawford, *Virtual Battlegrounds: Direct Participation in Cyber Warfare*, Sydney Law School Research Paper No. 12/10, 8 February 2012, available at: https://ssrn.com/abstract=2001794.
105 See Lockheed Martin, "Gaining the Advantage, Applying Cyber Kill Chain Methodology to Network Defense", 2015, available at: www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/Gaining_the_Advantage_Cyber_Kill_Chain.pdf.
106 *Ibid.*

The belligerent nexus element is more a matter of fact than of law. Certainly, it requires that "the act must be specifically designed to directly cause the required threshold of harm in support of a party to the conflict and to the detriment of another".[107] It is not a *mens rea*-like element. What matters is the purpose of the act, which must be objectively designed to directly cause harm. This leads to the inference that hostile acts carried out under coercion or without knowledge will not satisfy the element of belligerent contact. In light of the fact that botnet attacks occur frequently, it must be noted that there should be an exemption to the loss of immunity if a civilian computer is hacked by a botnet and the relevant user is unaware of the virus and the attack. In this case, the relevant user should not be regarded as performing an action, and consequently, as they lack any manifestation of action, they would not fulfil the belligerent nexus element.

If a civilian merely writes a malware program which would result in the shutdown of critical infrastructures, this action should not be regarded as direct participation in cyber hostilities, since it would normally fail to satisfy all three elements, and in any case, the causality would be too remote. Similarly, civilian scientists and weapons experts are generally regarded as protected from direct attack.[108] If the civilian sends this self-written malicious program to the armed force that he or she supports, such action still does not constitute direct participation in hostilities – this is similar to the transportation of weapons. However, if this malicious program is aimed at conducting a specific hostile act, this action would become an integral part of a cyber military operation, thus fulfilling the proximate causality requirement. When a civilian, no matter whether they are under a contract with the armed forces or acting unilaterally, executes such a malicious program, they would probably fulfil the criteria and thus would lose their protected status and become a lawful target, at least during the period when the program was being executed.

Article 51 of AP I also stipulates the temporal scope of specific acts amounting to direct participation in hostilities – that is, civilians lose protection against direct attack "for such time" as they directly participate in hostilities.[109] If "such time" has passed, the protection granted to the civilian returns. This should be distinct from the rules set for members of armed wings of organized armed groups and for those who belong to a party to the conflict; these individuals are no longer civilians and, therefore, lose their protection against direct attack for the duration of their continuous combat function, while civilians lose their protection for the duration of specific acts amounting to direct participation in hostilities.[110]

107 Interpretive Guidance, above note 14, p. 46.

108 ICRC, *Fourth Expert Meeting on the Notion of Direct Participation in Hostilities: Summary Report*, Geneva, 27–28 November 2006, p. 48. The present authors note that some doubts were expressed as to whether this assessment could be upheld in extreme situations – namely, those in which the expertise of a particular civilian is of very exceptional and potentially decisive value for the outcome of an armed conflict, such as the case of nuclear weapons experts during the Second World War.

109 AP I, Art. 51(3); ICRC Customary Law Study, above note 12, Rule 6, pp. 19–24.

A particularly important issue in the cyber context is that of how to calculate the temporal scope of civilian loss of protection when dealing with repeated cyber operations in a relatively concentrated time period. If a civilian repeatedly launches cyber operations that could constitute direct participation in hostilities, what is the temporal scope, or period for that civilian of being targetable?

In a traditional battlefield setting, the Interpretive Guidance takes the position of treating those actions separately,[111] but the *Targeted Killings* case expresses concern about the "revolving door" phenomenon in this regard.[112] In the eyes of the Interpretive Guidance, the "revolving door" of civilian protection prevents attacks on civilians who do not, at the time, represent a military threat.[113] As the concept of direct participation in hostilities refers to specific hostile acts, IHL restores the civilian's protection against direct attack each time his or her engagement in a hostile act ends.[114] Considering that a large amount of cyber operations, such as DDoS attacks, are conducted multiple times within a time period, this strict time demarcation makes little operational sense. Yet the present authors also hold a sceptical attitude about calculating the period from the first operation throughout the whole intermittent activity. This is because civilians who directly participate in hostilities are not the same as members of organized military groups: though they are both targetable, they are two types of human targets. As mentioned before, members of organized military groups are targetable for the duration of their continuous combat function, but civilians who directly participate in hostilities are targetable only for the duration of their specific acts. "A civilian taking a direct part in hostilities one single time, or sporadically, who later detaches himself from that activity, is a civilian who, starting from the time he detaches himself from that activity, is entitled to protection from attack."[115] So, presuming that a civilian engages in repeated cyber attacks, if the whole period of time (from the beginning of the first attack to the end of the last attack) is continuously calculated as the period during which the civilian can be attacked, in a sense we are treating the civilian who directly participates in hostilities by the standard of combatants (continuous combat function), because we are directly regarding the intermission as an attackable period as well. Strictly speaking, civilians who directly participate in hostilities lose their protected status because of their specific acts, and are not considered to have carried out any hostile actions in the intermission. On the other hand, a civilian who has joined a military organization and commits a chain of hostile acts, with short periods of rest between them, loses his immunity

110 Interpretive Guidance, above note 14, p. 73.
111 *Ibid.*, pp. 70–71.
112 Israel High Court of Justice, *Targeted Killings*, above note 88, para. 40.
113 Interpretive Guidance, above note 14, pp. 70–71.
114 See the description of direct participation in hostilities as potentially "intermittent and discontinuous" in ICTR, *The Prosecutor v. Strugar*, Case No. IT-01-42-A, Judgment (Appeals Chamber), 17 July 2008, para. 178.
115 Supreme Court of Israel, *Public Committee against Torture in Israel v. Government of Israel*, Case No. HCJ 769/02, 13 December 2006, para. 39.

from attack for the entire time of his activity. For such a person, the rest between hostile acts is nothing more than preparation for the next hostile act.[116]

In conclusion, in interpreting direct participation in hostilities, the threshold of harm requires objective likelihood instead of mere subjective intention, and the belligerent nexus must be confirmed while the causal link should be proximate. The temporal scope is of great importance, but is quite tricky to establish. So far, absent international jurisprudence on the matter, clarification of the concept is still left for academic scholarship, future State practice and judicial decisions.

The principle of distinction concerning non-human targets in cyber warfare

All non-human targets[117] can be divided into two categories: military objectives and civilian objects. Civilian objects are all objects which are not military objectives.[118] Only military objectives can be the object of attacks.[119] This part will discuss what can be attacked under the law by applying the principle of distinction in the cyber domain – that is, what constitutes a military objective in the cyber context. It is worrying that almost everything in cyberspace has huge military potential, and the issue of dual-use objects plays a more important role in targeting than ever. With the increasing importance of data in a cyber armed conflict, the question of whether data itself could be regarded as a military objective will also be addressed.

The notion of "military objective": Two equivalent elements

The widely accepted definition of all non-human military objectives is as follows: insofar as objects are concerned, military objectives are limited to those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.[120]

The notion of "military objective" is critical since it directly determines what can or cannot be attacked pursuant to the principle of distinction. In reality, the term "military objective" has been interpreted in vastly different ways. Some hold that it means war-fighting or war-sustaining capability for military action in the definition of Article 52(2) of AP I and includes targets that "indirectly but

116 *Ibid.*, para. 39; Daniel Statman, "Targeted Killing", *Theoretical Inquiries in Law*, Vol. 5, No. 1, 2004, pp. 179, 195.

117 The present authors try not to use the term "objects" here because the question about whether there are non-human targets which are not "objects" will be discussed in the following paragraphs.

118 AP I, Art. 52(1); ICRC Customary Law Study, above note 12, Rule 9, pp. 31–32.

119 AP I, Art. 52(2); ICRC Customary Law Study, above note 12, Rule 7, pp. 24–28.

120 AP I, Art 52(2); ICRC Customary Law Study, above note 12, Rule 8, pp. 29–32; Jacob Kellenberger, "International Humanitarian Law at the Beginning of the 21st Century", statement given at the 26th Round Table on Current Problems in International Humanitarian Law, Sanremo, 5–7 September 2002.

effectively support and sustain the enemy's war-fighting capability".[121] In practical terms, compliance with the first criterion of "effective contribution" will generally result in the advantage required in the second criterion of "definitive military advantage".[122] Others argue that only when these two elements are cumulatively present is there a military objective in the sense of the Protocol.[123] In other words, the test for the military status of an object is twofold and the two requirements are equivalent.[124]

The present authors disagree with the view that "effective contribution" includes targets that "indirectly but effectively support and sustain the enemy's war-fighting capability", especially in the cyber domain. This interpretation is far too broad and defeats the philosophy behind the limitation of military objectives – indeed, by characterizing the contribution as "effective" and the advantage as "definite", the drafters of AP I tried to avoid such a wide-ranging interpretation of what constitutes a military objective.[125] And the broad interpretation would make the distinction even more confusing in the context of cyber warfare;[126] given that almost everything has a military potential in cyberspace, if indirect support could count as effective contribution, the interpretation would become nearly unlimited since it would allow "any of the adversary's information functions that have a bearing on his capability to fight to qualify as a legitimate target".[127] It is therefore at odds with the object and purpose of Article 52(2) of AP I.

Thus, the definition of military objective should contain two equally important elements: effective contribution and definite advantage. The fulfilment of the former element does not automatically lead to the fulfilment of the latter, since these two elements are independent. The definite advantage element was discussed at length when AP I was drafted. The adjectives considered and rejected included the words "distinct" (*distinct*), "direct" (*direct*), "clear" (*net*), "immediate" (*immediat*), "obvious" (*evident*), "specific" (*specifique*) and "substantial" (*substantiel*).[128] It is clear that the word "definite" has its own value and should not be ignored – the advantage has to be definite and concrete.[129]

121 DoD, above note 41, p. 210; Charles J. Dunlap, "The End of Innocence: Rethinking Non-Combatancy in the Post-Kosovo Era", *Strategic Review*, Vol. 9, 2000, p. 17; US Department of the Navy and Department of Homeland Security, *The Commander's Handbook on the Law of Naval Operations*, July 2007, para. 8.2. There are also some opposite views, such as Laurent Gisel, "The Relevance of Revenue-Generating Objects in Relation to The Notion of Military Objective", in ICRC, *The Additional Protocols at 40: Achievements and Challenges*, 18th Bruges Colloquium, 19–20 October 2017.

122 Program on Humanitarian Policy and Conflict Research at Harvard University, *HPCR Manual on International Law Applicable to Air and Missile Warfare*, Cambridge, MA, 2010, p. 49.

123 Yves Sandoz, Christophe Swinarski and Bruno Zimmermann (eds), *Commentary on the Additional Protocols*, ICRC, Geneva, 1987 (ICRC Commentary on APs), para. 2018.

124 E. Mavropoulou, above note 51, p. 44.

125 Marco Sassòli, "Military Objectives", in *Max Planck Encyclopedia of Public International Law*, 2015, para. 7.

126 J. Kelsey, above note 33, p. 1440.

127 M. Roscini, above note 65, p. 186.

128 ICRC Commentary on APs, above note 123, para. 2019.

129 Robert Kolb and Richard Hyde, *An Introduction to the International Law of Armed Conflicts*, Hart Publishing, Oxford, 2008, pp. 60, 131.

Potential and indeterminate forms of advantage are not acceptable; neither are political ones.[130] In other words, it is prohibited to launch an attack which only offers potential or indeterminate advantages.[131]

These two elements, effective contribution and definite military advantage, are also equivalent. It is often difficult to identify the military advantage anticipated for a given attack, especially in the cyber context, where measuring the effects of a cyber operation can be challenging.[132] In the cyber domain, where the military uses the same cyber infrastructure as the civilian population for its military activity, the second requirement of the definition becomes even more inclusive and one should be cautious with a sweeping conclusion that seriously underestimates the importance of the second element.[133] Cyberspace is relatively resilient compared to other targets. In the case of an attack against a cyber infrastructure like a communication network, the data flow is so flexible that even if certain communication paths are destroyed by the cyber attack, the data packages will have various other possible paths to follow so as to reach their intended destination.[134] In this case, the partial destruction of the network might effectively contribute to military action but will hardly offer a definite advantage in the end. Thus the judgment on a definite military advantage is complex and cannot be automatically satisfied once the effective contribution element is fulfilled.

Definite military advantage in the cyber context is always hard, if not impossible, to measure and quantify. After the Stuxnet incident, while Iran denied that the incident had caused significant damage, the International Atomic Energy Agency reported that Iran had stopped feeding uranium into thousands of centrifuges at Natanz. No one knows what consequences were caused by Stuxnet on the Iranian nuclear programme, and it is still unclear whether the decision to stop using the Natanz centrifuges was due to Stuxnet or to technical malfunctions inherent to the equipment.[135]

What is particularly worth mentioning in the context of cyberspace is that the requirement to identify a definite military advantage associated with attacking a particular target arises most often with respect to potential dual-use objects. A facility can either support solely civil or solely military purposes, but it can also support both purposes simultaneously, making it a dual-use object.[136] Essential infrastructure such as bridges, electricity-generating installations and oil-refining

130 ICRC Commentary on APs, above note 123, para. 2024.
131 *Ibid.*, paras 2024–2025.
132 M. Roscini, above note 65, p. 188.
133 R. Geiss and H. Lahmann, above note 34, p. 388.
134 *Ibid.*
135 Marco Roscini, "Military Objectives in Cyber Warfare", in Mariarosaria Taddeo and Luciano Floridi (eds), *Ethics and Policies for Cyber Operations: A NATO Cooperative Cyber Defence of Excellence Initiative*, Springer, Cham, 2017, p. 108; Katharina Ziolkowski, *Stuxnet – Legal Considerations*, CCD COE, Tallinn, 2012, p. 5, available at: https://ccdcoe.org/uploads/2018/10/Ziolkowski_Stuxnet2012-LegalConsiderations.pdf.
136 Dominik Steiger, "Civilian Objects", in *Max Planck Encyclopedia of Public International Law*, 2011, para. 12.

facilities may also have the potential to serve civil and military purposes at the same time.[137]

The fundamental difference in cyber warfare lies in the *sui generis* nature of cyberspace – namely, the "systemic inter-connectivity of civilian and military infrastructure".[138] For example, it is estimated that approximately 98% of US government communications[139] use civilian-owned and civilian-operated networks.[140] Civilian satellites, routers, cables, servers and even computers are all potential dual-use cyber facilities. The reality is that "every component of the cyber infrastructure, every bit of memory capacity has a military potential", and this blurs the line between civilian objects and military objectives.[141] One Chinese professor, Zhu Lixin of Air Force Engineering University, pointed out that the US military attaches great importance to building resilient intelligence, reconnaissance and surveillance (ISR) systems supported by artificial intelligence and quantum computing, and actively procures weapons such as smart small-diameter bombs, unmanned swarm systems, hypersonic weapons and directed-energy weapons to ensure lethality. So-called ISR systems require expensive machines such as quantum computers, satellites and artificial intelligence systems, many of which serve both military and civilian purposes.[142] Despite all the challenges, for the law, dual-use objects are not a separate category; they must equally fulfil the two-pronged test of Article 52(2) of AP I. The idea that the Internet itself could constitute a military objective is probably untenable, because the use of military code through the Internet might make some military contribution, but it is hardly effective, and it would not justify an attack because the mere disruption of its operations would be highly unlikely to offer the necessary "definite military advantage".[143] In any event, an attack on the whole Internet would breach the principle of proportionality,[144] ergo it would by no means be legal.

Furthermore, as the dual-use concept is not an innovation of cyber warfare, AP I provides a remarkable assumption for *lex scripta*: in case of doubt regarding the object's military status, it shall be presumed not to be so used.[145] Rule 102 of the Tallinn Manual 2.0 also provides that "[i]In case of doubt as to whether an object and associated cyber infrastructure that is normally dedicated to civilian purposes is being used to make an effective contribution to military action, a determination that it is so being used may only be made following a careful assessment".[146]

137 *Ibid.*
138 R. Geiss and H. Lahmann, above note 34, p. 385.
139 To avoid ambiguity, in terms of the numbers noted, we would like to remind readers that not all government communication is equal to military communication or military objectives.
140 Eric Talbot Jensen, "Cyber Warfare and Precautions against the Effects of Attacks", *Texas Law Review*, Vol. 88, No. 7, 2010, pp. 1522, 1542.
141 R. Geiss and H. Lahmann, above note 34, p. 388.
142 L. Zhu, "Competition for International Rules in Cyberspace", above note 2, p. 40.
143 International Criminal Tribunal for the former Yugoslavia, *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign against the Federal Republic of Yugoslavia*, 13 June 2000, para. 75.
144 AP I, Arts 51(5)(b), 57(2)(iii).
145 *Ibid.*, Art. 52(3).

Whether data falls within the ambit of military objectives

Data has become a cornerstone of life in many societies. During an armed conflict, the manipulation of data to cause physical harm undoubtedly requires the restraint of IHL, but the question of whether the data *per se* may constitute a military objective is also controversial. Cyber attacks are capable of being directed at data without causing physical effects, such as those targeting civilian financial systems. There are some views which hold that only a material, tangible thing can be a military objective in order to qualify as a legitimate target for attacks.[147] In the Tallinn Manual 2.0, only a minority of experts considered that certain data should be regarded as objects, thus constituting a military objective.[148] It is important to illustrate the relationship between the term "military objective" and the term "object". In a nutshell, from the wording of Article 52(2) of AP I – "in so far as *objects* are concerned, *military objectives* are limited to those *objects* which …" – a military objective is an object that meets certain criteria. The disputed point here is whether data *per se* could constitute an object. There are two main reasons to doubt that data could constitute a military objective, and both of them are related to the notion of "object". First, the intangible character of data fails to fit in the ordinary meaning of "object". Second, the ICRC Commentary on the Additional Protocols observes that "an object is characterized … as something visible and tangible".[149] Thus, data obviously does not qualify. Some scholars argue that data should be treated as objects.[150] Their argument is that cyber operations against civilian data are, on a factual level, illegal attacks on civilian targets. It is important to emphasize, in the view of these scholars, that any impact, direct or indirect, on civilian data in actions directed against lawful cyber targets must be measured in the principle of proportionality analysis and subject to the requirement to seek to minimize civilian collateral damage.[151] The advantage of this interpretation is that it protects civilian populations from the potential negative effects of cyber operations, but it is too broad, too inclusive, and would even include cyber operations, such as psychological operations, in which some countries are already engaged in practice on a regular basis.[152] In short, these criticisms and doubts

146 Tallinn Manual 2.0, above note 31, p. 448.

147 Yoram Dinstein, "Legitimate Military Objectives under the Current *Jus in Bello*", *International Law Studies*, Vol. 78, 2002, p. 142.

148 Tallinn Manual 2.0, above note 31, Rule 100, p. 437; M. N. Schmitt, above note 30, p. 269; Michael N. Schmitt, "Rewired Warfare: Rethinking the Law of Cyber Attack", *International Review of the Red Cross*, Vol. 96, No. 893, 2015, p. 200.

149 ICRC Commentary on APs, above note 123, paras 2007, 2008.

150 See, for example, Kubo Mačák, "Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law", *Israel Law Review*, Vol. 48, No. 1, 2015; ICRC Cyber Operations Paper, above note 32, p. 8; ICRC Challenges Report 2015, above note 32, pp. 41–42; ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts: Recommitting to Protection in Armed Conflict on the 70th Anniversary of the Geneva Conventions*, Geneva, 2019 (ICRC Challenges Report 2019), p. 28.

151 Michael N. Schmitt, "International Cyber Norms: Reflections on the Path Ahead", *Netherlands Military Law Review*, 17 September 2018, available at: http://puc.overheid.nl/doc/PUC_248171_11;

about the position held by most experts on the Tallinn Manual 2.0 focus on the exclusion of data from the protection provided by the law of targeting in AP I. According to this view, even cyber operations without physical consequences should at least be tested by the principle of proportionality and precaution[153] as long as they involve damage to or destruction of data, even if they may only have a potential impact on the civilian population.[154] Other scholars disagree and suggest that data should be regarded as a military objective once it fits the criteria. For those scholars, interpreting data as an object would "greatly expand the class of permissible targets in warfare",[155] and is counter to the object and purpose of enhancing the protection of civilians during situations of armed conflict. Furthermore, the interpretation of the ordinary meaning of "object" is debatable. There are translation discrepancies in the six authentic languages of AP I,[156] including French and Spanish, in which the term "*un bien*" may be translated into English as "a good" or "a property", and in French the legal term includes both tangible and intangible property.[157] As a matter of fact, in the Chinese context, the term "object"[158] generally refers to those items composed of materials that occupy a certain amount of space,[159] and thus intangible data does not count.

Some scholars also hold the opinion that data should be divided into two categories: "operational-level" data and "content-level" data.[160] According to that view, content-level data, such as the text of this article or the contents of medical databases, library catalogues and the like, are largely excluded from the ambit of military objective.[161] Operational-level data, the type of data that gives hardware

152 *Ibid.*; Michael N. Schmitt, "Notion of Objects during Cyber Operations: A Riposte in Defence of Interpretive and Applicative Precision", *Israel Law Review*, Vol. 48, No. 1, 2015.

153 As noted in Y. Dinstein, above note 46, pp. 164–174, the principle of precaution includes both active precautions in attack (AP I, Art. 57) and passive precaution (AP I, Art. 58). Active precautions in attack mandate "(a) [d]oing everything feasible to verify that the targets to be attacked are lawful [and] (b) [t]aking all feasible precautions in the choice of means and methods of attack, with a view to avoiding – or, at least, minimizing – collateral damage to civilians and civilian objects". Passive precaution requires belligerent parties, "'to the maximum extent feasible', (i) to endeavour to remove civilians and civilian objects under their control from the vicinity of military objectives; (ii) to avoid locating military objectives within or near densely populated areas; and (iii) otherwise to protect civilians and civilian objects against the dangers resulting from military operations".

154 Paul Ducheine and Terry Gill, "From Cyber Operations to Effects: Some Targeting Issues", *Netherlands Military Law Review*, 17 September 2018, available at: https://puc.overheid.nl/doc/PUC_248377_11/1.

155 K. Mačák, above note 150.

156 AP I, Art. 102: "The original of this Protocol, of which the Arabic, Chinese, English, French, Russian and Spanish texts are equally authentic, …".

157 K. Mačák, above note 150.

158 The Chinese version of AP I uses the term "物体". See www.icrc.org/zh/doc/assets/files/other/mt_070116_prot1_c.pdf.

159 "由物质构成的, 占有一定空间的个体". See 当代汉语词典, 上海辞书出版社, 2001 (*Contemporary Chinese Dictionary*, Shanghai Dictionary Publishing House, 2001); 现代汉语大词典, 下册, 上海辞书出版社, 2009 (*Modern Chinese Dictionary*, Vol. 2, Shanghai Dictionary Publishing House, 2009);新华汉语词典, 崇文书局, 2006 (*Xinhua Chinese Dictionary*, Chongwen Publishing House, 2006); 近现代词源, 上海辞书出版社, 2010 (*Etymology of Modern Times*, Shanghai Dictionary Publishing House, 2010).

160 Heather Harrison Dinniss, "The Nature of Objects: Targeting Networks and the Challenge of Defining Cyber Military Objectives", *Israel Law Review*, Vol. 48, No. 1, 2015, p. 41.

its functionality and ability to perform the tasks required of it, would be considered a military objective.[162]

Regrettably, the question of whether civilian data should be considered as civilian objects and therefore be protected under IHL seems to have received little attention from Chinese scholars. Zhu Yanxin, an associate professor from the Political College of the PLA National Defence University, holds the view that data could be defined as a military objective while not being an object.[163] He argues that data is a "non-object" military objective.[164] This argument is based on the language at the beginning of the second sentence of Article 52 of AP I:

> Attacks shall be limited strictly to military objectives. In so far as *objects* are concerned, military objectives are limited to those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.[165]

The literal wording of the provision clearly permits the existence of military objectives which are objects and non-objects.

The present authors' views on this point are basically in line with the ICRC's position paper of 2019.[166] Certain data, at least essential civilian data,[167] should fall within the ambit of civilian objects since the ordinary meaning of "object" is evolving and it would suit the object and purpose of the Geneva Conventions and their Additional Protocols. The term "object" does not necessarily exclude data from the scope of military objectives; we must bear in mind that the ordinary meaning of "object" should not be limited to that of the time when the treaty was adopted, and will evolve over time.[168] A treaty interpretation based solely on a textual approach ignores other methods of interpretation enshrined in the Vienna Convention on the Law of Treaties.[169] For instance, from the perspective of the object and purpose of AP I, the idea that "deleting or tampering with essential civilian data would not be prohibited by IHL in today's data-reliant world seems difficult to reconcile with the object and purpose of IHL".[170] It is a convincing argument to state that the replacement of paper files and documents with digital files in the form of data should not

161 *Ibid.*
162 *Ibid.*
163 朱雁新, 数据的性质: 对军事目标法律含义的重新解读, 载黄志雄主编, 网络空间国际规则新动向: "塔林手册2.0版" 研究文集, 社会科学文献出版社, 2019: 410–413 (Yanxin Zhu, "The Nature of Data: A Reinterpretation of the Legal Meaning of Military Objective", in Zhixiong Huang (ed.), *New Trends in International Rules for Cyberspace: Collection of Papers on Tallinn Manual 2.0*, Social Sciences Academic Press, China, 2019, pp. 410–413, only available in Chinese).
164 *Ibid*, p. 410.
165 AP I, Art. 52(2).
166 ICRC Cyber Operations Paper, above note 32, p. 8.
167 *Ibid.*, p. 8.
168 K. Mačák, above note 150.
169 See Vienna Convention on the Law of Treaties, 1155 UNTS 331, 23 May 1969 (entered into force 27 January 1980), Art. 31(3)(a).
170 ICRC Cyber Operations Paper, above note 32, p. 8.

decrease the protection that IHL affords to them.[171] If data is not an object, cyber operations against civilian data become a vacuum in IHL, and cyber operations that cause substantial damage to civilian life are not prohibited by law.[172]

The Tallinn Manual 2.0 equates military objectives with objects. To illustrate, the definition of military objectives proposed in Rule 100 leaves no space for non-objects: "Military objectives are those objects which …".[173] The viewpoint that data could constitute a military objective while not being an object is questionable for two main reasons. Firstly, this idea would shake the traditional persons–objects dichotomy, which, insofar as the construction of these provisions is concerned, appears to be correct; States have even rejected a third category such as "places".[174] Secondly, it would consequently leave no valid criterion for assessing whether a specific data set would be a military objective.[175] The persons–objects dichotomy provides the criteria of effective contribution and definite advantage for non-living things, while there are other requirements for living targets.[176] If data is not an object, this would lead to the unreasonable position that data needs to be assessed on the same basis as living targets. Therefore, the idea that data could be defined as a military objective while not being an object is not persuasive.

Conclusion

Cicero's aphorism, "during war, the laws are silent" (*silent enim legis inter arma*), does not reflect the modern reality. Despite all the challenges involved, the *jus in bello* principle of distinction is applicable to cyber warfare. Because of the lack of treaty provisions and judicial decisions specific to the cyber realm, the interpretation of existing law is based on the available academic discussion and limited State practice. There is a need for general clarification and further development of the principle of distinction in the cyber context; for example, the definitions of "cyber military objective" and "cyber combatant" remain controversial. Just as the UN Secretary-General mentioned at the World Economic Forum, "we need to find a minimum of consensus in the world on how to integrate these new technologies in the laws of war that were defined decades ago in a completely different context".[177]

Up to now, the Chinese government has not been clear about the application of IHL in cyberspace. There are indeed Chinese academic papers that look at the application of IHL in cyberspace, but the discussion of the principle of

171 ICRC Challenges Report 2019, above note 150, p. 28.
172 See M. N. Schmitt, above note 151.
173 Tallinn Manual 2.0, above note 31, p. 435.
174 M. Bothe, K. J. Partsch and W. A. Solf, above note 87, pp. 301–304.
175 K. Mačák, above note 150.
176 AP I, Art. 52(2).
177 World Economic Forum, "António Guterres: Read the UN Secretary-General's Davos Speech in Full", 24 January 2019, available at: www.weforum.org/agenda/2019/01/these-are-the-global-priorities-and-risks-for-the-future-according-to-antonio-guterres/.

distinction in cyberspace is limited both in length and in academic depth. Compared with the West, the research of Chinese scholars on this issue is still in a relatively preliminary stage. At present, there is no specific deconstruction or clarification of the application of the principle of distinction in cyberspace in Chinese academia.

Despite the potential technical challenges and uncertainties involved, the principle of distinction should be applied to cyberspace. It should also be carefully re-examined and clarified from the standpoint of preventing over-militarization and maximizing the protection of the interests of civilians. For human targets, the elements identified in customary international law and relevant IHL treaties to determine who is a combatant are not well suited to the digital battlefield. Nevertheless, cyber combatants are still obligated to distinguish themselves from civilians. In applying the principle of distinction, the present authors argue that it makes more sense to focus on substantive elements rather than formal elements such as carrying arms openly or having a fixed distinctive sign recognizable at a distance. In interpreting "direct participation in hostilities", the threshold of harm requires an objective likelihood instead of mere subjective intention, and the belligerent nexus should be confirmed while the causal link should be at least proximate. Applying the "cyber kill chain" model by analogy helps us to grasp the whole process of direct participation in hostilities during cyber warfare. For non-human targets, all military objectives must cumulatively fulfil both the effective contribution and definite military advantage criteria, which are equally indispensable. The same requirements apply to dual-use objects. As for the status of data, the ordinary meaning of "object" is debatable. There are translation discrepancies in the six authentic languages of AP I; in French the legal term includes both tangible and intangible property, while under the Chinese context, the term generally refers to those items composed of materials that occupy a certain amount of space, and thus intangible data does not count. Furthermore, one Chinese scholar argues that certain data belongs in the category of "non-object" military objective.

With the popularization of internet technology, unprecedented changes have taken place in the twenty-first century. The future of IHL in cyberspace still lies in the hands of States, particularly as they interpret the extant provisions and norms. War, technology and the *jus in bello* have been substantively intertwined and have interacted with each other since the beginning of organized human conflict, but the law has been constantly forced to adjust and is seemingly always "one war behind reality".[178] Therefore, faced with changes in technology and science, it is preferable to use methods of dynamic and evolving interpretation of international treaties and principles of international law in order to give them their full effect. It must be recognized that the increasing evolution of weapons and the rapid development of science and technology will have a tremendous

178 Jimena M. Conde Jiminián, "The Principle of Distinction in Virtual War: Restraints and Precautionary Measures under International Humanitarian Law", *Tilburg Law Review*, Vol. 15, No. 1, 2010. See also Marco Sassòli, Antoine Bouvier and Anne Quintin, *How Does Law Protect in War?*, 3rd ed., Vol. 1, ICRC, Geneva, 2011, p. 52.

impact on human society and that the *jus in bello* will adjust and adapt accordingly. However, it would be naive to assume that changes to IHL will be timely and effective.

It is probably too early to advocate for the adoption of a new treaty in this area. In any event, the prospects that States will agree on a comprehensive convention on cyber warfare in the near future are quite slim. Instead, the existing *lex lata* provides the basic regulation on targeting in the cyber domain. State practice, judicial decisions and scholars' views and teachings should take the lead on the interpretation of the existing legal framework and the assessment of whether the humanitarian concerns served by it are satisfied or undermined in the interconnected domain of cyberspace. Predictably, in the course of this evolution, States may try to analogically reason, induce or creatively fill in the gaps of the existing IHL, or push the *lex lata* concerning the principle of distinction beyond its normative boundaries when implementing new strategies in the era of cyber warfare. This trend needs to be strictly limited; however, it would be arbitrary to exclude the possibility of setting new rules. From the standpoint of preventing over-militarization and maximizing the protection of the interests of civilians, it is necessary to re-read the principle with great caution. While there have admittedly been no mass-casualty cyber events so far, when the interpretation and clarification of the existing rules are not enough, new rules need to be proposed before a "cyber Pearl Harbor" incident occurs.[179]

179 DoD, above note 4; J. J. Wirtz, "The Cyber Pearl Harbor", above note 4; J. J. Wirtz, "The Cyber Pearl Harbor Redux", above note 4.

International Review of the Red Cross (2020), **102** (913), 367–387.
Digital technologies and war
doi:10.1017/S1816383121000151

IRRC

Hacking humanitarians: Defining the cyber perimeter and developing a cyber security strategy for international humanitarian organizations in digital transformation

Massimo Marelli*
Massimo Marelli is Head of the Data Protection Office at the International Committee of the Red Cross

Abstract

Digitalization and new technologies have an increasingly important role in today's humanitarian activities. As humanitarian organizations become more active in and reliant on new and digital technologies, they evolve from being simple bystanders to being fully fledged stakeholders in cyberspace, vulnerable to adverse cyber operations that could impact on their capacity to protect and assist people affected by armed conflict or other situations of violence.

* The opinions and views expressed in this article are the author's own and do not necessarily represent those of the ICRC. The author is grateful to Bruno Demeyere, Kubo Mačák, Tilman Rodenhäuser, Andrea Raab, Eve La Haye, Gilles Cerutti, Delphine Van Solinge, Pierrick Devidal, Vincent Graf Narbel, Fabien Leimgruber, Martin Schuepp, Adrian Perrig, Sai Sathyanarayanan Venkatesh and Saman Rejali for their valuable feedback on an earlier draft. All errors are the author's own.

This shift makes it essential for humanitarian organizations to understand and properly map their resulting cyber perimeter. Humanitarian organizations can protect themselves and their activities by devising appropriate cyber strategies for the digital environment. Clearly defining the digital boundaries within which they carry out operations lays the groundwork for humanitarian organizations to develop a strategy to support and protect humanitarian action in the digital environment, channel available resources to where they are most needed, and understand the areas in which their operational dialogue and working modalities need to be adapted for cyberspace.

The purpose of this article is to identify the unique problems facing international humanitarian organizations operating in cyberspace and to suggest ways to address them. More specifically, the article identifies the key elements that an international humanitarian organization should consider in developing a cyber security strategy. Throughout, the International Committee of the Red Cross and its specificities are used as an example to illustrate the problems identified and the possible ways to address them.

Keywords: cyber, cyber strategy, cyber security, cyber operations, cyber attack, digital services, international organizations, humanitarian organizations, humanitarian action, digital transformation.

Introduction and "setting the scene"

Digitalization and new technologies have an increasingly important role in today's humanitarian activities.[1] This is happening for a number of reasons and in response to a number of new challenges. For example, armed conflicts are more and more fragmented and difficult to read, and security and acceptance are more and more fragile, making it harder for international humanitarian organizations[2] to access conflict areas and affected people.

Some of the topics considered in this article first appeared as part of a series of blog articles on the ICRC's *Humanitarian Law and Policy Blog*: see Massimo Marelli, "Hacking Humanitarians: Moving Towards a Human Cybersecurity Strategy", 16 January 2020, available at: https://blogs.icrc.org/law-and-policy/2020/01/16/hacking-humanitarians-cybersecurity-strategy/; Massimo Marelli and Adrian Perrig, "Hacking Humanitarians: Mapping the Cyber Environment and Threat Landscape", 7 May 2020, available at: https://blogs.icrc.org/law-and-policy/2020/05/07/hacking-humanitarians-mapping-cyber-environment/; Massimo Marelli and Martin Schüepp, "Hacking Humanitarians: Operational Dialogue and Cyberspace", 4 June 2020, available at: https://blogs.icrc.org/law-and-policy/2020/06/04/hacking-humanitarians-dialogue-cyberspace/.

1 See Anja Kaspersen and Charlotte Lindsey-Curtet, "The Digital Transformation of the Humanitarian Sector", *Humanitarian Law and Policy Blog*, 5 December 2016, available at: https://blogs.icrc.org/law-and-policy/2016/12/05/digital-transformation-humanitarian-sector/ (all internet references were accessed in January 2021).

2 This article's scope of analysis is restricted to international humanitarian organizations – i.e., organizations that have international organization or equivalent status and that have a humanitarian mandate. This does not include non-governmental organizations. The major difference between international humanitarian organizations and non-governmental organizations, for the purposes of this

It is against this backdrop that humanitarian organizations have strived to evolve and adapt in order to be better able to respond to humanitarian crises. They have started looking with interest at the possibility of complementing physical proximity with digital proximity – for example, by being accessible and responding to requests for information and assistance through social media or messaging apps.[3] They are developing new digital channels to deliver existing humanitarian services as well as new, natively digital services to affected populations that might already access other public and private services online and might expect the same of humanitarian organizations. They also see the positive role of digital platforms in consolidating existing resilience mechanisms of affected populations or enabling new ones, and are asking themselves how they can play a role in facilitating or enhancing such resilience mechanisms.[4]

Moreover, an increasing number of armed conflicts and other situations of violence are taking place in urban, connected environments[5] where it is often not the lack of data that makes it difficult to get proper situational awareness, but, rather, the abundance thereof and the difficulty in making sense of it. Humanitarian organizations are therefore considering the advantages of using new technologies, such as artificial intelligence (AI), machine learning and big data, to try and make sense of the complex environments in which they need to operate.[6] These technologies are sometimes built into commercially available products which can be acquired off the shelf from companies that are often interested in partnering with humanitarian organizations.

In addition, armed conflicts are lasting longer. The average length of time that the International Committee of the Red Cross (ICRC) has been present in the countries hosting its ten largest operations is more than forty years.[7] In protracted conflicts, humanitarian action may be required to plan for a long-term response that goes well beyond immediate and one-off distribution of food and non-food items or war surgery, and encompasses repeat distributions of aid in the long term. This

analysis, is the extent to which an international humanitarian organization enjoys privileges and immunities to ensure that it can perform its mandate in full independence. The existence and work of international humanitarian organizations is central to the functioning of the international community, and the international community relies on international humanitarian organizations to take care of tasks which individual States or groups of States cannot achieve alone. This makes international humanitarian organizations very relevant, but at the same time, potentially very vulnerable as cyber targets. However, the specific status they enjoy, and their privileges and immunities, can provide important safeguards for the protection of the organization if properly applied in a cyber environment.

3 See International Committee of the Red Cross (ICRC), in collaboration with The Engine Room and Block Party, *Humanitarian Futures for Messaging Apps: Understanding the Opportunities and Risks for Humanitarian Action*, January 2017, available at: www.icrc.org/en/publication/humanitarian-futures-messaging-apps.

4 See A. Kaspersen and C. Lindsey-Curtet, above note 1.

5 See David Kilcullen, *Out of the Mountains: The Coming Age of the Urban Guerrilla*, Oxford University Press, Oxford, 2015, available at: www.kilcullenstrategic.com/out-of-the-mountains/.

6 See Kristin Bergtora Sandvik, Katja Lindskov Jacobsen and Sean Martin McDonald, "Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation", *International Review of the Red Cross*, Vol. 99, No. 904, 2017.

7 See Ellen Policinski and Jovana Kuzmanovic, "Protracted Conflicts: The Enduring Legacy of Endless War", *International Review of the Red Cross*, Vol. 101, No. 912, 2019, p. 965.

response also includes working on systems and infrastructure such as water, sanitation and electricity. In this context, digital identification of beneficiaries – including, to some extent, biometric technology – becomes of interest for the humanitarian sector.

This process of "digital transformation", with humanitarian services being offered and made accessible digitally, is taking the collection and generation of personal data to a new scale. When combined with the introduction of commercial and/or technical third-party stakeholders, which are usually necessary to deliver relevant services digitally, it becomes a paradigm shift in the dynamics of humanitarian action delivery which organizations must take into account in relation to their bilateral interactions between humanitarian actors and their interlocutors. This shift brings into the picture technology providers, financial institutions, mobile network operators, and stakeholders involved in large-scale mass surveillance of telecommunications networks or targeted digital surveillance.

Personal data protection is an essential tool to enable humanitarian organizations to fully understand and dissect data flows, identify external stakeholders, map new risks and help identify mitigating measures. Therefore, it is crucial to enable the adoption of new technologies in a way that respects the rights, dignity and agency of affected populations and ensures accountability of and trust for humanitarian organizations, as well as upholding the responsibility to "do no harm" in the digital environment.[8]

As a consequence, data protection and ethics are key elements informing how an organization shapes the ways in which it carries out its work in favour of affected people in cyberspace, and therefore, its cyber perimeter. However, this analysis aims to go beyond strictly exploring the personal data protection aspects of digital transformation and humanitarian data ethics. Rather, it aims to unpack the unique problems faced by international humanitarian organizations operating in cyberspace and to propose solutions to address them. More specifically, the article intends to look at how the combination of an increased digital footprint, on the one hand, and the legal, technical and geopolitical implications of digitalization in the humanitarian sector, on the other, shape the cyber perimeter of an international humanitarian organization. For the purposes of this article, the cyber perimeter of an organization is defined as all the elements that jointly shape the presence and behaviour of the organization in cyberspace: its mandate, the activities it carries out in cyberspace, and how it goes about implementing

8 On data protection in humanitarian action, see Christopher Kuner and Massimo Marelli (eds), *Handbook on Data Protection in Humanitarian Action*, 2nd ed., ICRC, Geneva, 2020, available at: https://shop.icrc.org/handbook-on-data-protection-in-humanitarian-action-print-en. On the implications of metadata generation through third-party interactions in delivering humanitarian programmes, see Tina Bouffet and Massimo Marelli, "The Price of Virtual Proximity: How Humanitarian Organizations' Digital Trails can Put People at Risk", *Humanitarian Law and Policy Blog*, 7 December 2018, available at: https://blogs.icrc.org/law-and-policy/2018/12/07/price-virtual-proximity-how-humanitarian-organizations-digital-trails-put-people-risk/. On of the use of biometric data by the ICRC, see Ben Hayes and Massimo Marelli, "Facilitating Innovation, Ensuring Protection: The ICRC Biometrics Policy", *Humanitarian Law and Policy Blog*, 18 October 2019, available at: https://blogs.icrc.org/law-and-policy/2019/10/18/innovation-protection-icrc-biometrics-policy/.

and protecting those activities, particularly in anticipation of and in response to specific threats.

Understanding and conceptualizing their cyber perimeter is essential for organizations working in humanitarian action and undergoing a process of digital transformation of the magnitude mentioned above. As these entities become more active in and reliant on cyberspace, they are moving away from being bystanders and towards being fully fledged stakeholders in this domain, itself vulnerable to adverse cyber operations, or to being caught up in "cross-fire", which might impact their capacity to carry out humanitarian activities for those most in need.

This shift makes it essential for humanitarian organizations to understand and properly map their resulting cyber perimeter. Doing this effectively can allow them to delineate a strategy to support and adequately protect the delivery of humanitarian action in the digital environment; to channel the resources available to where they are most needed; and to understand the areas in which their operational dialogue and working modalities need to be adapted to be fit for cyberspace.

In this sense, an international humanitarian organization's cyber perimeter may be analyzed in light of: (1) what the organization wants to do in the digital environment and the organization's digital humanitarian operations; (2) the identity, mandate and *modus operandi* of the organization, and the affected people it serves; and (3) the cyber environment, particularly regarding the challenges and threats that the organization faces in the digital space.

This paper argues that these aspects, and the challenges arising under each of them due to the organization being active in cyberspace, should shape an organization's cyber security strategy. Such a strategy would thereby set out the following (non-exhaustive) protections and affiliations: (1) the legal protections it needs to seek; (2) the technical protections it is entitled to or can seek for its data and for its data flows; and (3) the stakeholders it needs to engage with and the operational dialogue it deploys with them. Each of these elements is analyzed, in turn, below.

What the organization wants to do in the digital environment, and the organization's digital humanitarian operations

To accurately determine an organization's cyber perimeter, the first step is to analyze precisely what it is that the organization wants to do in cyberspace, and to map the organization's current or envisaged digital humanitarian operations. This will be essential to determining the other elements of the organization's cyber perimeter and what the organization can do to secure it, as will be seen further below.

In the case of the ICRC, offering digital services directly to beneficiaries is at the core of this organization's institutional strategy for 2019–22.[9] This strategy is

9 See ICRC, *ICRC Strategy 2019–2022*, Geneva, September 2018, available at: www.icrc.org/en/publication/4354-icrc-strategy-2019-2022.

dictated primarily by (1) the increased challenges in having physical access to conflict areas, and the consequent need to complement physical proximity with digital proximity and accessibility; and (2) the fact that conflicts increasingly take place in areas where people are more likely to have access to connectivity, are used to accessing services online, and expect to interact with humanitarian organizations digitally. This reality requires the ICRC to engage in significant digital transformations in order to meet its objectives. This, in turn, will lead the organization to exponentially increase its digital footprint, and this is a trend that is common to most organizations, both in the humanitarian sector and beyond, in the digital era. This trend is one that comes with a natural increase in attack surface and exposure, and attractiveness as a target of adverse cyber operations.[10]

An important objective of the ICRC is linked to leveraging data (both data generated as part of its digital growth and data generated, acquired or available externally) by, for example, enabling predictive analytics and big data analysis, or developing or fine-tuning AI and machine learning tools to help solve problems that are specific to humanitarian action. This is important as it may help to inform the organization's decisions and improve its operational readings of armed conflicts and other situations of violence – for example, by informing its readings of anticipated displacement patterns, identifying influencers within parties to a conflict that can be relevant interlocutors to ensure access, or improving logistics and supply chain management.[11] Leveraging data can also be useful to support humanitarian action through various data science tools and techniques, from statistics to AI (for example, by using facial recognition for the determination of the whereabouts of missing persons).[12]

Thus, in the case of the ICRC, the organization wants to use the cyberspace domain to (1) achieve digital proximity to complement physical proximity, offering humanitarian services and being reachable digitally and remotely by affected populations that are increasingly connected; (2) facilitate and leverage new resilience mechanisms of affected populations enabled by digital platforms; and (3) leverage data to better inform its decision-making, which then feeds into how its cyber perimeter will be shaped.

The organization's identity, mandate and *modus operandi*

To determine an organization's cyber perimeter, it is important to place the organization's identity, mandate and *modus operandi* at the centre of the analysis, in order to determine what needs to be protected and how it can be protected

10 See ICRC, *The Potential Human Cost of Cyber Operations*, Geneva, 29 May 2019, available at: www.icrc.org/en/document/potential-human-cost-cyber-operations.

11 See, for example, "Big Data, Migration and Human Mobility", *Migration Data Portal*, available at: https://migrationdataportal.org/themes/big-data-migration-and-human-mobility.

12 See, for example, ICRC, "Rewards and Risks in Humanitarian AI: An Example", *Inspired*, 6 September 2019, available at: https://blogs.icrc.org/inspired/2019/09/06/humanitarian-artificial-intelligence/.

with respect to the cyberspace domain. Each organization should start its analysis from the specificities of the organization.

Taking again the example of the ICRC, this organization is a neutral, impartial, independent organization with the exclusively humanitarian mission to protect the life and dignity of victims of armed conflicts and other situations of violence. The work of the ICRC is based on the Geneva Conventions of 1949, their Additional Protocols, the organization's Statutes and those of the International Red Cross and Red Crescent Movement (the Movement), and the resolutions of the International Conferences of the Red Cross and Red Crescent.

The ICRC enjoys a special legal status and privileges and immunities under both international and domestic law.[13] The purpose of the privileges and immunities is to enable the ICRC to effectively carry out its mandate, and to do so in full conformity with its Fundamental Principles and standard working modalities.[14]

As will be seen further below, the ICRC's neutrality, impartiality and independence, the exclusively humanitarian nature of its work, and the privileges and immunities it enjoys in most of the countries in which it operates, enable it to carry out its mandate and are all essential elements shaping the organization's cyber perimeter and clearly distinguishing it from the cyber perimeters of other organizations.

The ICRC is entrusted by governments, through international humanitarian law and the Statutes of the Movement,[15] to assist and protect people during armed conflict and other situations of violence. To be able to carry out this mandate today, as outlined above, the organization also needs to be present and act in cyberspace by, for example, providing digital services. The same commitment from governments that would apply in the physical world to respect the ICRC's work as well as its working modalities for the benefit of populations affected by armed conflicts and other situations of violence should apply, *mutatis mutandis*, in cyberspace.

In carrying out its mandate, the ICRC adopts a proximity-based approach, through its approximately 20,000 staff members across eighty countries, in order to respond to the humanitarian needs of affected populations and to engage with key stakeholders on the application of international humanitarian law.[16] Unlike other humanitarian organizations that often operate through implementing partners, the ICRC's type of work requires direct proximity with affected populations (e.g. displaced populations, people deprived of their liberty, the wounded and sick, separated family members and unaccompanied minors, and families of the

13 See Els Debuf, "Tools to Do the Job: The ICRC's Legal Status, Privileges and Immunities", *International Review of the Red Cross*, Vol. 97, No. 897/898, 2016, available at: https://international-review.icrc.org/articles/tools-do-job-icrcs-legal-status-privileges-and-immunities.

14 See ICRC, "Fundamental Principles", available at: www.icrc.org/en/fundamental-principles.

15 See Statutes of the International Red Cross and Red Crescent Movement, adopted by the 25th International Conference of the Red Cross, Geneva, 1986 (amended 1995, 2006), available at: www.icrc.org/en/doc/resources/documents/misc/statutes-movement-220506.htm.

16 See ICRC, "What We Do", available at: www.icrc.org/en/what-we-do.

missing) as well as a physical presence in the areas where those affected populations are located.

An essential precondition for access is trust. This relates to the trust of both (1) affected populations and (2) parties to the armed conflict and actors in other situations of violence. As far as affected populations are concerned, trust is established by the guarantee that any engagement between them and the ICRC will be exclusively humanitarian. In particular, affected people expect that the information they provide for exclusively humanitarian purposes is treated as such and is not used or handled in a way that is detrimental to their safety or to humanitarian action more generally, such as when non-humanitarian stakeholders use the information for the furtherance of conflict-related objectives, counterterrorism agendas, migration flow controls or commercial exploitation. The importance of ensuring that data collected for humanitarian purposes are not used for other purposes is acknowledged in both the Resolution on Privacy and International Humanitarian Action (adopted by the International Conference of Privacy and Data Protection Commissioners in Amsterdam in 2015)[17] and the Movement's 2019 resolution on "Restoring Family Links while Respecting Privacy, Including as it Relates to Personal Data Protection".[18]

As far as parties to an armed conflict and actors in other situations of violence are concerned, to establish trust, they need to be confident that the work of the organization is neutral, impartial, independent and exclusively humanitarian. This entails that the organization take measures to minimize the risk that the data it collects will be accessed by such actors and to ensure that it does not end up being used to further conflict-related purposes, used by law enforcement or intelligence agencies, used as evidence in criminal proceedings, or otherwise made public. One of the key working modalities enabling the ICRC to have access to conflict areas and people affected by conflict is, therefore, that of confidentiality.[19] In particular, the ICRC does not share with any third parties information relating to its confidential bilateral dialogue with the authorities and other actors involved in conflicts and other situations of violence. This working modality is also safeguarded by the privilege of non-disclosure, a specific protection under customary international law from which the ICRC is the only organization to benefit.[20]

Although no academic study appears to be available to support this finding, it is the experience of the author and other humanitarian operators that in the

17 See "Resolution on Privacy and International Humanitarian Action", 37th International Conference of Data Protection and Privacy Commissioners, Amsterdam, 27 October 2015, available at: http://globalprivacyassembly.org/wp-content/uploads/2015/02/Resolution-on-Privacy-and-International-Humanitarian-Action.pdf.

18 See ICRC, "Restoring Family Links while Respecting Privacy, Including as it Relates to Personal Data Protection", 33IC/19/R4, Resolution 4 adopted at the 33rd International Conference of the Red Cross and Red Crescent, Geneva, 9–12 December 2019, available at: https://rcrcconference.org/app/uploads/2019/12/33IC-R4-RFL-_CLEAN_ADOPTED_en.pdf.

19 See ICRC, "Confidentiality Q&A", 15 January 2018, available at: www.icrc.org/en/document/confidentiality-q.

20 See E. Debuf, above note 13.

physical world, and to achieve physical proximity, trust is gained through a number of factors, including vulnerability.[21] In order to ensure and secure its presence, the ICRC does not generally rely on armed escorts and armoured vehicles, or on physical barriers; rather, and with all the vulnerability that this involves, it relies only on the acceptance of its humanitarian work and trust in its neutral, impartial and independent approach. If a stakeholder is not convinced about this, the organization's staff and assets would be very easy to target. The fact that the organization and its staff expose themselves and are so vulnerable vis-à-vis any possible ill-intentioned third parties leads interlocutors to trust that the organization stands for what it says and does not have ulterior motives.

In the digital world, however, vulnerability is not a strength but a weakness. The idea that the systems of the organization could easily be breached if anyone wanted to attack them would, alone, destroy any trust in the organization and discourage stakeholders from engaging with it. Therefore, to establish trust, to ensure digital proximity and to avoid causing detriment to physical proximity, the ICRC must be able to demonstrate the security and resilience of its cyber infrastructure. It is therefore essential for an international humanitarian organization to have full awareness of its cyber environment, challenges and threats.

The cyber environment, and the challenges and threats an organization faces therein

The cyber environment in which an international humanitarian organization operates presents a number of threats. These are often analyzed through the "confidentiality, integrity, availability" (CIA) triad.[22] As discussed below, in the case of international humanitarian organizations, the "classic" CIA analysis is not sufficient and needs to be adapted to take into account specific security threats arising from "jurisdictional" considerations – i.e., the fact that access may take place by virtue of authorities exercising jurisdiction over processors or sub-processors. Additional and specific considerations should be developed concerning the security of the supply chain. Each of these aspects is analyzed, in turn, below.

Confidentiality

A humanitarian organization may have to deal with situations in which individuals or groups supporting one party to an armed conflict (or actor in other situations of violence) may try to access sensitive data held by the organization. This is because the data in question may relate to specific individuals of interest, or populations

21 See Philippe Dind, "Security in ICRC Field Operations", *Secure 02*, Finnish Red Cross, June 2002, p. 27, available at: www.icrc.org/en/doc/assets/files/other/secure02_dind.pdf.

22 See Michael Nieles, Kelley Dempsey and Victoria Yan Pilliterri, *An Introduction to Information Security*, NIST Special Publication 800-12, National Institute of Standards and Technology, June 2017, available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-12r1.pdf.

linked to or of the same ethnic origin or national or political affiliation as that party's enemy. Health information, for example, may indicate a medical condition that is linked to a high-value target.[23]

"Big data theft" attacks are also a possible, important, confidentiality-type challenge. These may be aimed at collecting as many large data sets as possible, which are then correlated, analyzed, and used to profile individuals of interest to the attacker.[24] Such individuals might include beneficiaries of humanitarian action or other interlocutors of the humanitarian organization's neutral and impartial dialogue. Individuals so profiled could then be put under targeted surveillance, and data about them possibly used to inform additional actions in furtherance of conflict-related objectives. This concern may relate to large data sets, including metadata (that is, data about data), held both by humanitarian organizations themselves and by their third-party service providers (such as telephone companies or financial institutions), which may generate and use these data in the framework of humanitarian programmes such as mobile cash transfers.

Collaboration with or engagement of third-party technology service providers to handle or process data, such as through certain types of cloud-based solutions or in cash transfer programmes involving financial service providers and/or mobile network operators, is also extremely significant for the discussion on confidentiality. International humanitarian organizations can benefit from certain privileges and immunities regarding the data they collect. Where they do, the authorities cannot lawfully use due process to seek to access data they hold, thereby preserving confidentiality. It is very important that similar protections are acknowledged where it is a third-party service provider that processes data for the organization, though specific challenges involved in the generation and processing of data by third-party providers through digital tools make this principle difficult to apply.

To understand how third-party service providers can pose a threat to the data security of an international humanitarian organization, it is necessary to have a clear appreciation of the application of the principle of sovereignty to cyberspace, in particular by analyzing how States see their jurisdiction over technology providers, the infrastructure that supports data flows, and the data flows themselves, whether on their territory or outside. It is key for an international humanitarian organization, and particularly one like the ICRC, to ensure that no authority can by due process legitimately seek access to data held by the organization, whether directly or through third-party processors.

A digitalization of the scale and magnitude highlighted above, however, is most likely not going to be possible without leveraging the public cloud for at least

23 See, for example, C. Currier, "The NSA Plan to Find Osama Bin Laden by Hiding Tracking Devices in Medical Supplies", *The Intercept*, 21 May 2015, available at: https://theintercept.com/2015/05/21/nsa-plan-find-osama-bin-laden-infiltrating-medical-supply-chain/.

24 See, for example, Bill Gertz, "Cybercom: Big Data Theft at OPM, Private Networks Is New Trend in Cyber-Attacks", *Washington Free Beacon*, 27 July 2015, available at: https://freebeacon.com/national-security/cybercom-big-data-theft-at-opm-private-networks-is-new-trend-in-cyber-attacks/.

part of the organization's service offering.[25] Technology companies are increasingly and rapidly pushing their offering of software and storage to the public cloud and are no longer supporting non-cloud-based alternatives, often rendering them obsolete. In addition, certain tools enabling maximization of information – through, for example, AI – may be procured and deployed more efficiently on the public cloud. Because of this, the non-cloud-based model involving solutions held, managed and supported on the premises of the organization, traditionally favoured by security-conscious organizations, is harder and harder to sustain over the medium term. Even software that is procured as an on-premise solution today is likely to be linked to public cloud applications and/or sharing diagnostic or telemetry data across jurisdictions.[26] This means that data collected and generated by an organization will most likely be processed by third-party technology providers at some point. This brings significant new challenges in guaranteeing confidentiality.

It is therefore important for humanitarian organizations to carefully analyze this area and find solutions that are suitable for the sensitive work they are doing. Such considerations ought to bear in mind the specific architectural features of the public cloud[27] and legislation allowing authorities to access data generated and/or stored outside of their territory, such as the US CLOUD Act and other equivalent legislation elsewhere. CLOUD Act-type legislation and its impacts are spreading fast around the world,[28] due primarily to two factors: (1) other countries replicating the Act in order to assert jurisdictional control over data, and (2) agreements between the United States and third countries, under the CLOUD Act itself, allowing both parties to seek access to data under one another's jurisdictional control.

Integrity

From the point of view of integrity, an important challenge comes from the increasing use of AI and machine learning in supporting decision-making and situational awareness. This situation raises the threat that third parties might tamper with the accuracy and integrity of data used to train algorithms and develop models, as well as data sets used for the analysis, thereby interfering with the outcome of the analysis and decision-making.[29] Humanitarian organizations may, consequently, be manipulated into wrongly prioritizing certain affected populations over others or operating in particular areas over others, or be

25 For a description of the public cloud and why it can be an important asset to leverage, see Microsoft, "What Are Public, Private, and Hybrid Clouds?", available at: https://azure.microsoft.com/en-us/overview/what-are-private-public-hybrid-clouds/.

26 See, for example, Dutch Ministry of Justice, *DPIA Office 365 ProPlus Version 1905: Data Protection Impact Assessment on the Processing of Diagnostic Data*, June 2019, available at: www.government.nl/documents/publications/2019/07/22/dpia-office-365-proplus-version-1905.

27 See Microsoft, "What Is Cloud Computing? A Beginner's Guide", available at: https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/.

28 See US Department of Justice (DoJ), "CLOUD Act Resources", available at: www.justice.gov/dag/cloudact.

29 See C. Kuner and M. Marelli (eds), above note 8, Chap. 16.3.5.

otherwise manipulated in ways that may be detrimental to affected populations, or to the neutrality, impartiality and independence of their action.

Availability

From the point of view of availability, or ensuring timely and reliable access to and use of information, the concern is with situations in which the humanitarian organization offers digital services to affected populations. This can happen in a situation in which digital proximity is successfully deployed to complement physical proximity, or in a situation in which physical access is impossible and digital access is used instead. If affected populations rely on the availability of digital services from humanitarian organizations for their livelihood or for humanitarian protection, any cyber operation affecting availability of these services will have humanitarian consequences. In these cases, cyber operations affecting the availability of (digital) humanitarian services, like distributed denial-of-service (DDoS) operations or operations involving ransomware, raise very serious humanitarian concerns. Within this category, humanitarian organizations should also consider the implications for their capacity to deliver digital humanitarian services and for the possibility for affected populations to access them.[30] In addition, although not directly a type of challenge affecting the systems and infrastructure of an organization, humanitarian organizations also need to consider in their cyber perimeter the possibility that operations may be carried out by a stakeholder that would use cyber means against its adversaries by impersonating the organization or exploiting its name[31] or reputation, thereby attacking the sense of trust that individuals may have in it.

Supply chain security

Specific challenges are presented in ensuring the security of the supply chain.[32] This means, for example, that no back doors are present in the hardware or software procured and used by the humanitarian organization to deliver digital humanitarian services and/or to operate its systems. As far as hardware is concerned, while it may be possible for an organization, going forward, to effectively invest in the security of some key components of the hardware it procures,[33] it will still be unrealistic to aim for security of all the components it requires.

30 See Berhan Taye and Sage Cheng, "The State of Internet Shutdowns", *Access Now*, 8 July 2019, available at: www.accessnow.org/the-state-of-internet-shutdowns-in-2018/.

31 See, for example, Bill Marczak and John Scott-Railton, "The Million Dollar Dissident: NSO Group's iPhone Zero-Days Used against a UAE Human Rights Defender", *Citizen Lab*, 24 August 2016, available at: https://citizenlab.ca/2016/08/million-dollar-dissident-iphone-zero-day-nso-group-uae/.

32 See, for example, ICT Switzerland, "Supply Chain Security", available at: https://ictswitzerland.ch/en/topics/cyber-security/supply-chain/.

33 See Fabio Bergamin, "Open-Source Microprocessor", *ETH Zürich*, 30 March 2016, available at: https://ethz.ch/en/news-and-events/eth-news/news/2016/03/open-source-microprocessor.html.

A comprehensive strategy to address supply chain security concerns may need to be developed by the organization. Such a strategy would need to cover a combination of elements such as open-source hardware components, procurement practices, usage awareness and practices (such as staff training, but also minimizing the capacity of the hardware and software so that they process only the data and perform only the operations that are strictly required for the purposes of the processing), and partnerships with academia on solutions to monitor performance of hardware in order to detect possible anomalies linked to a compromised piece of hardware.[34] As far as software is concerned, some software companies may provide access to source code to countries and international organizations so that they can audit it and verify that no back doors are present.[35]

Although an international organization may seek access to such programmes, this may not be a solution available with all suppliers. In addition, even if the organization did have access to the source code, it may not have the means to effectively review all the lines of code of the software procured and thereby ensure its own protection.

The legal protections that an international humanitarian organization needs to seek

The first countries to explore the legal implications of hosting data for another subject of public international law were the governments of Estonia and Luxembourg, with the establishment of Estonia's "data embassy" in Luxembourg in 2017.[36] The interpretation of the law in this area is not fully settled, and in this section of the article, a number of unanswered questions are raised along with recommendations for possible clarifications that can be sought in relation to privileges and immunities to ensure that headquarters agreements fully reflect the specific needs raised by the hosting of data and applications in the strategic locations where the organization bases the most significant hosting of its data and applications.

The independence required by an international organization to fully and effectively implement its mandate is generally safeguarded in headquarters, status or host-State agreements. These provide for a series of privileges and immunities for the organization and its staff, including immunity from judicial or administrative process for the organization and its property, assets and staff, as well as inviolability of its premises, property, assets, correspondence and archives.

34 See Markus Gross, "A Booting Computer Is as Vulnerable as a Newborn Baby", *ETH Zürich*, 5 November 2019, available at: https://ethz.ch/en/news-and-events/eth-news/news/2019/11/project-opentitan.html.

35 See Microsoft, "Government Security Program", available at: www.microsoft.com/en-us/securityengineering/gsp.

36 See e-Estonia, "Estonia to Open the World's First Data Embassy in Luxembourg", available at: https://e-estonia.com/estonia-to-open-the-worlds-first-data-embassy-in-luxembourg/.

Clarifications may be required as to the interpretation of these agreements and their application in the digital environment.

It is important to clarify the application of the privileges and immunities of international organizations to include data (in transit, at rest and in processing) stored and processed not only by the humanitarian organization directly, but also by third-party service provider(s) or separate organization(s), including when hosted or otherwise processed by third-party technology providers on behalf of the organization, as well as the servers and networks used by the organization, whether they belong to the organization or to a third-party service provider.[37] Surprisingly, and to the best of this author's knowledge, no literature exists on this very important question.

Other provisions typically found in headquarters agreements generally involve guarantees that the host State will permit the organization's free use of the means of communication that the organization deems most appropriate, for official purposes and without any interference. Data flows required for and generated by the deployment of digital humanitarian services are covered by these guarantees, in addition to immunity and inviolability provisions. The agreements also cover the freedom for the organization to deploy specific technical protections in order to give practical effect to these provisions. Such protections could include sophisticated encryption algorithms or technologies incorporating them, and technologies aimed at preventing interference with, or interception of, communications and data flows involving the organization.

The agreement between an international humanitarian organization and a host State may also need to clarify that permitting and protecting free use of the means of communication includes, for instance, not interfering with access to the Internet and not interrupting or slowing down the internet connection of the organization or of a third-party service provider in a targeted manner. Considering that some measures of this type may be necessary to deal with DDoS attacks, however, and to avoid unintended consequences, it would also be important to clarify how these guarantees apply in scenarios where the organization is subject to a DDoS operation. While guaranteeing free communications may require that a host State does not block or reduce data traffic to the organization, such measures may be necessary if the host State is to protect such communications in cases of DDoS operations affecting the organization.

Specific considerations may need to be taken into account in cases where the organization processes data through cloud providers established in the host State. In particular, in addition to the considerations listed above, it would be necessary to clarify whether and to what extent staff of the third-party technology provider may also be covered by the immunities of staff linked to the tasks

37 For a reference to the US State Department's position supporting such application of privileges and immunities of States, see *Implementation of the Virtual Data Embassy Solution: Summary Report of the Research Project on Public Cloud Usage for Government, Conducted by Estonian Ministry of Economic Affairs and Communications and Microsoft Corporation*, 2015, p. 14, n. 12, available at: https://tinyurl.com/3rucylfy.

carried out in the performance of their functions as staff of the organization, insofar and to the extent that they process data of the organization and have access to clear, unencrypted data. These individuals may come into contact with sensitive information, for example to provide technical support or maintenance, and thus should be granted some limited functional immunity. Certain technical solutions are currently under way that could potentially address this issue;[38] some, like homomorphic encryption,[39] seem to be promising. However, their functionality, effectiveness and scalability still need to be fully tested.

In addition, due consideration may need to be given to the application of agreements for the sharing of data between the host country and third countries, as well as to the possibility that third countries may seek to access data held by technology companies through US CLOUD Act-type legislation and other relevant domestic laws having extraterritorial implications.[40]

Finally, such agreements should consider the implications of internet shutdowns for the digital operations of the organization and seek specific protections against them. The organization should, for example, seek to negotiate specific guarantees that all traffic directed to or from the humanitarian organization will not be blocked. This may not be sufficient, however, to ensure that beneficiaries can have access to humanitarian services provided by digital means in cases where entire mobile and telecommunications networks are shut down, where it is prohibited for affected populations to obtain SIM cards, or where mobile data traffic is restricted for them and where the problem is therefore not one of network traffic but one of network access. Alternative strategies will need to be developed by the organization in order to address these cases, as part of the development of the organization's cyber strategy.

In cases where the humanitarian organization processes data through third-party technology providers, such as a cloud solution provider, the organization would then need to ensure that any clarifications between itself and the host State, as highlighted above, are also reflected in the contractual arrangements with the technology company, to ensure that the company commits to defending them and the company's staff is prepared to give effect to them.

The legal measures described above are primarily aimed at ensuring an organization's independence. Indeed, safeguarding the confidentiality of an organization's data through its privileges and immunities plays an important role in ensuring that the organization can carry out its mandate effectively – and in the case of the ICRC, also in line with the Fundamental Principles of the Movement. In this sense, it is important to stress that solutions which may be described as very secure, and accepted as such in highly regulated industries characterized by high sensitivity of data and confidentiality requirements (such as the banking industry), may nonetheless be totally inadequate for use by the

38 See Microsoft, "Confidential Computing", available at: www.microsoft.com/en-us/research/theme/confidential-computing/.

39 See Andy Greenberg, "Hacker Lexicon: What Is Homomorphic Encryption?", *Wired*, 11 March 2014, available at: www.wired.com/2014/11/hacker-lexicon-homomorphic-encryption/.

40 See DoJ, above note 28.

ICRC, since such – very secure – solutions may still involve ways for organizations to be obliged to hand over data to States, may be subject to encryption back-door legal requirements, and so on.[41]

Legal protections are not enough: The technical protection a humanitarian organization is entitled to/can seek for its data, and for data flows

The legal protections described above, alone, are however insufficient to ensure that no authority can lawfully access the data of international humanitarian organizations. Three aspects are of particular concern in this sense: (1) surveillance practices are not always in line with privileges and immunities, (2) data traffic may also be caught and intercepted as part of large-scale/bulk data collection, and (3) the data of an organization may be hosted and processed through commercial technology providers.

The consequence of these issues is that an organization needs to act on two different levels. The first is the legal level, aiming to ensure that no third actor may successfully claim access to its data by application of the law; the second is the technical and organizational level, with specific measures aimed at ensuring secure data flows, hosting, and processing. As highlighted above, these measures may not, at present and for some types of cloud architectures, be available from the market, and may need to be procured as part of research and development partnerships with academia and other partners, to be then converted into sustainable solutions. Considering costs and available resources, it may be necessary for international humanitarian organizations to pool resources with other organizations with similar mandates and status, particularly to ensure the conversion into sustainable solutions of the research and development technical aspects.

The operational dialogue deployed by the organization

As highlighted in detail above, an organization like the ICRC seeks to establish its presence and work based on acceptance, and this in turn is based on the trust that derives from its neutrality, impartiality and independence, from the fact that it furthers exclusively humanitarian objectives, and from its confidential approach. In this sense, being able to establish bilateral, confidential dialogue with all stakeholders, irrespective of whether they are State or non-State actors and whether they may be accepted as lawful groups or not, is an essential requirement in order to ensure performance of the mandate.

41 See, for example, Julia Carrie Wong, "US, UK and Australia Urge Facebook to Create Backdoor Access to Encrypted Messages", *The Guardian*, 4 October 2019, available at: www.theguardian.com/technology/2019/oct/03/facebook-surveillance-us-uk-australia-backdoor-encryption.

These are the features that shape the dialogue which the organization, in this case the ICRC, needs to have, also in the cyber realm.

Dialogue with "cyber host States"

Developing and deploying digital humanitarian services, as discussed, requires an organization to identify one or more key jurisdictions where it can safely host such services and procure the necessary ingredients to then offer them globally. These "cyber host States" are likely to be stable countries where no active conflict or other situation of violence is present and where, therefore, the humanitarian organization would otherwise be unlikely to run any humanitarian programmes. They are likely to be identified among technologically advanced countries with a strong cyber industry, capabilities, academia and infrastructure. One example is the recently updated agreement between the ICRC and the Swiss Confederation.[42]

Operational dialogue with cyber host States is framed, first of all, in the host State agreement itself, and any further memoranda of understanding, documents, or practices existing between the two. This dialogue should be shaped as to cover, at least, the aspects set out below.

First, dialogue should address potential cooperation regarding the anticipation, detection and attribution (an essential precondition to bilateral confidential dialogue) of cyber operations, as well as the identification of the appropriate response to them. Because of its control over the network on its territory and flows of data going through it, the resources and expertise available, and the international cooperation networks it is likely to be involved in, a cyber host State may have much better means than the organization alone to anticipate, detect, attribute and respond to cyber operations. Defining the perimeters of this dialogue will be a very sensitive task and will be important in order to ensure that, on the one hand, the dialogue is effective, while, on the other hand, it does not make the organization over-reliant on the cooperation of the cyber host State, thereby creating a risk that the neutrality, impartiality and independence of the organization will be compromised.

Second, due consideration must be given to how to deal with "cyber criminals" – i.e., cases in which an operation affecting the organization is attributed to criminal groups and is not linked to State or State-sponsored actors. To what extent can or should the organization rely on law enforcement by the host State to protect its activities, and what type of cooperation does this require? How can the organization and the host State deal with the cross-border and international nature of cyber criminals, whereby the cyber criminals may not be found in the jurisdiction of the host State, and the impact of the action may reveal itself in third countries where the organization deploys its humanitarian action? What types of international cooperation mechanisms does the host State

42 ICRC and Swiss Federal Council, "Accord entre le Conseil fédéral suisse et le Comité international de la Croix-Rouge en vue de déterminer le statut juridique du Comité en Suisse", 19 March 1993, available at: www.fedlex.admin.ch/eli/cc/1993/1504_1504_1504/fr#sidebarLink.

engage in, and are these suitable for the nature, mandate and working modalities of the organization?

Third, the dialogue should also clarify how to deal with adverse cyber operations attributed to third countries, including by State-sponsored actors. This is also a sensitive area that may need to be specifically discussed and agreed between the organization and the host State, since it may raise sensitive questions of public international law and international relations. These questions may relate to the violation of sovereignty of the host State, possible countermeasures available to the host State, and reliance on due diligence obligations of third countries under international law to support bringing the adverse operation to an end, on the one hand, and the neutrality, impartiality and independence of the organization, on the other.

While some of these questions, relating in particular to a host State's failure to assist an international organization and the availability of countermeasures, have been analyzed in detail,[43] many others remain. In particular, while questions around sovereignty, countermeasures and due diligence in cyberspace have been discussed in different fora[44] and in certain governments' cyber security policies and/or statements,[45] these have so far looked more at the implications on sovereignty when it comes to operations impacting on the territory of the State affected, and, with the notable exception mentioned above, not so much when it comes to the relationship between an international organization and its host State. In this area, different States may have different and diverging views as to how they interpret those concepts, and some may not have a clear, public position as to their interpretation of this area of law. It is therefore important to ensure that questions which may affect an organization's capacity to operate are addressed by it in its dialogue with its host State.

In other words, would a cyber host State consider an operation targeting an organization hosted on its territory as a violation of its sovereignty? If so, under what conditions? Could the cyber host State in that case seek countermeasures against the perpetrators? If so, which countermeasures? If the operation is being run through infrastructure on the territory of a third State, would the cyber host State seek to get the cooperation of the third State in order to bring the operation to an end? Would the cyber host State refer to a due diligence obligation of the third State to bring the operation to an end? Would any of the above constitute a concern for the organization, insofar as the intervention of the cyber host State may affect and compromise its neutrality, impartiality and independence?

43 See "Scenario 04: A State's Failure to Assist an International Organization", in Kubo Mačák, Tomáš Minárik and Taťána Jančárková (eds), *Cyber Law Toolkit*, available at: https://tinyurl.com/3m4nm6nv.

44 See, for example, Michael N. Schmitt and Liis Vihul (eds), *Tallinn Manual 2.0 on International Law Applicable to Cyber Operations*, 2nd ed., Cambridge University Press, Cambridge, 2017, available at: https://ccdcoe.org/research/tallinn-manual/.

45 See French Ministry of Defence, *International Law Applied to Operations in Cyberspace*, 2019, available at: www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf.

Dialogue with the State/government where the organization wants to deploy/offer digital services

For an organization like the ICRC, working in areas of armed conflict and other situations of violence, dialogue with the State in which it would operate is an essential step to ensuring acceptance of the deployment of digital humanitarian services.

This is not an anodyne statement, particularly taking into account that, as set out above, such services must be exclusively humanitarian services, and offered in a neutral, impartial and independent way. This means that affected people expect that any communication with or data provided to the humanitarian organization will not be accessed and used by third parties for non-humanitarian purposes. Similarly, the State in question must accept this protected digital humanitarian space and not interfere with it or with the technical measures used by the humanitarian organization to protect it.[46]

Similarly, this dialogue should also aim at ensuring that "humanitarian data flows" directed to the organization are not affected by internet shutdowns, and that affected populations have access, to the maximum extent possible, to connectivity.

Dialogue with State and State-sponsored attackers

Securing the organization's cyber perimeter against the technical capabilities of State-led or State-sponsored attackers, and in some cases also of certain groups linked to non-State armed groups, is a major challenge. A humanitarian organization will most likely never have sufficient resources to counter the offensive power of these types of adversaries. From the point of view of an organization like the ICRC, which bases its security on acceptance and respect of its humanitarian mandate, the primary objective would be to ensure acceptance of a protected digital humanitarian space.

Just like the organization routinely does in the non-digital world, this requires it to consider how to securely carry out a bilateral confidential dialogue with States, State-sponsored groups and groups linked to non-State armed groups with sophisticated capabilities, potentially including hacker groups, in order to explain its work, mandate and *modus operandi*, to establish respect for its digital humanitarian space, to prevent adverse cyber operations and, thereby, to negotiate and obtain "digital access". In this respect, key questions will arise as to how, technically, the organization can in practice set up a bilateral confidential

46 See Group of Friends of the Protection of Civilians in Armed Conflict, statement submitted to the UN Security Council Arria-Formula Meeting on Cyber-Attacks against Critical Infrastructure, New York, 26 August 2020, available at: www.eda.admin.ch/dam/mission-new-york/en/speeches-to-the-un/2020/20200826-new-york-POC-GoF%20PoC%20statement_E.pdf. "The trust of the people they serve is the currency of humanitarian organizations. This trust is a precondition for humanitarian action. Therefore, we, as Members [*sic*] States, must create an environment, including a safe information infrastructure that allows humanitarian organizations to successfully carry out their mandate. The Resolution on Restoring Family Links adopted at the 33rd International Conference of the Red Cross and Red Crescent in 2019 constitutes an important step in this direction."

dialogue with these actors, and in particular with State-sponsored hacker groups (and be sure it is with them that it is indeed having the dialogue). In order to maintain the trust of all stakeholders in the international community, it is also important that the organization is transparent about the existence, reasons and objectives of this dialogue. As explained on the ICRC web pages that clarify who the ICRC engages in dialogue with, and why:

> It is those who carry weapons who can kill – and be killed. It is also they who can facilitate or hinder humanitarian action. The ICRC therefore maintains a dialogue with all weapon bearers, State and non-State, as part of our mandate to protect and assist people affected by war and other violence.[47]

This is true both in the physical world and in cyberspace.

This confidential dialogue should be complemented with state-of-the-art security[48] and, where possible, research and development partnerships with academia to go one step further than state-of-the-art security. Although it is most likely very difficult to ensure security at a level sufficient to counter a State-sponsored actor in all circumstances, the level of security to be put in place should be guided by (1) due diligence – i.e., applying a level of security that can be expected from an organization handling highly sensitive data, taking into account the cost of technology, sensitivity of the information, and state of the art; and (2) the aim of raising the cost (in terms of financial resources, time, and staff required to carry out adverse cyber operations, as well as reputational repercussions) of adverse operations that successfully affect the organization, to a level that such operations are not worth the cost of achieving them. It is suggested that a combination of these two elements is necessary to ensure effective protection.

Conclusion

An international humanitarian organization going through a process of digital transformation and aiming to offer digital services directly to beneficiaries faces numerous questions that are extremely novel. These questions range from the legal and organizational to the technical and operational, and relate to issues that are transversal and highly interdependent – and at present none of them have any clear and unequivocal answers.

It is fundamental, therefore, that any organization which becomes an actor in cyberspace carries out an in-depth analysis of the questions discussed in the present paper and identifies the answers that are suitable for the organization based on its status, mandate and working modalities. Furthermore, these answers

47 See ICRC, "Dialogue with Weapon Bearers", available at: www.icrc.org/en/what-we-do/building-respect-ihl/dialogue-weapon-bearers.

48 See ENISA, "What Is 'State of the Art' in IT Security?", 7 February 2019, available at: www.enisa.europa.eu/news/enisa-news/what-is-state-of-the-art-in-it-security.

need to be formulated in a clear cyber security strategy informing the organization's stance in cyberspace, as well as its decisions to prioritize investment areas and its allocation of resources.

In addition to a cyber strategy developed on these bases, international humanitarian organizations need to consider unique and specific technical solutions to their specificities, such as the creation of a "digital humanitarian space" along the model of a "sovereign cloud" or a "digital embassy". These do not currently exist as part of any commercial offering, primarily because technological commercial offerings are developed based on the demands of the majority of customers, who, unlike international humanitarian organizations, are not entitled to rely on privileges and immunities from the jurisdictional control of at least one State.

Partnerships with academia and industry are an important part of this effort, but they alone are not sufficient – what is essential is both (1) wider political will on the part of external stakeholders to guarantee the protection of a digital humanitarian space, and (2) the awareness, knowledge, focus and determination of internal stakeholders to genuinely preserve the independence, impartiality and neutrality of international humanitarian organizations in cyberspace. Without this, international humanitarian organizations will inevitably be pushed into accepting solutions that are unsuitable for the work they are mandated to carry out.

International Review of the Red Cross (2020), **102** (913), 389–416.
Digital technologies and war
doi:10.1017/S1816383121000035

IRRC

The updated ICRC Commentary on the Third Geneva Convention: A new tool to protect prisoners of war in the twenty-first century

Jemma Arman, Jean-Marie Henckaerts, Heleen Hiemstra and Kvitoslava Krotiuk*

Jemma Arman is a Regional Legal Adviser for the ICRC based in Nairobi, Jean-Marie Henckaerts is Head of the Commentaries Update Unit of the Legal Division of the ICRC, Heleen Hiemstra is a Legal Adviser in the Commentaries Update Unit, and Kvitoslava Krotiuk is an Adviser in the Office of the President of the ICRC. Jemma Arman and Kvitoslava Krotiuk were Legal Advisers in the Commentaries Update Unit before their current assignments.

Abstract

Since their publication in the 1950s and 1980s respectively, the Commentaries on the Geneva Conventions of 1949 and their Additional Protocols of 1977 have become a major reference for the application and interpretation of those treaties. The International Committee of the Red Cross, together with a team of renowned experts, is currently updating these Commentaries in order to document developments and provide up-to-date interpretations of the treaty texts. This article highlights key points of interest covered in the updated Commentary on the Third

* The authors wish to acknowledge that this article summarizes some of the key findings of the updated Commentary and as such reflects the input of many experts involved in the drafting and review of the Commentary, including the authors' colleagues in the Commentaries Update Unit Bruno Demeyere, Yvette Issar, Eve La Haye and Heike Niebergall-Lackner.

Geneva Convention. It explains the fundamentals of the Convention: the historical background, the personal scope of application of the Convention and the fundamental protections that apply to all prisoners of war (PoWs). It then looks at the timing under which certain obligations are triggered, those prior to holding PoWs, those triggered by the taking of PoWs and during their captivity, and those at the end of a PoW's captivity. Finally, the article summarizes key substantive protections provided in the Third Convention.

Keywords: international humanitarian law, Geneva Convention III, updated Commentary, prisoners of war, internment, captivity, Detaining Power, humane treatment, protection of person and honour, equal treatment, non-discrimination, principle of assimilation, transfer, release and repatriation, seriously wounded and sick prisoners, quarters, food, clothing, medical care and sanitation, recreation, religion, relations with the exterior, labour, complaints, prisoners' representatives, disciplinary and judicial proceedings.

In 2011, the International Committee of the Red Cross (ICRC) embarked on an ambitious project to update the Commentaries on the Geneva Conventions of 1949 and their Additional Protocols of 1977.[1] The updated Commentaries seek to reflect developments in how the law is applied and interpreted in practice, recognizing that over seventy years have passed since the Geneva Conventions were adopted. Previous milestones of this project include the completion of the updated Commentaries on Geneva Conventions I and II (GC I and GC II) in 2016 and 2017 respectively.[2] In 2020, the project reached another major milestone with the completion of the updated Commentary on Geneva Convention III relative to the Treatment of Prisoners of War (GC III).[3]

GC III protects members of the armed forces and other defined categories of persons who fall into the power of the enemy in times of international armed

1 Jean Pictet (ed.), *Commentary on the Geneva Conventions of 12 August 1949*, Vols 1–6, ICRC, Geneva, 1952–60. The ICRC has engaged in the writing of the original Commentaries, and the updating of those Commentaries, pursuant to its role as guardian and promoter of international humanitarian law (IHL). This role is recognized in the Statutes of the International Red Cross and Red Crescent Movement, in particular the ICRC's role "to work for the understanding and dissemination of knowledge of international humanitarian law applicable in armed conflicts and to prepare any development thereof". See Statutes of the International Red Cross and Red Crescent Movement, 1986, Arts 5(2)(c), 5(2)(g). On the ICRC's role in the interpretation of IHL, see also François Bugnion, *The International Committee of the Red Cross and the Protection of War Victims*, ICRC and Macmillan Education, Oxford, 2003, pp. 914–922.

2 For more details, see Bruno Demeyere, Jean-Marie Henckaerts, Heleen M Hiemstra and Ellen Nohle, "The Updated ICRC Commentary on the Second Geneva Convention: Demystifying the Law of Armed Conflict at Sea", *International Review of the Red Cross*, Vol. 98, No. 902, 2016; Lindsey Cameron, Bruno Demeyere, Jean-Marie Henckaerts, Eve La Haye and Heike Niebergall-Lackner, "The Updated Commentary on the First Geneva Convention – a New Tool for Generating Respect for International Humanitarian Law", *International Review of the Red Cross*, Vol. 97, No. 900, 2015.

3 ICRC, *Commentary on the Third Geneva Convention: Convention (III) relative to the Treatment of Prisoners of War*, 2nd ed., 2020 (ICRC Commentary on GC III), available at: https://ihl-databases.icrc.org/ihl/full/GCIII-commentary (all internet references were accessed in January 2021). The print edition of the updated Commentary will be published by Cambridge University Press in 2021.

conflicts.[4] Prisoners of war (PoWs) are not to be punished for their mere participation in hostilities; their detention is not a punishment but an act to prevent their further participation in hostilities. This understanding underpins the whole of GC III.[5]

GC III sets out a number of fundamental protections that apply to all PoWs. These fundamental protections serve as a foundation for the more prescriptive articles, which provide that PoWs must at all times be treated humanely, with respect for their person and their honour, and treated equally, without discrimination.[6] These principles in turn are supplemented by detailed provisions regulating the treatment of PoWs. These include provisions relating to the beginning of captivity, the provision of prisoners' basic needs, the transfer of prisoners, the use of prisoners' labour, the imposition of disciplinary or judicial proceedings, and the final release and repatriation of prisoners. The level of detail provided for the protection of PoWs at the time of drafting GC III in 1949 was unprecedented, and GC III continues to provide comprehensive protection to PoWs.

Updating the Commentaries on each of the 142 articles of GC III required consideration of a wide range of historical, legal, military, ethical, socio-cultural and technological issues. As with the updated Commentaries on GC I and GC II, the development of the updated Commentary on GC III involved a collaborative effort, with input from ICRC and non-ICRC lawyers, specialists with subject-matter expertise (including military personnel, protection officers specializing in detention, and academics), and others. In addition, the development of this Commentary has benefited from the fact that the ICRC has been able to draw on archival records of its work visiting PoWs over the last seventy years. This work has enabled the ICRC to witness measures taken to comply with GC III, and also challenges in its implementation.

GC III remains relevant today, as there continue to be prisoners of war. Its rules have informed parallel provisions protecting civilian internees under Geneva Convention IV (GC IV). No article of GC III was found to have fallen into desuetude, although it was sometimes more difficult to find recent practice in relation to certain topics, such as the financial resources of PoWs.[7]

The update of the Commentary on GC III follows the same methodology as that applied for the updated Commentaries on GC I and GC II, based on the rules of treaty interpretation set out in the 1969 Vienna Convention on the Law of Treaties, in particular Articles 31–33.[8] Pursuant to these rules, the contributors started from

4 It should be noted that in non-international armed conflicts, IHL foresees no entitlement to PoW status as it exists for international armed conflict.

5 ICRC Commentary on GC III, above note 3, Introduction, para. 20, and Art. 21, para. 1932.

6 Geneva Convention (III) relative to the Treatment of Prisoners of War of 12 August 1949, 75 UNTS 135 (entered into force 21 October 1950) (GC III), Arts 13, 14, 16.

7 In international armed conflicts since 1949, Article 61 on supplementary pay for PoWs does not appear to have been resorted to. On absence of practice and desuetude, see also ICRC Commentary on GC III, above note 3, Introduction, section C.8.

8 Vienna Convention on the Law of Treaties, 1155 UNTS 18232, 23 May 1969. Articles 31–33 are generally considered to reflect customary international law. See, for example, International Court of Justice (ICJ), *Kasikili/Sedudu Island (Botswana v. Namibia)*, Judgment, 13 December 1999, *ICJ Reports 1999*, paras

the ordinary meaning of the terms of GC III in their context and in light of the object and purpose of the treaty. Although the updated Commentary has been drafted in English, the authors have consistently consulted and compared the English text of the Convention with the French text, which is equally authentic.[9] Close examination was also made of the preparatory work for each article of the Convention.

Where relevant, the updated Commentary also takes into account developments in branches of international law other than international humanitarian law (IHL), such as international criminal law and international human rights law. Other treaties are referred to on the understanding that they apply only to States which have ratified or acceded to them, and only if the conditions relating to their geographic, temporal and personal scope of application are fulfilled. Reference is made to international human rights law where relevant to interpret shared concepts (for example, cruel, inhuman and degrading treatment), as well as to provide practitioners with further information about certain topics, and in certain circumstances where GC III may be affected by international human rights obligations.[10] This does not mean that international human rights law and interpretations can be transposed mechanically to IHL provisions, and differences have also been pointed out where relevant.[11]

This article highlights key points of interest covered in the updated Commentary on GC III. It is divided into three parts. The first part covers the fundamentals of GC III: the historical background, the personal scope of application of the Convention, and the fundamental protections that apply to all PoWs. The second part provides a framework for understanding when certain obligations are triggered; these may be broadly grouped as the obligations of a Detaining Power prior to holding PoWs, the obligations triggered by the taking of PoWs and during their captivity, and the obligations that arise at the end of a PoW's captivity. The third part summarizes key substantive protections provided in GC III, providing examples of the depth of detail in the Convention when it comes to the protection of PoWs.

18–20; ICJ, *Case Concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro)*, Merits, Judgment, 26 February 2007, *ICJ Reports 2007*, para. 160; International Law Commission, "Subsequent Agreements and Subsequent Practice in Relation to the Interpretation of Treaties", Conclusion 2.1 (adopted on second reading), *Report of the International Law Commission on the Work of its Seventieth Session*, UN Doc. A/73/10, 2018, p. 13.

9 GC III, Art. 133; Vienna Convention on the Law of Treaties, Art. 33.

10 For example, a discussion on the application of Article 100 on the death penalty would not be complete without acknowledging the existence of international treaties by which many States have committed to abolishing the death penalty. See ICRC Commentary on GC III, above note 3, Art. 100, para. 3979.

11 For an example in relation to the definition of torture, see *ibid.*, Art. 3, section G.2, and Art. 130, section D.2.a. For more information on the use of other relevant rules of international law, see *ibid.*, Introduction, section C.5.

The fundamentals of Geneva Convention III

The historical background of Geneva Convention III

Customs and codes regulating the capture and detention of enemy soldiers have existed for thousands of years, drawing on a variety of cultural, religious and ethical frameworks.[12] The development of an international treaty, however, began in earnest in the eighteenth and nineteenth centuries, at which time many States began to establish and consolidate professional armies, to enter bilateral agreements regarding the conditions of warfare,[13] and to include protections for PoWs in their military manuals.[14]

In 1874, a conference of fifteen European States adopted a draft text submitted by the Russian government, now known as the Brussels Declaration, which included twelve articles on the protection of PoWs. The Brussels Declaration never became a binding treaty, but many of its definitions were adopted essentially without change at the 1899 Hague Peace Conference.[15] The Hague Regulations annexed to the Hague Conventions of 1899 were the first binding multilateral agreement dealing with PoWs.[16] Seventeen articles of the Regulations dealt with PoWs, addressing, *inter alia*, the obligation to treat prisoners humanely and without distinction, to feed and clothe prisoners at a standard at least on par with the soldiers of the Detaining Power, and to ensure speedy repatriation of prisoners upon the end of the conflict.[17]

The provisions in the Hague Regulations proved to be insufficiently detailed, and during World War I some belligerents signed temporary agreements to clarify disputed points.[18] Further, the changing character of warfare, technological developments and the increased size of armies and wars led to significantly larger numbers of persons being taken captive during armed

12 *Ibid.*, Introduction, para. 4.

13 For example, during the Napoleonic Wars, the United Kingdom and France entered into an agreement which allowed for a "protecting power" to visit prisoners and provide additional food. In 1896 Italy and Ethiopia entered into the Treaty of Addis Ababa, which included the requirement of release of all prisoners, as well as an obligation on the part of Ethiopia to allow a detachment of the Italian Red Cross to facilitate this process. Alexander Gillespie, *A History of the Laws of War*, Vol. 1: *The Customs and Laws of War with Regards to Combatants and Captives*, Hart Publishing, Oxford, pp. 149, 164; James Molony Spaight, *War Rights on Land*, Macmillan, London, 1911, p. 37.

14 See Allan Rosas, *The Legal Status of Prisoners of War: A Study in International Humanitarian Law Applicable in Armed Conflicts*, Institute for Human Rights, Åbo Akademi University, Turku/Åbo, 1976 (reprinted 2005), pp. 69, 72–73; and, in particular, the Lieber Code of 1863: Instructions for the Government of Armies of the United States in the Field, prepared by Francis Lieber, promulgated as General Order No. 100 by President Abraham Lincoln, Washington, DC, 24 April 1863.

15 Regulations concerning the Laws and Customs of War on Land, Annexed to Convention (II) with respect to the Laws and Customs of War on Land, The Hague, 29 July 1899, Section II. Provisions dealing with PoWs can also be found in Institute of International Law, *The Laws of War on Land*, Oxford, 9 September 1880, e.g. Arts 21–22, 61–78.

16 A. Rosas, above note 14, p. 70.

17 ICRC Commentary on GC III, above note 3, Introduction, para. 7.

18 See, for example, the Agreement between the British and Ottoman Governments respecting Prisoners of War and Civilians, signed in Bern in December 1917 (HM Stationery Office, London, 1918).

conflicts of this period, most notably in World War I.[19] On the basis of general principles developed by the Tenth International Conference of the Red Cross and Red Crescent, the 1929 Convention relative to the Treatment of Prisoners of War was adopted, considerably supplementing the Hague Regulations.[20] Its eighty substantive articles included provisions on the prohibition of measures of reprisal and collective penalties, the organization of labour of PoWs, the ability of prisoners to elect their representatives, the codification of judicial procedures and punitive measures, and the official recognition of the role of the ICRC, generally and in regard to the organization of a central information agency. Forty-seven States were party to the 1929 Convention at the outbreak of World War II.[21] While the protections conferred by the 1929 Convention had an important impact in several theatres of World War II, in others they did not, in part because they were interpreted not to be applicable. For example, a narrow interpretation of the definition of PoW was used to deny PoW status to soldiers of several countries who surrendered following the capitulation of their State.[22]

The negotiations for what would become GC III were in turn heavily influenced by the experiences of World War II. As in World War I, quarter was regularly denied on a devastating scale.[23] Further, World War II witnessed the use of detention itself as a means to enable the killing of innumerable soldiers, including by summary execution, extreme acts of violence, ill-treatment, starvation and malnutrition.[24] Prisoners were treated differently depending on their nationality and on which State detained them, and at the end of the war, the repatriation of prisoners was significantly drawn out.[25]

In the immediate aftermath of World War II, several expert conferences were convened in Geneva, where preparatory material gathered by the ICRC and first drafts for the new conventions were discussed. The most important of these conferences were the Preliminary Conference of National Red Cross Societies in 1946 and the Conference of Government Experts in 1947. The drafts prepared by these conferences were presented to the 1948 International Conference of the Red Cross and Red Crescent in Stockholm, where further amendments were adopted. The Stockholm Drafts served as the basis for negotiation at the Diplomatic Conference that met in Geneva from 21 April to 12 August 1949. Fifty-nine

19 During World War I, for example, it is estimated that an unprecedented 7 to 8 million soldiers were taken as PoWs. On treatment issues for PoWs in World War I, see A. Gillespie, above note 13, pp. 166–172.

20 See François Bugnion, above note 1, p. 121, for more detail on the preparatory steps that led to the adoption of the 1929 Convention.

21 In addition, Japan declared that it was ready to apply the Convention during World War II "under conditions of reciprocity and *mutatis mutandis*". ICRC, *Report on Activities during the Second World War*, Vol. 1, Geneva, 1948, p. 229.

22 ICRC Commentary on GC III, above note 3, Art. 4, section D.1. See also *ibid.*, para. 1041, in relation to the denial of PoW status to soldiers of governments or authorities not recognized by the Detaining Power.

23 See, for example, A. Gillespie, above note 13, p. 186.

24 See, for example, A. Rosas, above note 14, p. 78; A. Gillespie, above note 13, pp. 192–200; Sandra Krähenmann, "Protection of Prisoners in Armed Conflict", in Dieter Fleck (ed.), *The Handbook of International Humanitarian Law*, 3rd ed., Oxford University Press, Oxford, 2013, p. 362.

25 For example, it is estimated that there were still 630,000 German prisoners of war in France in 1947. S. Krähenmann, above note 24, p. 363.

States were officially represented by delegations with full powers to discuss the texts; four States sent observers.

In general terms, the 1949 GC III is considerably more detailed than the 1929 Convention. It clarifies and expands the scope of persons to whom it applies; it provides clearer regulation to keep prisoners in good health; it elaborates on the guarantees they are due in cases of disciplinary or penal sanction; it provides stricter regulation on the use of PoW labour; and it clarifies the obligation to repatriate prisoners at the end of active hostilities. Like the other three Geneva Conventions, GC III also contains a system for the suppression of breaches of the Convention, by defining the concept of "grave breaches" against PoWs, by creating obligations on States to pass legislation criminalizing grave breaches, and by obliging States to search for and to try or extradite those who are suspected of having committed such breaches. It provides for a greater role for relief societies and acknowledges the "special position" of the ICRC in this respect. Finally, GC III allows for the ICRC to visit PoWs and forms the basis for its Central Tracing Agency.[26]

The personal scope of application of Geneva Convention III

Article 4 is perhaps the best known and most debated provision of GC III. This article defines PoWs and, accordingly, is central to understanding the personal scope of application of the Convention. It provides, in short, that a PoW is a person belonging to one of six categories defined in Article 4(A) at the time that they "fall into the power of the enemy" in an international armed conflict.[27] Article 4(A) mirrors the list of protected persons in Article 13 of GC I and GC II, which provide protection for wounded, sick and shipwrecked military personnel. Wounded, sick and shipwrecked persons covered by GC I or GC II who fall into the power of the enemy are simultaneously protected by GC III as well as GC I or CG II.[28]

Members of the armed forces

The first of the six categories is "members of the armed forces". Numerically, this is likely to be the most significant category. "Members of the armed forces" refers to all military personnel under a command responsible to a party to the conflict. The requirement for membership in the armed forces is not prescribed in

26 The requirement that the ICRC be allowed to visit "all places where prisoners of war may be" is provided for in GC III, Art. 126. The creation of a Central Tracing Agency, operating under the responsibility of the ICRC, is established in GC III, Art. 123.

27 For a discussion on the expression "fallen into the power of the enemy", see ICRC Commentary on GC III, above note 3, Art. 5, paras 1100–1101.

28 For details, see Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of 12 August 1949, 75 UNTS 31 (entered into force 21 October 1950) (GC I), Art. 14; Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea of 12 August 1949, 75 UNTS 85 (entered into force 21 October 1950) (GC II), Art. 16.

international law and is instead a matter of domestic regulation.[29] This first category also includes members of militia or volunteer corps forming part of the armed forces – that is to say, formally incorporated into the armed forces and under the responsible command of a party to the conflict.[30] It may also include paramilitary and armed law enforcement agencies that are formally incorporated into the armed forces through the national law of a State.[31] Members of the armed forces are required to distinguish themselves from the civilian population during military operations. Under customary IHL, failure to do so while engaged in an attack or in a military operation preparatory to an attack results in their forfeiting the right to PoW status.[32] This provides an example of how the updated Commentary refers to customary IHL where it may be considered a "relevant [rule] of international law applicable in the relations between the parties".[33]

Members of other militias and members of other volunteer corps, including those of organized resistance movements, belonging to a party to the conflict and fulfilling the four conditions

The second category of PoWs consists of members of "other militias and members of other volunteer corps, including those of organized resistance movements, belonging to a Party to the conflict"[34] and fulfilling four prescribed conditions. This category concerns groups that are not incorporated into the armed forces but otherwise "belong" to a party to the conflict.

A group belongs to a party to the conflict for the purpose of Article 4(A)(2) if the group fights on behalf of that party and that party accepts this fighting role. This acceptance can be express – for example, when a party gives a formal authorization to the group or acknowledges that the group fights on its behalf. It can also be implicit or tacit, such as when a group fights alongside the State and claims to be fighting on its behalf and the State does not deny this relationship when given the opportunity. The acceptance of a "belonging to" relationship can also be demonstrated by the overall control that the party exercises over the group.[35]

For members of such militia or volunteer corps to be considered PoWs upon falling into the power of the enemy, the militia and volunteer corps must collectively fulfil four conditions, each of which serves a protective purpose: they must be commanded by a person responsible for his or her subordinates, they

29 ICRC Commentary on GC III, above note 3, Art. 4, para. 977.
30 *Ibid.*, Art. 4, para. 979.
31 *Ibid.*, Art. 4, paras 979–982.
32 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 106, available at: https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1. For further discussion on this point, see ICRC Commentary on GC III, above note 3, Art. 4, paras 983–987.
33 Vienna Convention on the Law of Treaties, Art. 31(3). See also ICRC Commentary on GC III, above note 3, Introduction, paras 92–95.
34 GC III, Art. 4(A)(2).
35 For a more in-depth discussion on the meaning of "belonging to" under Article 4(A)(2), see ICRC Commentary on GC III, above note 3, Art. 4, paras 1001–1009.

must have a fixed distinctive sign recognizable at a distance, they must carry arms openly, and they must conduct their operations in accordance with the laws and customs of war. A structured hierarchy can ensure internal discipline and that operations are planned, coordinated and carried out in a way that is consistent with the laws and customs of war. Having a fixed distinctive sign and carrying arms openly facilitates the distinguishing of combatants from the civilian population. The condition that the militia or volunteer corps conduct operations in accordance with the laws and customs of war serves as an additional encouragement for the groups to comply with IHL, in order for their members to receive protective PoW status in the event that they fall into the power of the enemy.

The Commentary also considers the question of whether these four conditions, which appear in Article 4(A)(2) but not in 4(A)(1), also apply to 4(A)(1) forces.[36] In the ICRC's view, while the four conditions are obligations for regular armed forces, they are not *collective conditions* for PoW status.[37] The four conditions reflect the usual practice of State armed forces. By definition, such forces are commanded by a person responsible for their subordinates.[38] Further, as mentioned above, members of the armed forces are under an obligation to distinguish themselves sufficiently from the civilian population and not to conceal their weapons during military operations. The ICRC's understanding is that a combatant loses eligibility for PoW status if he/she fails to distinguish him/herself. Such loss of eligibility, however, applies only on an individual basis and not to the group as a whole.[39] Compliance with the laws and customs of war is a standard requirement under the Geneva Conventions and general international law,[40] and Article 85 of GC III makes it clear that PoWs keep their protected status if convicted for acts committed prior to capture.[41] If and when regular armed forces are perceived as not fulfilling these obligations, avenues other than a collective denial of PoW status are available to States under international law to induce compliance.[42] The ICRC recognizes, however, that there are diverging views as to whether the four conditions under Article 4(A)(2) are collective conditions for PoW status for members of a State's regular armed forces.[43]

Members of regular armed forces who profess allegiance to a government or authority not recognized by the detaining power

The third category consists of members of regular armed forces who profess allegiance to a government or authority not recognized by the detaining power. Members of the regular armed forces of a party to an international armed conflict

36 *Ibid.*, Art. 4, paras 1028–1039.
37 *Ibid.*, para. 1039.
38 *Ibid.*
39 *Ibid.*, paras 983, 1039.
40 *Ibid.*, para. 1039.
41 See also *ibid.*, para. 1033.
42 *Ibid.*, para. 1039.
43 *Ibid.*, para. 1036.

are included within the definition of PoWs under the first category described above, but World War II saw the denial of PoW status to certain groups on the basis that the authorities or governments to whom those armed forces pledged allegiance were not recognized by the enemy State.[44] To avoid a repetition of this abusive interpretation, the definition of PoWs in GC III expressly includes all members of regular armed forces, irrespective of whether the enemy recognizes the legitimacy of their government or authority.[45]

Persons authorized to accompany and in fact accompanying the armed forces without being members thereof

The fourth and fifth categories of PoWs are the only two categories of persons entitled to PoW status without equally being entitled to combatant status, immunity or privileges. The fourth category consists of persons authorized to accompany and in fact accompanying the armed forces without being members thereof.[46] The inclusion of this category recognizes that the proximity of such persons to the armed forces increases the risks of their being interned with combatants and makes explicit the protective framework that applies to them. It might include, for example, civilian contractors authorized to accompany the armed forces providing services such as laundry or transportation.[47] The authorization of a person to accompany the armed forces is evidenced by the provision of an identity card of a similar model to that annexed to GC III in Annex IV(A), and could also be evidenced by co-location, shared logistical arrangements, contractual arrangements and/or apparel.[48]

Members of the crew of the merchant marine or civil aircraft of the parties to the conflict who do not benefit from other more favourable treatment in international law

The fifth category consists of members of the crew of the merchant marine or civil aircraft of the parties to the conflict who do not benefit from other more favourable treatment in international law. The inclusion of the crew of the merchant marine sought to remedy uncertainty as to their status and inconsistencies in the protection provided to such persons during World War II. Civilian members of aircraft crews were also included, recognizing the increasing role of aircraft in providing deliveries to combat areas.[49]

44 For example, PoW status was denied by Germany to French forces operating under the command of General de Gaulle, and to Italian units in southern Italy following the signing of an armistice between the Allies and Italy in September 1943. ICRC, *Preliminary Documents submitted by the ICRC to the Conference of Government Experts of 1947*, Vol. 2, Geneva, 1947, p. 4.

45 *Ibid.*, p. 4.

46 For more information, see ICRC Commentary on GC III, above note 3, Art. 4, paras 1047–1050.

47 For a more detailed discussion, see *ibid.*, Art. 4, paras 1047–1050. See para. 1051 in relation to private military and security companies.

48 *Ibid.*, para. 1050.

49 *Ibid.*, paras 1052–1060.

In relation to treatment, GC III makes no distinction between PoWs who are combatants and those who are civilians. However, some provisions of the Convention presume the existence of membership in the armed forces and are silent as to their application in relation to the other categories of PoWs. For example, certain provisions in relation to the use of PoW labour, such as the rate of payment, are framed around the rank of PoWs.[50] If a Detaining Power interns PoWs who are civilians, it must apply these provisions in good faith and in line with the rationale behind the provisions in question.[51]

Levée en masse

The sixth and final category of PoWs consists of participants in what is commonly referred to as a *levée en masse*. More precisely, this category comprises any inhabitants of a non-occupied territory who, on the approach of the enemy, spontaneously take up arms to resist the invading forces, without having had time to form themselves into regular armed units, provided they carry arms openly and respect the laws and customs of war. This is the only category of PoWs which is entirely autonomous from the State. The persons under this category do not "belong to" the State, nor do they require any level of organization, command structure or fixed distinctive sign.[52]

The circumstances in which the conditions for a *levée en masse* apply are limited. First, the term refers only to those that take up arms during an invasion period, where territory is not yet occupied, or in an area where the previous Occupying Power has lost control over the administration of the territory and is attempting to regain it. Second, the persons in question must have spontaneously taken up arms in response to the invading army. This category does not include persons or groups who organize or are organized in advance of the invasion. Finally, persons in this category must carry their arms openly and must respect the laws and customs of war.

In addition to these six categories, Article 4 also sets out two categories of persons who are not PoWs *per se*, but are to be treated as PoWs.[53] The definition in Article 4 is supplemented in this respect by Additional Protocol I (AP I) (and customary IHL), which excludes spies, saboteurs and mercenaries from PoW status.[54]

The drafters of GC III gave considerable attention to defining which categories of persons qualify for PoW status, and in many ways reduced the uncertainties that existed within previous definitions under the Hague Regulations and 1929 Convention. Notwithstanding this, doubt as to the status of persons may still arise. An important innovation in GC III was to provide a

50 See GC III, Art. 60.
51 ICRC Commentary on GC III, above note 3, Art. 4, para. 1046.
52 *Ibid.*, para. 1062.
53 See GC III, Art. 4(B)(1)–(2); ICRC Commentary on GC III, above note 3, paras 1069–1090.
54 For more information on the exclusion of "spies and saboteurs", see ICRC Commentary on GC III, above note 3, paras 988–991. For more information on the exclusion of "mercenaries", see para. 998.

mechanism to address these situations. Article 5(2) of GC III provides that in case of doubt regarding the status of persons who have committed belligerent acts and fall into the hands of the enemy, such persons enjoy the protection of the Convention until a determination of their status has been made by a competent tribunal. The term "competent tribunal" was employed in order to encompass review by a court or military tribunal, and to prevent "arbitrary decisions [being made] by a local commander, who may be of a very low rank".[55] In practice, the status of individuals has been decided by civil courts, military tribunals or courts, and boards of inquiry.[56]

A determination should be made within a reasonable time frame, in good faith, and on a case-by-case basis; the requirement that determinations be made by a "competent tribunal" prevents arbitrary, "on-the-spot" decision-making. That noted, the particular procedural guarantees applicable to status determinations are not regulated by IHL and are a matter of domestic law.[57]

Doubt as to a person's status arises when it is not clear whether the person belongs to any of the categories discussed above. For example, it can arise in relation to persons who accompany the armed forces and have lost their identity card, persons engaged in belligerent acts without wearing a uniform in zones of active hostilities, or persons suspected of being spies. It may also arise where a person or the Power on which he or she depends asserts PoW status and this is not immediately accepted. Conversely, it may arise where a person asserts that they are not a PoW. The existence of a doubt that triggers a determination by a competent tribunal must not depend solely on the subjective belief of the Detaining Power; rather, a Detaining Power must consider each situation in good faith, on a case-by-case basis, with a proper assessment of the facts.[58]

Any person determined to be a PoW will continue to enjoy the protections of GC III. A person determined not to fall within the categories of Article 4 of GC III will otherwise be considered a civilian and is protected by GC IV (including Articles 43 and 78), and/or Article 75 of AP I, as applicable, and customary IHL.[59]

Fundamental principles for the protection of prisoners of war

GC III embodies a balance between the requirements of humanity and military necessity. Its overall object and purpose is to ensure that PoWs are humanely treated at all times, while allowing belligerents to intern captured enemy

55 *Final Record of the Diplomatic Conference of Geneva of 1949*, Vol. II-B, 1949, p. 270. See also H. W. William Caming, "Nuremberg Trials: Partisans, Hostages and Reprisals", *Judge Advocate Journal*, Vol. 4, 1950, p. 19, in relation to the infamous Barbarossa Jurisdiction Order issued on 13 May 1941. This order directed that "partisan suspects" be brought before an officer who would determine whether they were to be shot. This was considered during the Nuremberg Trials as "patently criminal" as it "permitted the immediate killing of alleged partisans and 'partisan suspects' without investigation and at the discretion of a junior officer".

56 ICRC Commentary on GC III, above note 3, Art. 5, para. 1126.

57 *Ibid.*, para. 1127.

58 *Ibid.*, paras 1119–1121.

59 *Ibid.*, para. 1115.

combatants in order to prevent them from returning to the battlefield.[60] The authorization to intern, contained in Article 21 of the Convention, gives expression to military necessity: interning PoWs for the duration of active hostilities aims to ensure that captured enemy personnel are not able to participate again in the hostilities, which would pose a military threat to the Detaining Power.[61]

Reflecting the requirements of humanity, on the other hand, GC III provides a set of general protections for PoWs, setting standards below which the treatment afforded to and conditions enjoyed by such prisoners must not fall. These overarching protections include the obligations of humane and equal treatment, the prohibition of adverse distinction, and respect for prisoners' persons and honour. GC III deals with an extremely broad range of issues, and many articles in the Convention are more specific iterations of these obligations. The drafters did not intend, however, to set out detailed rules or codes for every single area covered. Instead, the Convention refers in certain articles, through the principle of assimilation, to rules and regulations that are applicable to the Detaining Power's own armed forces. In those cases, PoWs are to be treated in accordance with these rules and regulations, while the Convention's standards on humane treatment continue to apply and act as a minimum standard.

The requirement to treat PoWs humanely is stated in Article 13 of GC III. It is complemented by the obligations in Article 14 to respect PoWs' person and honour, as well as the requirement to treat PoWs equally and the prohibition on discrimination in Article 16. These provisions provide for the minimum standard of treatment. They are interconnected and underpin all protections owed to PoWs.

The requirement to treat a PoW humanely (or in the equally authentic French version, "avec humanité") requires the respect of the prisoner's inherent human dignity and inviolable quality as a human being.[62] Article 13 provides certain express articulations of what this requires, including the prohibition of physical mutilation, medical or scientific experiments, acts of violence, intimidation, insults and public curiosity. The protection against public curiosity is particularly relevant in the age of mass media and social media, given the ease with which images and comments can be spread around the world.[63]

This obligation clearly cannot be separated from the obligation to respect a prisoner's person and their honour. Respect for the person of the PoW relates not only to the physical integrity of the prisoner, prohibiting acts of violence and physical torture, but also to their moral integrity – namely, the essential attributes that make up a person, including their religious, political, intellectual and social convictions, their gender and their sexual orientation.[64] Respect for the honour of

60 *Ibid.*, Introduction, para. 89.
61 *Ibid.*, Art. 21, para. 1932. Further expressions of military necessity can be found in the rules that serve the maintenance of security, discipline and good order in PoW camps. See, for example, GC III, Arts 42 (use of weapons against PoWs), 76 (censorship and examination), 92 (unsuccessful escape) and 95 (disciplinary procedures), which specifically mentions "camp order and discipline".
62 ICRC Commentary on GC III, above note 3, Art. 13, para. 1570.
63 *Ibid.*, para. 1563.
64 *Ibid.*, Art. 14, para. 1665.

a PoW more specifically entails due respect for the sense of value that every person has of themselves.[65] GC III expressly protects certain aspects of honour with regard to military structures, distinctions and codes of honour – for example, in providing that badges of rank and decorations may not be taken away from PoWs, and that they may not be deprived of their rank.[66] How the person and honour of the PoW is to be respected depends on a wide range of factors, including their cultural, social or religious background, their gender and their age.[67]

This, in turn, relates to the protection contained in Article 16 of GC III, which provides for the equality of treatment of PoWs and the prohibition of "adverse distinction based on race, nationality, religious belief or political opinions, or any other distinction founded on similar criteria".

Equal treatment does not necessarily require identical treatment. PoWs in different situations and with different needs may need to be treated differently in order to achieve substantive equality of treatment.[68] Article 16 expressly lists health, age and professional qualifications as potential grounds for "privileged treatment", and also requires consideration of provisions relating to rank and sex in GC III.[69] These considerations should not be taken as an exhaustive list upon which non-adverse distinction may be permitted or is required.[70]

In relation to discrimination, the prohibition in Article 16 identifies a number of grounds on which adverse differentiated treatment is prohibited: race, nationality, religious belief or political opinion, as well as "any other distinction founded on similar criteria". AP I provides a longer list of prohibited grounds: "race, colour, sex, language, religion or belief, political or other opinion, national or social origin, wealth, birth or other status, or … any other similar criteria".[71] Adverse distinctions founded on other grounds, such as ethnicity, disability, level of education or family connections of a PoW and, as noted above, age or state of health, would equally be prohibited. Any list of prohibited criteria will necessarily be incomplete and should be interpreted in light of legal and social developments. The residual category of "any other distinction based on similar criteria" makes express provision for this.

It is in conjunction with the minimum standards and safeguards provided in GC III that the principle of assimilation operates. This principle reflects an understanding that, with respect to certain issues, PoWs are to be treated in the same or a similar manner as members of the Detaining Power's own forces.[72] In

65 *Ibid.*, para. 1658.
66 GC III, Arts 18(3), 44.
67 ICRC Commentary on GC III, above note 3, Art. 14, para. 1659.
68 *Ibid.*, Art. 16, para. 1742.
69 With regard to different treatment in relation to sex, GC III, Art. 14(2) provides that female PoWs are to be treated "with all the regard due to their sex" and, most importantly, that their treatment may in no case be inferior to that of male PoWs.
70 See ICRC Commentary on GC III, above note 3, Art. 16, paras 1743–1744.
71 See GC III, Art. 9(1); 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) (AP I), Art. 75(1). See also ICRC Customary Law Study, above note 32, Rule 88.
72 See, in particular, GC III, Arts 20 (conditions of evacuation), 25 (quarters), 46 (conditions for transfer), 82 (applicable legislation), 84 (courts), 87 (penalties), 88 (execution of penalties), 95 (confinement awaiting

this way, it complements the prohibition on adverse distinction as it ensures that all PoWs interned by a Detaining Power are subject to the same or similar conditions and standards, irrespective of their country of origin. This would not necessarily be the case if the Detaining Power treated PoWs from different countries according to the standards and conditions prevailing in each of those different armed forces.

The principle of assimilation also facilitates the task of administering the internment of PoWs, since the Detaining Power has to apply to them some of the rules and standards that are already in force for its own troops. The Detaining Power is necessarily familiar with and has pre-existing experience with implementing those rules and standards and thus can readily apply them to PoWs as well.

The principle of assimilation does not operate in a vacuum; rather, it operates in conjunction with the minimum standards and safeguards spelled out in the rest of GC III, in particular those concerning the humane treatment of PoWs discussed above.[73] This is made explicit in several rules, including Article 82.[74] The approach to protecting PoWs by reference to the rules of both national and international law is also reflected in the provisions on penal and disciplinary sanctions (discussed below). Several of these provisions expressly make the principle of assimilation subject to compliance with minimum standards that must be applied to all PoWs, irrespective of the standards or conditions applicable to members of the armed forces of the Detaining Power. Accordingly, when the treatment afforded by a Detaining Power to its own armed forces falls short of the minimum standards set out in the Convention, the latter standards apply with respect to PoWs.

Timing of obligations

Planning and preparation

Because of the wide range of issues dealt with in GC III, proper planning and preparation, including making sure the domestic legal framework is up to date, are indispensable for its successful implementation.[75] An important part of this planning and preparation is the requirement for the Detaining Power to instruct the armed forces of their duties.[76] In this respect, Article 127(1) provides for the

hearing), 102 (conditions for validity of sentence), 103 (confinement awaiting proceedings), 106 (right to appeal) and 108 (premises and conditions for serving a sentence). The principle is also implicit in Articles 33 (rights and privileges of retained personnel), 52 (dangerous or humiliating labour) and 60 (advances of pay).

73 See *ibid.*, Art. 13, and the provisions that give expression to the requirement of humane treatment in specific areas, such as quarters (Art. 25), food (Art. 26), clothing (Art. 27) and hygiene (Art. 29).

74 "However, no proceedings or punishments contrary to the provisions of this Chapter shall be allowed." For further examples, see *ibid.*, Arts 25 (quarters), 46 (conditions of transfer), 50 (authorized work), 84 (courts), 87 (penalties), 95 (confinement awaiting hearing), 102 (conditions for validity of sentence), 103 (confinement awaiting trial) and 108 (execution of penalties).

75 This is consistent with the reference to provisions to be implemented in peacetime in *ibid.*, Art. 2(1).

76 This point was emphasized in the United Kingdom's Baha Mousa Public Inquiry Report. Although the Inquiry concerned the treatment of Iraqi civilian internees by UK armed forces, it contained general

dissemination of the text of the Convention in time of peace and in time of armed conflict "so that the principles thereof may become known to all their armed forces and to the entire population". Article 127(2) requires that authorities who assume responsibility for PoWs must possess the text of the Convention and be specially instructed in its provisions.

The implementation of some provisions of GC III requires action to be taken prior to the capture of prisoners. For example, the Convention requires that PoWs be interned on land, with every guarantee afforded for their hygiene and health; that they must not be held in penitentiaries except in particular cases where it is in the interests of the prisoners themselves;[77] and that they be quartered under conditions as favourable as those of the Detaining Power.[78] Providing accommodation that meets these standards requires infrastructure, equipment, logistics, trained staff, a budget and operating procedures. This may be challenging for the Detaining Power once it is engaged in an international armed conflict. Successfully establishing humane, compliant internment of PoWs requires States to develop plans, even in peacetime, with regard to how they would hold such potential prisoners, including the types and location of interment facilities.

On taking prisoners of war captive

Once a person in one of the categories of Article 4 falls into the power of the enemy, GC III applies as a whole. GC III does, however, take the different stages of captivity into consideration. For example, it contains a section dedicated to the beginning of captivity, outlining the obligations of the Detaining Power immediately after prisoners fall into its hands: Article 17 deals with the questioning of prisoners, Article 18 addresses the property of prisoners, and Articles 19 and 20 concern the evacuation of prisoners from the combat zone. While these articles are most relevant soon after combatants fall into the power of the enemy and during the initial processing of prisoners, they remain relevant beyond the immediate time and location of the point of capture, and in some cases, throughout captivity. For

conclusions and recommendations that are also relevant to PoWs. With regard to training, the Inquiry concluded that the general training the soldiers received in the law of armed conflict "lacked specific guidance on how to handle a prisoner; what the permitted treatment of a prisoner actually was in practical terms; and most importantly what type of treatment was expressly forbidden" (Vol. 2, para. 6.67). In addition, the Inquiry identified deficiencies in specific teaching courses, including the training given to tactical questioners and interrogators. Accordingly, it made several recommendations, both general (Recommendations 47–58) and specific (Recommendations 59–73), on training soldiers in the handling of prisoners. See Sir William Gage, *The Report of the Baha Mousa Inquiry*, HM Stationery Office, London, September 2011, Vol. 2, paras 6.66–6.73, 6.339–6.349, and Vol. 3, pp. 1279–1282, 1282–1286, available at: www.gov.uk/government/publications/the-baha-mousa-public-inquiry-report. The Al Sweady Public Inquiry Report referred to several of these recommendations; see Sir Thayne Forbes, *The Report of the Al Sweady Inquiry*, Vol. 2, HM Stationery Office, London, December 2014, para. 5.101, available at: www.gov.uk/government/publications/al-sweady-inquiry-report.

77 GC III, Art. 22(1).

78 *Ibid.*, Art. 25. This is discussed further below.

example, the prohibition on torture and coercion during questioning set out in Article 17 remains valid during the entire time of internment.[79]

The principle of humane treatment discussed above underpins these articles as they seek to ensure that, where prisoners are taken, they are brought to safety and are properly identified and processed. Often the first obligation for a Detaining Power is to evacuate the persons who have fallen into its power to an area that is far enough removed from the combat zone for the prisoners to be out of danger.[80] This evacuation must be carried out humanely and in conditions similar to those for the forces of the Detaining Power when they change positions.[81]

Depending on the circumstances, such as the distance and available means of transport, it may be that PoWs pass through transit camps during their evacuation. Such camps may be established temporarily and even close to the combat zone. Considering these *ad hoc* circumstances, it will usually be difficult for a Detaining Power to fulfil all the material conditions of the entire Convention. Accordingly, the stay in such camps must be as brief as possible.[82] These camps can be distinguished from permanent transit camps. If a Detaining Power has such permanent establishments which it uses to screen and process prisoners, they must offer conditions similar to those of other PoW camps, and prisoners therein must benefit from the same treatment as in other camps.[83]

After their evacuation and processing, PoWs typically arrive in a permanent PoW camp. However, they do not always stay in one place, nor under the responsibility of the same Power. During their captivity, they may be transferred to other camps and/or to other Powers. GC III regulates both the physical transfer of PoWs to another location, irrespective of whether they remain under the control of the same Power, and the transfer of PoWs from one Power to another.[84] For the transfer of a prisoner to another location, the Convention has a similar provision as for evacuation: the transfer must be effected in a humane manner and in conditions not less favourable than those under which the Detaining Power's own forces are transferred. This provision is slightly more stringent, however, than the provision on evacuation, as the conditions in the former case must only be "similar". This is understandable considering the more predictable nature of a transfer compared to an evacuation from the battlefield.[85]

During captivity, a prisoner may also be transferred to another Power if that receiving Power is also a party to GC III[86] and after the original Detaining Power can satisfy itself of the willingness and ability of the receiving Power to

79 See ICRC Commentary on GC III, above note 3, Art. 17, para. 1822.
80 GC III, Art. 19(1).
81 *Ibid.*, Art. 20(1).
82 *Ibid.*, Art. 20(3).
83 ICRC Commentary on GC III, above note 3, Art. 24, paras 2058, 2063–2065.
84 GC III, Arts 12(2)–(3), 46–48.
85 Keiichiro Okimoto, "Evacuation and Transfer of Prisoners of War", in Andrew Clapham, Paola Gaeta and Marco Sassòli (eds), *The 1949 Geneva Conventions: A Commentary*, Oxford University Press, Oxford, 2015, p. 965, quoted in ICRC Commentary on GC III, above note 3, Art. 46, n. 16.
86 This does not serve as a limitation today, as the Geneva Conventions are universally ratified.

apply the Convention.[87] Because of the general understanding that only States can be High Contracting Parties to the Geneva Conventions, this means that PoWs may not be transferred to entities other than States, such as non-State armed groups and paramilitary and non-military organizations.[88]

An important obligation due to the transferring Power is that if the receiving Power "fails to carry out the provisions of the Convention in any important respect", it must "take effective measures to correct the situation or shall request the return of the prisoners of war".[89] The Convention does not explain what "important respect" means. One benchmark for determining whether a breach is "important" is whether it violates the general obligation of humane treatment as articulated in Article 13. This covers acts that qualify as grave breaches. Failure to provide for the basic needs of prisoners with respect to their quarters, food, water and medical care, in a way that would endanger the health of the prisoners, or denying prisoners contact with the outside world, including visits from the ICRC, would also be covered. These examples are not exhaustive.[90]

There are different ways for a transferring Power to rectify such a failure to comply. As the Convention itself specifies, in some situations, the transferring Power must request the return of the prisoners. Where the failure is due to inadequate material conditions of internment, such as lack of space, food, water or medical care, measures to correct the situation may consist of direct assistance provided by the transferring Power, such as food, medical staff and equipment. In situations where the failure is more systemic, for example when it relates to a denial of judicial guarantees or ill-treatment by camp staff, a request for the return of the prisoner may be the only adequate measure.[91]

On the end of captivity

GC III also regulates the end of captivity of PoWs. For most PoWs, captivity will cease at the end of active hostilities. Article 118 establishes a unilateral and non-reciprocal obligation on Detaining Powers to release and repatriate PoWs without delay after the cessation of active hostilities.[92] This obligation logically follows from the purpose of internment, which is to prevent further participation in hostilities.[93] Once hostilities between the two or more States have ended, there is no longer a need to keep PoWs interned.[94] Release and repatriation at the end of

87 GC III, Art. 12(2). This includes neutral States. See ICRC Commentary on GC III, above note 3, Art. 12, section C.2.a.

88 See, however, ICRC Commentary on GC III, above note 3, Art. 12, for a discussion on transfers to non-State entities, including armed groups under the overall control of a State, international organizations or international courts and tribunals, at paras 1530–1532.

89 GC III, Art. 12(3).

90 For a discussion and examples, see ICRC Commentary on GC III, above note 3, Art. 12, section E.2.

91 For further discussion, see *ibid.*, section E.4.

92 PoWs against whom criminal proceedings are pending or who are serving a criminal sentence may be kept back: see GC III, Art. 119(5).

93 See ICRC Commentary on GC III, above note 3, Art. 21, section C.1., and in particular para. 1932.

94 *Ibid.*, Art. 118, para. 4444.

active hostilities marks the end of application of GC III for these prisoners. Repatriation at the end of hostilities must take place "without delay". While this implies that repatriation does not have to be instantaneous, it must happen as soon as feasible considering the circumstances. This may depend, for example, on the number of persons to be repatriated, the security situation, the location of the camp(s) and available logistical means, and the ability of the State on which the prisoners depend to receive the prisoners.[95]

GC III does not address the situation in which a PoW does not want to be repatriated. As already recognized in the 1960 Pictet Commentary, and reiterated in the updated Commentary, an exception to the obligation to repatriate PoWs may be made if, as determined on an individual case-by-case-basis, there are

> serious reasons for fearing that a prisoner of war who is himself opposed to being repatriated may, after his repatriation, be the subject of unjust measures affecting his life or liberty, especially on grounds of race, social class, religion or political views, and that consequently repatriation would be contrary to the general principles of international law for the protection of the human being.[96]

An interpretation of Article 118 allowing for such an exception is in line with the principle of *non-refoulement* under international law, by which a State cannot transfer persons within its control to another State if there is a real risk that they may face violations of certain fundamental rights.[97]

The updated Commentary on Article 118 also discusses the obligation to release and repatriate in situations where the legal classification of a conflict changes from an international to a non-international armed conflict, because of a change of circumstances on the ground. In such circumstances, a party to the conflict is unlikely to be willing to release and repatriate any PoWs that it holds. This is an example of a situation where the updated Commentary indicates divergent views and highlights issues not yet settled. There are two main approaches to this issue. Under the first approach, the obligation to release and repatriate PoWs is not triggered because the hostilities between the same actors have not ceased, even if the legal classification of the armed conflict has changed. Accordingly, GC III remains the legal basis for the internment of PoWs and for their protection. Under the second approach, the hostilities related to the international armed conflict and the non-international armed conflict are

95 *Ibid.*, para. 4462.

96 Jean Pictet (ed.), *Commentary on the Geneva Conventions of 12 August 1949*, Vol. 3: *Geneva Convention relative to the Treatment of Prisoners of War*, ICRC, Geneva, 1960, p. 547; ICRC Commentary on GC III, above note 3, Art. 118, para. 4469.

97 See Cordula Droege, "Transfers of Detainees: Legal Framework, *Non-refoulement* and Contemporary Challenges", *International Review of the Red Cross*, Vol. 90, No. 871, 2008, p. 671; Emmanuela-Chiara Gillard, "There's No Place Like Home: States' Obligations in Relation to Transfers of Persons", *International Review of the Red Cross*, Vol. 90, No. 871, 2008, p. 704; Christopher Michaelsen, "The Renaissance of Non-refoulement? The Othman (Abu Qatada) Decision of the European Court of Human Rights", *International and Comparative Law Quarterly*, Vol. 61, No. 3, 2012, p. 753. See also ICRC Commentary on GC III, above note 3, Art. 3, section G.7.

considered to be distinct. Because the hostilities related to the international armed conflict have ceased, the obligation to release and repatriate PoWs is triggered on the basis of Article 118 of GC III. In that case, the latter no longer provides a legal basis for the internment of the prisoners, and if the detaining party believes it must continue to hold such persons for imperative reasons of security, another legal basis for their internment is required.[98]

In addition to the obligation to release and repatriate PoWs at the end of active hostilities, certain PoWs must be released and repatriated earlier than this. GC III dedicates a number of articles to the repatriation of seriously wounded or sick PoWs during hostilities.[99] Again, this is a logical consequence of the purpose of internment. The assumption is that such prisoners are no longer able to participate in hostilities and therefore their continued internment would no longer be justified by military necessity.[100] A safeguard is built into the Convention though, as it includes an explicit prohibition against re-employing such repatriated prisoners on active military service; this is particularly important in modern warfare, given the wider variety of assignments that might make the redeployment of seriously wounded or sick prisoners possible.[101]

Finally, GC III also contains rules applicable to the Detaining Power when a prisoner dies during captivity. Needless to say, full compliance with GC III may reduce instances of fatalities, both through proper care of prisoners and through ensuring the repatriation of the seriously wounded and sick. In the event that prisoners do pass away during internment, the Detaining Power retains certain obligations towards the deceased, which indirectly benefit their family. First, as an important means of accountability and to prevent people going missing, death certificates or certified lists must be prepared for any person who dies while a PoW. These documents should record the identity of the dead, the circumstances of death, and the burial site (or details of cremation, if applicable).[102] They must be forwarded to the national information bureau as rapidly as possible, which today, generally means electronically.[103] At the same time, if not done previously, the will of the deceased should also be transmitted to the Protecting Power, and a certified copy sent to the Central Tracing Agency.[104] These processes are not only important for families' "closure" but may also have important legal implications.

Respecting the honour of a prisoner extends to the dead: Detaining Powers are required to ensure that PoWs who have died in captivity are honourably buried, if possible according to the rites of their religion, and that their graves are respected and suitably maintained and marked.[105] AP I goes further, requiring parties to

98 For a detailed discussion, see ICRC Commentary on GC III, above note 3, Art. 118, paras 4459–4460, and Art. 5, section C.4.
99 GC III, Art. 110.
100 ICRC Commentary on GC III, above note 3, Art. 109, para. 4245.
101 On the temporal scope of the obligation, see *ibid.*, Art. 117, section C.3.
102 Details should be included as to why cremation was chosen (e.g. religious reasons, the wishes of the deceased), given the presumption in GC III in favour of burial. See *ibid.*, Art. 120, para. 4576.
103 *Ibid.*, Art. 120, para. 4563.
104 GC III, Art. 120(1).
105 This includes the establishment of an official grave registration service: *ibid.*, Art. 120(6).

conclude agreements as soon as circumstances permit "to facilitate the return of the remains of the deceased and of personal effects to the home country".[106] The ICRC can and has acted as a neutral intermediary in the return of bodies to the families of the deceased.[107]

Substantive protections

As discussed above, in addition to setting out fundamental principles for the protection of PoWs, GC III elucidates express protections on many facets of the life of a PoW. The following discussion summarizes a number of these protections.

Internment in a PoW camp

In the event that a PoW is interned, he or she should not be held in a penitentiary unless it is in his or her best interests.[108] Further, unless they are subjected to penal or disciplinary sanctions, which are further discussed below, or when necessary to safeguard their health, PoWs may not be held in close confinement.[109]

While it is not an obligation for the Detaining Power to intern PoWs, generally it will choose to do so, and GC III provides detailed conditions for such internment. Below is a summary of some of the provisions provided for interned PoWs.

Quarters

Article 25 provides that PoWs who are interned must be "quartered under conditions as favourable as those of the forces of the Detaining Power who are billeted in the same area". Again, this provision is underpinned not only by the fundamental protections described above (including respecting the person and the honour of PoWs) but also by the consideration that holding PoWs is not intended to be for punitive reasons. Furthermore, allowance must also be made for the customs and habits of PoWs, and the Detaining Power must ensure that the accommodation is not "prejudicial to their health".[110]

While there can be wide disparities between the standards of quarters provided by the Detaining Power to its own forces, quarters provided for PoWs must be at least of the standard genuinely used by the Detaining Power to accommodate a significant number of those forces.[111] The quarters must also be protected from the vagaries of the weather and vermin and should be periodically

106 AP I, Art. 34(2)(c). See also ICRC Customary Law Study, above note 32, Rule 114.
107 ICRC Commentary on GC III, above note 3, Art. 120, para. 4598.
108 GC III, Art. 22(1).
109 *Ibid.*, Art. 21(1).
110 *Ibid.*, Art. 25(1).
111 ICRC Commentary on GC III, above note 3, Art. 21, para. 2076.

visited by a doctor or other suitably qualified person to ensure that they are not prejudicial to the health of the prisoners.[112]

According to Article 25, women must be provided with dormitories separate from men, but it is not necessarily required that the quarters as a whole be separated.[113] In the event that infants or very young children are present in PoW camps (for instance, because they were born there), they must be accommodated with their parents.[114]

Food

Article 26 requires the Detaining Power to allow for basic daily food rations that are "sufficient in quantity, quality and variety", as well as sufficient drinking water. Care must be taken of prisoners with health conditions by appropriately adapting food rations to their condition. Rations provided for older PoWs, pregnant or lactating prisoners, or any children present in PoW camps have to be adapted to their needs.[115] Where PoWs carry out physical work, they will also need to be provided with additional rations to permit them to remain in good health.[116]

Article 26 further requires the Detaining Power to take into account the habitual diet of the prisoners.[117] One means to implement this provision is to involve them in the preparation of their own meals.[118]

GC III requires that "[t]he use of tobacco shall be permitted." At the time of drafting the Convention, the health hazards of tobacco use were not commonly known. Today, it would be appropriate and consistent with the requirement to provide for a healthy environment for internees for a Detaining Power to impose reasonable restrictions on tobacco use, such as measures to protect people against passive smoking and to prevent minors from gaining access to tobacco.[119] This may also be required by other applicable rules of international law.[120]

Parties to an armed conflict shall also provide canteens where "foodstuffs, soap and tobacco and ordinary articles in daily use" must be available for purchase.[121] However, in certain situations, for example in conflicts of short duration or where PoWs are to be transferred to another camp or to another

112 *Ibid.*, Art. 25, paras 2078–2079.
113 In comparison, see GC III, Art. 108(2), which requires women PoWs undergoing confinement to be held in separate quarters. See also AP I, Art. 75(5), and ICRC Customary Law Study, above note 32, Rule 119, which refer to separate quarters for women.
114 See also AP I, Arts 75(5), 77(4), and also ICRC Commentary on GC III, above note 3, Art. 25, para. 2104.
115 ICRC Commentary on GC III, above note 3, Art. 26, para. 2113.
116 GC III, Art. 26(2); ICRC Commentary on GC III, above note 3, Art. 26, para. 2126.
117 See, for example, Canada, *Prisoner of War Handling Manual*, 2004, p. 3F-10: "Ration scales are to be tailored, as far as is possible, to the national dietary requirements of [PoWs], bearing in mind that a diet which is totally suited to [PoWs] from one nation may be inadequate or unsuitable for those from a different nation. There may also be religious or ethnic dietary requirements for which, whenever possible, provision should be made." See also ICRC Commentary on GC III, above note 3, Art. 26, para. 2121.
118 GC III, Art. 26(4).
119 ICRC Commentary on GC III, above note 3, Art. 26, para. 2131.
120 See, for example, the WHO Framework Convention on Tobacco Control, 2003.
121 GC III, Art. 28.

party to the conflict, it may be unnecessary or unreasonable to establish such a canteen.[122]

Clothing

Article 27 of GC III requires the Detaining Power to supply clothing, underwear and footwear to PoWs. To ensure that the health of captives is not affected, the Detaining Power must provide sufficient clothing adapted to the climate where the prisoners are interned, such as sweaters, hats and gloves in cold climates.[123] PoWs generally require at least two sets of clothing and sleepwear to enable a change when one set is being washed or repaired.[124]

The type of clothing provided must also be in line with the fundamental protections described above, in particular the obligation to respect the person's honour. Clothing must be adapted, for example, to the prisoner's age, gender, and religious and cultural background.[125] PoWs may not be compelled to wear the uniform of their enemies or other clothing that may negatively impact their sense of allegiance or honour.[126]

Medical care and sanitation

Every PoW camp must have its own infirmary to tend to the health-care needs of prisoners. PoWs requiring medical attention are entitled to receive it at the cost of the Detaining Power.[127] Meeting the health-care needs of prisoners may require, in some circumstances, transferring prisoners with health conditions that require specialized treatment to a military or civilian medical unit where such treatment can be given.[128] All medical care must comply with the applicable standards of medical ethics, which include the duty to provide medical care impartially and without adverse distinction. Such standards also address the principle of voluntary and informed consent.[129]

GC III refers to the use of isolation wards for "contagious or mental disease". Any decision to use isolation wards must be taken exclusively on the advice of a medical doctor or other appropriately qualified health professional, and should only be for as long as necessary. The reference in Article 30 to isolating people with mental health conditions "if necessary" should be read in line with the other obligations of the Detaining Power, including the fundamental protections mentioned above.[130] Imposing isolation on PoWs with mental health conditions should be avoided – it may aggravate the person's condition, may be

122 ICRC Commentary on GC III, above note 3, Art. 28, para. 2164.
123 *Ibid.*, Art. 27, para. 2149.
124 *Ibid.*, para. 2148.
125 *Ibid.*, para. 2151.
126 *Ibid.*, para. 2151.
127 GC III, Art. 15.
128 *Ibid.*, Art. 30(2).
129 ICRC Commentary on GC III, above note 3, Art. 30, paras 2232, 2245.
130 *Ibid.*, para. 2242.

inconsistent with the prohibition on adverse distinction, and may amount to torture or other ill-treatment as it can lead to psychotic symptoms and/or significant functional impairments, self-harm or even suicide.[131]

In order to prevent illness, GC III also provides an obligation on Detaining Powers to take all necessary sanitary measures to ensure cleanliness and "healthfulness" of camps and to prevent the spread of infectious diseases.[132] Upholding hygienic standards and reducing the risk of disease transmission within places of detention is of immediate practical value to the Detaining Power, as it reduces the risk of transmission to personnel of the Detaining Power, such as guards, as well as the neighbouring community.[133]

Recreation and religion

Maintaining the health of a PoW and ensuring respect for their person requires attention not only to the physical well-being but also to the mental well-being of the prisoner. One of the ways in which this is acknowledged and addressed in GC III is through the requirement of allowing prisoners "complete latitude" to exercise their religious duties (or in the equally authentic French, "l'exercise de leur religion"), provided this complies with any disciplinary routine prescribed by the military authorities. GC III also requires the Detaining Power to encourage "the practice of intellectual, educational, and recreational pursuits, sports and games amongst prisoners".[134]

Allowing prisoners to practice their faith is an important way through which the Detaining Power can enable PoWs to process their current situation and the hardships that come with it.[135] It is also consistent with the obligations both to treat PoWs humanely and to respect their person and honour. Detaining Powers must take religious practices into account in many aspects of camp life, such as setting up the place of internment (for example, providing facilities for washing), food preparation (consistent with religious precepts and taboos) and work schedules (for example, allowing time for prayer).[136] Various armed forces employ cultural advisers to help them better understand the human and cultural environments in which they operate.[137]

In relation to recreational activities, Article 38 specifies that the individual preferences of each prisoner must be respected to ensure that the provision is not used as a pretext to oblige prisoners to take place in ideological or political propaganda under the guise of "recreation".[138] The Detaining Power must

131 *Ibid.*, para. 2243.
132 GC III, Art. 29(1).
133 ICRC Commentary on GC III, above note 3, Art. 29, para. 2185.
134 GC III, Arts 34, 38.
135 ICRC Commentary on GC III, above note 3, Art. 34, para. 2359.
136 *Ibid.*, Art. 34, para. 2365.
137 *Ibid.*, para. 2366.
138 See also GC III, Art. 16; ICRC Commentary on GC III, above note 3, Art. 14, para. 1671.

provide prisoners with adequate premises and necessary equipment for this purpose, including sufficient open spaces for physical exercise.

Educational opportunities are particularly important for PoWs who are interned for long periods of time. In some international armed conflicts, the ICRC has been allowed to supply writing materials, notebooks, textbooks and other books, subject to the Detaining Power's approval, as well as sporting equipment.[139]

Relations with the exterior

Maintaining connection with the outside world is another vital means of maintaining morale for PoWs, as well as serving as a check on their treatment and preventing disappearances.

Article 70 of GC III provides that a PoW's capture, sickness, hospitalization and transfer should be communicated at the earliest possible moment to the prisoner's family and also to the Central Tracing Agency (previously known as the Central Prisoners of War Agency). This is facilitated by enabling PoWs to write "capture cards" which are forwarded as rapidly as possible to the Central Tracing Agency and to the family of the prisoner.

For more substantive communications, Article 71 provides for the right of PoWs to send and receive letters and cards. The importance of being connected to families was well understood at the time of drafting GC III. As expressed by the ICRC shortly after the adoption of the Geneva Conventions, "[e]ven the most favourable living conditions do not compensate, in the eyes of the prisoner, for absence of news or slowness in mail delivery".[140] In practice, where postal services are not functioning, the ICRC regularly facilitates correspondence through its "Red Cross messages" service, enabling families to connect and share content of a strictly private and familial nature.[141]

Article 71 also recognizes that in circumstances where PoWs have been without news for a long period of time or are unable to receive news from their next of kin or by the ordinary postal route, they "shall be permitted to send telegrams". Clearly this is a product of the time of drafting, but the purpose behind this provision should be respected with the use of more modern means of communication, such as email, telephone calls or video calls.[142]

Another additional protection provided for PoWs is that they may receive relief shipments. Detaining Powers cannot charge "import, customs or other dues" on such shipments, or "postal dues".[143]

139 See ICRC Commentary on GC III, above note 3, Art. 34, para. 2377, and Art. 38, para. 2461.
140 ICRC, *The Geneva Conventions of August 12, 1949: Analysis for the Use of National Red Cross Societies*, Vol. 2, Geneva, 1950, p. 27.
141 For more information, see ICRC Commentary on GC III, above note 3, Art. 71, para. 3215.
142 *Ibid.*, para. 3218.
143 GC III, Art. 74.

The use of PoWs' labour

The ability to use a PoW's labour is of potential benefit to the Detaining Power. The framework provided for PoW labour also assists in supporting the well-being of prisoners, maintaining them in a good state of physical and mental health. The absence of meaningful activity, coupled with isolation and uncertainty about the future, can lead to boredom and impact prisoners' mental and physical well-being.[144]

PoWs may only be engaged in certain types of work and may not be engaged in work that is unhealthy or dangerous unless they volunteer. In addition, for permitted labour, certain health and safety measures are expressly prescribed, such as the requirement of suitable accommodation, food, clothing and equipment for the tasks in which the prisoners are employed; these may not be inferior to those enjoyed by nationals of the Detaining Power employed in similar work. In relation to the duration of labour, three essential safeguards are put in place: the duration of labour must not be excessive,[145] the maximum duration of work is fixed as the maximum allowed under the domestic legislation of the Detaining Power for civilians in the same work, and the time taken to travel to and from the place of work must be counted within the working hours. Detaining Powers must allow for a minimum of one hour's rest in the middle of the day, a day of rest per week, and a period of eight consecutive rest days every year.[146]

Article 62 provides that PoWs shall be paid "a fair working rate of pay by the detaining authorities direct", and that this rate shall be fixed by the authorities "but shall at no time be less than one-fourth of one Swiss franc for a full working day".[147] Even factoring in the Swiss consumer price index, 0.25 Swiss francs in 1949 corresponded to just 1.25 Swiss francs in 2019.[148] In many contexts around the globe, this amount would not be considered a fair working rate, and accordingly the Detaining Power must consider in good faith an adequate increase.[149]

Relations with the detaining authorities

GC III contains three categories of provisions regarding the relationship between PoWs and the detaining authorities. These cover circumstances where the prisoners have complaints about their conditions of captivity; the mechanism for facilitating communication between prisoners and the detaining authorities

144 ICRC Commentary on GC III, above note 3, Art. 49, para. 2675.
145 *Ibid.*, Art. 53, para. 2762.
146 GC III, Art. 53.
147 ICRC Commentary on GC III, above note 3, Art. 62, para. 2952.
148 See "Indice suisse des prix à la consommation: La calculatrice du renchérissement", available at: www.portal-stat.admin.ch/lik_rechner/f/lik_rechner.htm.
149 For further discussion on the fixing of a fair rate, see ICRC Commentary on GC III, above note 3, Art. 62, paras 2952–2955.

(namely, through PoW representatives); and circumstances where the detaining authorities have complaints about the conduct of detainees (penal and disciplinary sanctions).

PoWs have a right to "make known" their requests about the conditions of their captivity to the prison authorities, as well as to the prisoners' representative or even directly to the Protecting Powers. They cannot be punished for making these requests. In practice, complaints are often communicated to the ICRC, through channels including confidential interviews with ICRC delegates pursuant to Article 126. The role of the ICRC in this regard is important given the absence of Protecting Powers in most international armed conflicts since 1949.[150]

Prisoners' representatives are PoWs who are elected by the other prisoners and are tasked with representing prisoners before military authorities, Protecting Powers, the ICRC and other organizations. They work for the well-being of PoWs, and they carry out a number of other duties defined in GC III.[151] Prisoners' representatives are to be supported in this role by the Detaining Powers, by having the right to "all material facilities",[152] the ability to appoint advisers or assistants,[153] an exemption from work if it makes fulfilling their duties difficult,[154] and the freedom to move about the camp or visit other locations in order to fulfil their duties.[155]

In some circumstances, there may be cause for a Detaining Power to pursue disciplinary or judicial proceedings against a PoW. Underlying the framework for disciplinary and judicial proceedings is the principle of assimilation, according to which PoWs are "subject to the laws, regulations and orders in force in relation to the armed forces of the Detaining Power".[156]

In deciding whether to proceed with disciplinary or judicial proceedings, Detaining Powers are required to apply "the greatest leniency", recognizing that prisoners owe no allegiance to the Detaining Power.[157] The only four types of disciplinary punishments allowed for are fines, discontinuance of privileges, fatigue duties and confinement.[158] More arduous labour may not be imposed as a disciplinary punishment.[159]

If a PoW is to face prosecution for an offence, they can only be tried in a court that offers the essential guarantees of independence and impartiality, and in particular the procedure of which affords the accused the necessary rights and

150 *Ibid.*, Art. 78, para. 3433. On the absence of Protecting Powers in general, see also *ibid.*, Introduction, paras 49–51.
151 ICRC Commentary on GC III, above note 3, Art. 80, section D.
152 *Ibid.*, Art. 80, para. 3528.
153 See *ibid.*, Art. 80, para. 3525, in relation to the different usage of the terms "adviser" and "assistant".
154 Prisoners' representatives and their assistants/advisers are paid out of canteen funds, unless there are no such funds available, in which case they are paid by the detaining authorities: *ibid.*, Art. 62, para. 2944.
155 GC III, Art. 81(2).
156 GC III, Art. 82(1). For a detailed discussion on the principle of assimilation in relation to disciplinary or judicial proceedings, see ICRC Commentary on GC III, above note 3, Art. 82, section C. For a wider discussion of the principle of assimilation in GC III, see *ibid.*, Introduction, section A.3.c.
157 GC III, Art. 83.
158 *Ibid.*, Art. 89.
159 ICRC Commentary on GC III, above note 3, Art. 51, paras 2737–2738.

means of defence.[160] While the principle of assimilation will ordinarily ensure the application of a robust framework for judicial guarantees and due process, GC III expressly sets out a number of protections, including the prohibition against double jeopardy, the principle of legality (the prisoner may not be tried or sentenced for an act which was not prohibited by the law of the Detaining Power or by international law at the time the act was committed) and the right to present one's own defence with the assistance of a qualified advocate or counsel.[161] In the event that the standards provided for in the domestic law of the Detaining Power fall short of these minimum standards, the rules of the Convention prevail and PoWs must benefit from the protections that it offers.

Conclusion

GC III provides a robust framework for the protection of PoWs, whereby prisoners must be treated humanely, their person and honour is to be respected, they must be treated equally, and discrimination is prohibited.

The articulation of these principles within GC III is detailed. Learning from the experiences of previous conflicts, in particular World War II, the drafters recognized the essential applications of these principles that are needed to ensure the humane treatment of PoWs and respect for their person and honour. The drafters understood from experience what was essential to keep PoWs in good mental and physical health. They also understood the realities of providing for the care of prisoners while in the midst of active hostilities. The 142 articles of GC III provide a rich framework of realistic but essential protections covering all aspects of a prisoner's capture until their final release and repatriation. Some articles refer to outdated technologies or understandings of science, but after many decades of visiting PoWs, the ICRC is firmly convinced that the provisions remain as relevant and important for prisoners today as they were when first drafted.

The updated ICRC Commentary on GC III is the third of the series of updated Commentaries to be published by the ICRC. Research continues with respect to the protections of civilians in times of war (GC IV), and an updated Commentary will be published on this and on Additional Protocols I and II consecutively over the coming years.

160 GC III, Art. 84(2). See also Art. 105.
161 *Ibid.*, Arts 86, 99.

International Review of the Red Cross (2020), **102** (913), 417–443.
Digital technologies and war
doi:10.1017/S1816383121000072

IRRC

The camera and the Red Cross: "Lamentable pictures" and conflict photography bring into focus an international movement, 1855–1865

Sonya de Laat*
Dr Sonya de Laat is a member of the Humanitarian Health Ethics research group, an Academic Adviser in the Global Health graduate programme at McMaster University, Canada, and a member of the Canadian Network on Humanitarian History.

Abstract

Henry Dunant's appeal for a neutral and impartial organization to provide care to wounded combatants aligned with growing criticism of mid-nineteenth-century European and North American conflicts. This article discusses the important convergence of Dunant's "lamentable pictures", laid out in his Memory of

* The author would like to thank her doctoral thesis committee: Sharon Sliwinski (Western University), Amanda Grzyb (Western University) and Lisa Schwartz (McMaster University). She would also like to thank Valérie Gorin (CERAH), Dominique Marshall (Carleton University) and members of the Canadian Network on Humanitarian History for comments on early drafts of this paper. Her participation in the Global Humanitarian Research Academy, along with her SSHRC and OGS doctoral awards, were invaluable in supporting research for this paper.

Solferino, *with spectators' passionate responses to them and to battlefield photographs that circulated between 1855 and 1865. Through these images and reactions, there emerged a shared, expanded vision of humanity worth caring for, which brought into focus an international humanitarian movement.*

Keywords: Red Cross history, combat photography, humanitarian visual culture, Henry Dunant, Solferino.

Introduction

The mid-nineteenth century was a time that saw an emergence of new media, leading historian and philosopher Richard Rorty to claim it as having contributed to an "unprecedented acceleration in the rate of moral progress".[1] New media, including the telegraph, the illustrated newspaper and the camera, circulated – at unprecedented speed and in unsurpassed detail – news and pictures of bloody conflicts in Europe and North America. While not everyone was reacting the same way to this information, the novel perspective offered by the camera began to alter people's perceptions. Together with a rising tide of democratic nationalism came a growing distaste for warfare and a recognition of the common soldier not as the "scum of the earth" – as they had been derided by some contemporary military leaders – but as a fellow citizen worthy of protection and care.[2]

Henry Dunant's appeal for an organization supported by an international convention to provide care to wounded combatants, regardless of rank or nationality, aligned with that of spectators who were responding in particular ways to novel conflict photographs, making it possible for Dunant's ideas to become a reality. As an example of the way in which history can provide perspective on the present, this paper includes a brief narrative of Dunant, as an unsuspecting tourist of the horrors of battle, and his translation of the suffering he witnessed into "lamentable pictures" meant to invoke passionate responses in his readers.[3] His book is regarded from within the context of the nineteenth-century equivalent of the "viral" spread of news and pictures of three major conflicts of the time: the Crimean Campaign, the Battle of Solferino and the American Civil War. These conflicts, and the unprecedented spread and detail of stories about them, are said to have contributed to an acceleration in humanitarian sentiment.

1 Richard Rorty, "Human Rights, Rationality and Sentimentality", in *Truth and Progress: Philosophical Papers*, Cambridge University Press, Cambridge, 1998, p. 121.

2 The Duke of Wellington made the following statement on 2 July 1813: "It is quite impossible for me or any other man to command a British Army under the existing system. We have in the service the scum of the earth as common soldiers." This reflected a common sentiment, particularly among the gentry. See Jonathan Marwil, "Photography at War", *History Today*, Vol. 50, No. 6, 2000, p. 35.

3 Following Darnton, I support close readings of the past to gain insights into the present. See Robert Darnton, *George Washington's False Teeth: An Unconventional Guide to the Eighteenth Century*, Norton, New York, 2003.

Paying particular attention to eyewitness accounts of mid-nineteenth-century battles and to recent visual scholarship on spectators' responses to images of atrocity and suffering, what follows is an invitation to gain a new perspective on Dunant's words through the lens of early combat photography and the humanitarian narrative. As humanitarian actors and agencies today actively work to assess the impact of their communication strategies on global audiences, this historical review provides insights on the role of photography in expanding a shared vision of who constitutes humanity, and who is worth caring for. As some of the examples in this article show, there are no guarantees when it comes to use of, and engagement with, photographs. Of interest here, however, are people's passionate responses to pictures, for it is this relationship that plays a vital role in developing humanitarian sentiment. In the case of Dunant's "lamentable pictures", passionate responses helped bring into focus an international humanitarian movement.

Translating suffering

Henry Dunant wrote these words near the end of his 1862 book, *A Memory of Solferino*:

> But why have I told of all these scenes of pain and distress, and perhaps arouse painful emotions in my readers? Why have I lingered with seeming complacency over lamentable pictures, tracing their details with what may appear desperate fidelity?[4]

The passage is a bridge to Dunant's call for a relief organization that would provide care to soldiers wounded in war. The quotation suggests that the text is filled with "lamentable pictures", yet, remarkably, the only figure illustrating the book's first edition was a line-drawn map.[5] Pictures in the physical sense of the word (photographs, prints or drawings) may not have appeared, but the book's pages are indeed filled with images in the notional sense (mental pictures, imagination).[6]

4 Henry Dunant, *A Memory of Solferino*, International Committee of the Red Cross (ICRC), Geneva, 1862.

5 Dunant would create three editions of his book within the first year: one for close acquaintances, followed by one for heads of State and political officials, and a third popular edition. Martin Grumpert, *Dunant: The Story of the Red Cross*, Oxford University Press, New York, 1938, p. 84.

6 Dunant was exposed to photography through Jean-Gabriel Eynard (1775–1863), a family friend, wealthy banker and early photography enthusiast. While photography was growing in popular use, reproduction technology at the time meant that photographs had to be translated into (woodblock) prints or hand-printed for each volume, which was prohibitively expensive. Dunant primarily considered himself a man of letters, making it natural that he would have gravitated to the written word to express himself. He did recognize the rhetorical force of visual pictures, using woodblock prints in his first pamphlets promoting his ideas for an aid organization. He would later coordinate the composite group portrait of the participants of the First Geneva Convention. See Roger Durand, *Henry Dunant, 1828–1910*, Slatkine, Geneva, 2011, p. 42; Roger Durand, personal communication, July 2015; Natalie Klein-Kelly, "Dot to Dot: Exploring Humanitarian Activities in the Early Nineteenth Century", *Human Rights and Humanitarianism Blog*, 13 October 2017, available at: https://hhr.hypotheses.org/1766#more-1766 (all internet references were accessed in January 2021).

Beginning already in the eighteenth century, "spectatorial sympathy" in poetry, theatre and the novel was a popular device marshalled to open eyes and soften hearts.[7] Dunant's choice was a style of writing with which he had become intimately familiar through an influential member of his social network.

In 1853, Dunant had the good fortune of meeting Harriet Beecher Stowe, the American author of *Uncle Tom's Cabin* (1852), during her stay in Geneva. She was on tour "pleading the cause of humanity in old Europe" and happened to stay at the home of a relative of General Dufour.[8] Stowe, and her literary peers Charles Dickens, Victor Hugo and Emile Zola, were writing a form of humanitarian narrative meant "to arouse people to a crusade against wrong through books or appeals".[9] Before photography's invention, graphic language and illustrations (e.g., prints) were increasingly incorporated into various popular and authoritative texts – from the novel, to government inquiries, to medical autopsies – with the express purpose of shaping perceptions and generating sympathies. For social historian Thomas Laqueur, these texts constituted a rubric of sorts that he termed "humanitarian narratives". Their frequency and persistence were a sign, according to Laqueur, that "some people [had] begun thinking and feeling in new ways" and had thus begun to regard a wider swath of the population around them as part of a humanity worth caring for.[10] Generally the arc of these chronicles would include a victim, almost always described as innocent, who would struggle with a villain (e.g., disease, disaster, or an individual or group causing suffering) and be saved by a hero, in most cases either a technology or a person of socio-economic privilege. The "sensationalistic" details of suffering experienced or inflicted on the affected person's body operated on two fronts: to represent the truth of the claims, and to "demand attention and sympathy" on the part of the readers.[11] Such compositions were expected to form a common bond between the reader and the victim, with the personalization of the victim's misery anticipated to nurture the "moral imagination", thus contributing to the generation of a more humanitarian outlook towards subjects who until that time had "been beneath notice".[12] As such, humanitarian narratives are central to the sort of "sentimental education" that Rorty has associated with the sharing of sad stories. Indeed, the purpose of such manipulations of feeling, particularly those aimed at "increasing [the] ability to see similarities between ourselves and people very unlike us", is "to expand the

7 Karen Halttunen, "Humanitarianism and the Pornography of Pain in Anglo-American Culture", *American Historical Review*, Vol.100, No. 2, 1995, p. 307.

8 See R. Durand, *Henry Dunant*, above note 6, p. 67. Dufour was in many ways Dunant's mentor: he would later help arrange the battlefield meeting between Dunant and Napoleon that would lead Dunant to stumble across – quite by accident – the Battle of Solferino, and he eventually became one of the five founding members of the Red Cross.

9 M. Grumpert, above note 5, p. 15. Grumpert continues: "During their lifetime, they were honoured and admired, and even now are accorded full rites by the motion picture industry. It was felt that even the shadow of pain and suffering was a last damnable blot on modern civilization."

10 Thomas W. Laqueur, "Bodies, Details, and the Humanitarian Narrative", in Lynn A. Hunt (ed.), *The New Cultural History*, MPublishing, University of Michigan Library, 1989, p. 200.

11 *Ibid.*, p. 184.

12 *Ibid.*, pp. 176, 191.

reference of the terms 'our kind of people' and 'people like us'".[13] Rorty sees this practice as a critical aspect of developing a more equitable and just world. Recent scholarship on humanitarian visual culture identifies a similar narrative association within humanitarian photography.[14]

In her text, Stowe relied heavily on graphic depictions of the bodies of slaves and the violence done to them, in order to represent their full humanity.[15] Her writing and advocacy skills substantially influenced the young Dunant.[16] Mobilizing imagery equivalent to that of Stowe's, he wrote: "I moistened his dry lips and hardened tongue, took a handful of lint and dipped it in the bucket they were carrying behind me, and squeezed the water from this improvised sponge into the deformed opening that had been his mouth."[17] Describing another soldier who had been peppered with grapeshot, he wrote: "The rest of his swollen body was all black and green, and he could find no comfortable position to sit or lie in. I moistened great masses of lint in cold water and tried to place this under him, but it was not long before gangrene carried him off."[18]

It was not until three years after the Battle of Solferino that Dunant would write and publish this account of his experience providing care to wounded combatants. The passing of time did little to diminish the details he would recall, or his emotions surrounding the events. Dunant's publication was an account of the *ad hoc* care he provided to soldiers wounded in the battle. Not medically trained, Dunant could do little more than provide comfort and company to those who might survive and to those who would surely die: essentially, rudimentary palliative care. For the bulk of the book, Dunant focuses on combatants overcome with pain and suffering, the eye-witnessing of which spurred him to action – and, he expected, would surely do the same for his readers. Dunant sets up his accounts of the battle as those of "a mere tourist with no part whatever in this great conflict[,] … witness [to] the moving scenes that I have resolved to describe".[19] He explicitly invites his readers to engage the narrative visually as he

13 R. Rorty, above note 1, pp. 129, 122–123.

14 Heide Fehrenbach and Davide Rodogno (eds), *Humanitarian Photography: A History*, Cambridge University Press, Cambridge, 2015.

15 Subsequent literary critics have differentially interpreted and debated Stowe's use of violence in her novel. For example, in her analysis "The Ecstasies of Sentimental Wounding in *Uncle Tom's Cabin*", Marianne Noble puts her ideas into dialogue with those of Laqueur and Halttunen, who each find different motivations and impacts of Stowe's mobilization of violence in the book. Noble reveals the ways in which the humanitarian narrative is a "double edged sword", precariously balancing liberation and repression, awareness-raising and objectifying. What this criticism points to is a long-standing, perpetual paradox of potentially doing harm while trying to do good by objectifying victims and their pain. See Marianne Noble, "The Ecstasies of Sentimental Wounding in *Uncle Tom's Cabin*", *Yale Journal of Criticism*, Vol. 10, No. 2, 1997.

16 Abraham Lincoln is said to have believed that Stowe's book played a part in bringing about the American Civil War, a conflict that was still nearly a decade away when she and Dunant met. M. Grumpert, above note 5, p. 15. According to Hamand, "Abraham Lincoln is alleged to have said upon meeting Harriet Beecher Stowe, 'so this is the little lady who made this big war'." Wendy F. Hamand, "'No Voice from England': Mrs. Stowe, Mr. Lincoln, and the British in the Civil War", *New England Quarterly*, Vol. 61, No. 1, 1988, p. 3.

17 H. Dunant, above note 4, p. 62.

18 *Ibid.*, p. 67.

19 *Ibid.*, p. 16.

lived it: to become affected witnesses alongside him. Dunant's writing, his selection of words and his choice of scenes to describe are meant to grip readers and encourage "the onlooker" and fellow "tourists" to visualize in their mind's eye repulsive panoramas of battle and the concomitant grievous suffering and death.[20]

Importantly, Dunant deviated from Stowe in one crucial way. Unlike Stowe's famous work, Dunant wrote himself in as the hero (or what would be characterized today as the hero), albeit not already in possession of heroic qualities. Dunant's story is about his transformation from disaffected tourist to advocate for the care of wounded soldiers. His text has many of the characteristics of a classic *Bildungsroman* or coming-of-age story. Joseph Slaughter has positioned the *Bildungsroman* as an ideal form of writing in the service of humanitarianism and human rights.[21] Conventional humanitarian narratives that centralized suffering and graphic descriptions of pain as the source of moral development, Slaughter argues, offered less guarantee that readers would respond in the ways that writers intended. Indeed, long-standing anxieties accompanied the creation and critique of these stories over their potential to be interpreted by audiences as entertainment or worse.[22] To centrally feature the main character's development took the focus off bodily suffering and instead concentrated on encouraging reader emulation through a "sense of responsibility to [the] moral integrity of one's own class of humanity".[23] The benefit of this was that readers were not left to their own devices to determine "proper sentiment" and what actions should be taken in response to that emotion.[24] Such a focus on the development of the protagonist is an important part of the sort of story which Rorty identifies as crucial in sentimental education that can help to bring about a more inclusive and diverse human family.

Repeatedly in *Solferino*, Dunant steers spectators past horrific scenes – but he also guides them along a direct and actionable path. In the vein of the classic *Bildungsroman*, Dunant provides examples of proper sentiment and encourages its development in his readers:

> The moral sense of the importance of human life; the humane desire to lighten a little the torments of all the furious and relentless activity which a man summons up at such moments: all these combine to create a kind of energy which gives one a positive craving to relieve as many as one can.[25]

20 *Ibid*., pp. 22, 65.
21 Joseph Slaughter, *Human Rights Inc.*, Fordham University Press, Brooklyn, NY, 2007.
22 Kevin Rozario, "'Delicious Horrors': Mass Culture, the Red Cross, and the Appeal of Modem American Humanitarianism", *American Quarterly*, Vol. 55, No. 3, 2003, available at: https://muse.jhu.edu/article/46647.
23 Lynn Festa, *Sentimental Figures of Empire in Eighteenth-Century Britain and France*, Johns Hopkins University Press, Baltimore, MD, 2006, p. 103; see also Jane Lydon, *Photography, Humanitarianism, Empire*, Bloomsbury, London, 2016.
24 K. Halttunen, above note 7, p. 307.
25 H. Dunant, above note 4, p. 73.

Dunant focuses on drawing attention to suffering as something equally felt by soldiers, officers, allies and enemies.[26] He mobilizes his text to show this consciousness for the benefit of his readers:

> [The] women of Castaglione, seeing that I made no distinction between nationalities, followed my example, showing the same kindness to all these men whose origins were so different, and all of whom were foreigners to them. "*Tutti fratelli*" ["All are brothers"], they repeated feelingly.[27]

Dunant's witnessing and his first-hand experiences in the aftermath of fighting transformed his perceptions when it came to battlefield care. His book, with its graphic text, became a proxy to his readers, a substitute for actually standing on the sidelines of the battle with him. In the final section of the book, Dunant presents a sketch of an unprecedented plan: "Would it not be possible, in time of peace and quiet, to form relief societies for the purpose of having care given to the wounded in wartime by zealous, devoted and thoroughly qualified volunteers?"[28] All of this occurs in a little over 100 pages.

In writing as graphically as he did, Dunant was tapping into the power of images – even if only ones appearing in the mind's eye of the reader – to affect people, to touch them emotionally so that they might act. "People" in this context refers, of course, to those within his social circle, people with the means to effect political or material change, to whom he was directing his message. Dunant's appeal was aimed at "comfortably situated citizens possessing an international spirit of good works",[29] adding credence to Rorty's argument that sentimental stories are best addressed to the people "on top", those with political and material means, as they are the most likely to be able to effect change. This is the same class of people who are in positions of sufficient security "to relax long enough to listen" to appeals such as those of Dunant.[30]

26 Florence Nightingale famously stated following her experience at Crimea: "Suffering lifts its victim above normal values. While suffering endures there is neither good nor bad, valuable nor invaluable, enemy nor friend. The victim has passed to a region beyond human classification or moral judgments and his suffering is a sufficient claim." See British Red Cross, "Florence Nightingale and the Red Cross", 20 August 2017, available at: www.redcross.org.uk/stories/health-and-social-care/health/how-florence-nightingale-influenced-the-red-cross##. Ironically, Nightingale was against Dunant's plan, arguing that it would lead to governments relaxing their responsibilities towards their fighting forces. Nightingale's caution has since been termed the "Nightingale risk" in conflict studies circles. See Katherine Davies, *Continuity, Change and Contest: Meanings of "Humanitarian" from the "Religion of Humanity" to the Kosovo War*, Humanitarian Policy Group Working Paper, Overseas Development Institute, 2012, p. 5; R. Durand, *Henry Dunant*, above note 6; Eleanor O'Gorman, *Conflict and Development*, Zed Books, New York and London, 2011.

27 H. Dunant, above note 4, p. 72.

28 *Ibid.*, p. 115.

29 David P. Forsythe, *The Humanitarians: The International Committee of the Red Cross*, Cambridge University Press, Cambridge, 2005, p. 17.

30 R. Rorty, above note 1, pp. 128, 130. See also D. P. Forsythe, above note 29, p. 17: those who had "the money and leisure time to make that spirit count for something, proved receptive to Dunant's ideas. Dunant would ultimately draw upon what might be termed Genevan exceptionalism: the collective self-image, no doubt partially the product of Calvinism, that the citizens of Geneva constituted a special people with a positive role to play."

To achieve his aims, Dunant reconfigured and fictionalized some of his experiences for added affect. He also provided his readers with a clear path of action towards which they could direct their energies.[31] Much of the credit for the success of *A Memory of Solferino* has to go directly to Dunant. He was adept at working his social network and rallying influential people to a cause.[32] Building on these abilities, Dunant circulated *Solferino* across Europe and discussed its contents and ideas with influential leaders in his social circle.[33] He did so in person with members of the European nobility, political leaders, ministers of defence and military physicians. He also toured his book through the salons of Paris.

High praise followed from various arenas of the political and cultural elite. The Brothers Goncourt, popular social commentators of the time, noted in their *Journal* of 8 June 1863: "One finished this book by damning war."[34] Among other philosophers and philanthropists, Victor Hugo wrote in a letter to Dunant: "I have read your book with the greatest interest. You are arming humanity and are serving freedom … I endorse your noble efforts with enthusiasm, and I send you my heartiest good wishes."[35] Also in 1863, "England's most popular author", Charles Dickens, published "The Man in White" in his weekly journal, *All the Year Round*; this account "was devoutly read by the English-speaking world, [presenting] a detailed analysis of the book of the 'travelling amateur' and his difficult and courageous attempt to alleviate the misery of war".[36]

Dunant's book and the emotions it aroused might have been destined to fade from the memories of the influential people he met with had it not been, again, for the intrepid and sympathetic General Dufour. It was Dufour who connected Dunant with Gustave Moynier, then president of the Geneva Society for Public Welfare;[37] together with Louis Appia and Théodore Maunoir, these would be the five founders of what eventually became the Red Cross.

A Memory of Solferino has been credited as having sparked the creation of the International Committee of the Red Cross (ICRC),[38] but Durant's call for a relief society to be viewed by his readers as a "damnation of war" and an "armament" in

31 D. P. Forsythe, above note 29. See also N. Klein-Kelly, above note 6.

32 While at Brescia, where Dunant spent most of his time providing what care he could to wounded combatants from the Battle of Solferino, Dunant made appeals to philanthropists in Geneva to donate funds and supplies. Dunant's contacts also included the Dutch Royal Family, establishing a line of royal patronage with the Red Cross that continues to this day. R. Durand, *Henry Dunant*, above note 6; Caroline Moorehead, *Dunant's Dream: War, Switzerland, and the History of the Red Cross*, HarperCollins, London, 1998.

33 R. Durand, *Henry Dunant*, above note 6. Dunant developed his networking proficiency while internationalizing the Young Men's Christian Association (YMCA) in the decade before the battle of Solferino. He was employed by the YMCA while the First War of Italian Independence (1848–49) and the Russo-Turkish War (1853–56) raged. The Crimean War was a campaign in the Russo-Turkish War; the Battle of Solferino would be fought during the Second War of Italian Independence in 1859.

34 Cited in M. Grumpert, above note 5, p. 84.

35 Cited in *ibid.*, p. 85.

36 *Ibid.*, pp. 84–85. Dickens' title is a reference to the moniker given to Dunant by the soldiers he tended to; he had arrived at the aftermath of the battle in the tropical colonial suit he wore for his meeting with Napoleon II.

37 H. Dunant, above note 4, p. 35.

the service of freedom was made all the more possible because everyday people were already starting to think differently about warfare and its impact on infantrymen. The opinions and designs that Dunant proposed in his book could only take hold in a climate already receptive to them. An example of the growing distaste for war at the time can be seen in the work of Francisco Goya; some four decades earlier, the Spanish court painter, known for his idyllic scenes of everyday life among the gentry, became a *de facto* documentarian and commentator of the cruelties that accompanied Napolean Bonaparte's invasion of Spain, leading to the Peninsular War of 1808–14.

While his peers created works that glorified war or depicted the miseries of warfare as unavoidable, Goya broke from painterly traditions to present horrific scenes of brutality, famine and repression with a dramatically accusatory tone.[39] He combined each unflinching image of brutality and despair with titles that explicitly expressed strong sentiments against the conflict. With titles such as *They Do Not Want To*, *Bury Them and Keep Quiet* and *There Is No One to Help Them*, Goya guided spectators to a new interpretation of war as abhorrent and repulsive.

Despite Goya's reputation, the prints were never put into wide circulation in his lifetime.[40] By 1863, when Henry Dunant was circulating *Solferino* throughout Paris and among the European elite, changes in public opinion resulted in Goya's work taking on new significance. Sentiments that earlier had appeared unconventional and controversial had become widespread and accepted, as the "fatal consequences" of numerous battles taking part in quick succession were being made visible through the new medium of photography.[41] Goya's prints were newly published for mass circulation in 1863 under the collective title *The Disasters of War*. Goya has since been credited with having introduced the sentiment of revulsion for warfare in the nineteenth century.[42] Furthermore, ICRC senior medical officer Paul Bouvier has presented a convincing interpretation of Goya's prints as depicting the essence of humanitarian action.[43] Undoubtedly, Goya's work contributed to changing social relations in the face of conflict.

38 This credit comes not only from the ICRC, but also from critical humanitarian scholars such as David Rieff. See J. Slaughter, above note 21, p. 327.

39 In this regard, Goya is believed to have been influenced by Jacques Callot's *Les Grands Misères de la Guerre*, 1633.

40 It is speculated that the prints were too critical of the French, then the European powerhouse. See Jonathan Jones, "Look What We Did", *The Guardian*, 20 June 2002, available at: www.theguardian.com/culture/2003/mar/31/artsfeatures.turnerprize2003.

41 Originally published under the title *Fatal Consequences of Spain's Bloody War with Bonaparte, and Other Emphatic Caprices*, Goya's prints contain many images of violence toward civilians, mainly women. It would be almost a century from the time he made his pictures before humanitarian laws to protect women, children and other non-combatants would come into effect.

42 Sharon Sliwinski, *Human Rights in Camera*, University of Chicago Press, Chicago, IL, 2011; Susan Sontag, *Regarding the Pain of Others*, Picador, New York, 2004.

43 Paul Bouvier, "In Folio: '*Yo Lo Vi*'. Goya Witnessing the Disasters of War: An Appeal to the Sentiment of Humanity", *International Review of the Red Cross*, Vol. 93, No. 884, 2011.

Deep philosophical roots underpin the long-standing belief in the persuasive link between sight and moral sensibilities. Long before Dunant's time, a succession of Scottish Enlightenment thinkers had prioritized scenes of suffering as central to exciting the imagination, enabling spectators to "enter into" the body of the sufferer and develop a proper sense of what was required to end the other's distress.[44] More recently, visual theorist Sharon Sliwinski has reflected specifically upon the role of the camera in advancing humanitarian sentiment and the codification of human rights. Sliwinski outlines how spectators' "passionate responses" to visual images, such as outrage, disgust, sorrow or frustration, "can be read as a sign that demonstrates the 'moral character' of humanity".[45] It is these sensorial reactions, these aesthetic responses to images of suffering and calamity, that precede – indeed, spark – actions to better humanity. The role of images in this moral progression is to show examples of situations that call for human intervention, and also to show the viewer *how* to intervene. Dunant provided aesthetic scenes by way of his "lamentable pictures" revealing suffering that called for a response, and he extended those scenes to allow spectators a chance to see how such intervention could be done to the betterment of many. While Dunant's ideas generated support for concrete actions to ameliorate the conditions for common soldiers, this expansion of humanitarian sentiment was made all the more possible because contemporary conflict photography (and related commentary) heralded these passionate responses. It is this contextual moment that is the focus of the following section.

New media and the acceleration of moral sentiment

The Battle of Solferino has been described as the most modern conflict of the nineteenth century. The advent of steam technology, the telegraph and photography made it possible for communications to travel greater distances at unprecedented speeds and in unsurpassed detail.[46] With public opinion weighing in almost hourly, this swift movement of information – albeit within particular privileged circles with access to less commonplace technologies – was the nineteenth-century equivalent of news "going viral".[47] Alongside Solferino, two other major conflicts in the decade 1855–65 received much popular, political and media attention: the Crimean Campaign and the American Civil War.

44 For instance, David Hume, Francis Hutcheson, and Adam Smith: see K. Halttunen, above note 7, p. 307. See also K. Rozario, above note 22.

45 S. Sliwinski, above note 42, pp. 5, 23.

46 M. Grumpert, above note 5, p. 39.

47 Simon J. Potter, "Webs, Networks, and Systems: Globalization and the Mass Media in the Nineteenth- and Twentieth-Century British Empire", *Journal of British Studies*, Vol. 46, No. 3, 2007, available at: www.jstor.org/stable/10.1086/515446?seq=1. See also Jeremy Stein, "Reflections on Time, Time-Space Compression and Technology in the Nineteenth Century", in Jon May and Nigel Thrift (eds), *Timespace: Geographies of Temporality*, Routledge, London, 2001.

By the time that Dunant was writing *A Memory of Solferino*, photography had been in popular use for nearly two decades. The state of the art had moved away from one-off daguerreotypes and now consisted of a collodion process using glass-plate negatives to which a wet emulsion was applied just prior to exposure.[48] When exposed outside a studio, these plates had to be developed in travelling darkrooms. Photographs were circulated as individual prints, displayed in exhibitions and projected in lantern lectures. They were also readily translated into woodblock prints for mass circulation in newspapers and books.

Photographers, journalists and media moguls were quick to take advantage of the ability to bring the camera onto the battlefield: to impress their existing audiences, to entice new ones, or for the sheer challenge of taking pictures amidst a battle.[49] In 1842, the *Illustrated London News* launched with bold, centrally featured prints – often made from photographs – depicting calamities and conflict, with the express conviction that "disaster could push newspaper sales".[50] Meanwhile, governments and political leaders mobilized the new media technologies to try to sway public opinion in favour of their actions, including conflicts that put the lives of many of their citizens at risk. Regardless of the intent on the part of States, editors, writers or photographers, there was no guarantee that spectators would interpret what they heard, read or saw in the same way. Indeed, the advent of the camera enabled people's perceptions to change in unexpected ways;[51] this was no less true regarding conflict photography, beginning with one of the first conflicts where the camera was present, the Crimean War.

The war that would make Florence Nightingale famous was also a war that was tremendously unpopular in the UK after British involvement unexpectedly dragged on through an uncommonly harsh winter. The British suffered great losses in the Crimean War due to what was seen as mismanagement that resulted in troops being severely undersupplied. It is said that more soldiers were lost to

48 A dry collodion process was also possible, but most photographers of the time used the wet process because, as one photographer noted, the dry process was "too slow to be employed where the exposure must only occupy a short time". See J. L., "Photography at the Seat of War", *Photographic News*, Vol. 2 No. 42, 24 June 1859, p. 183.

49 Jason E. Hill and Vanessa Schwartz (eds), *Getting the Picture: The Visual Culture of the News*, Bloomsbury Academic, London and New York, 2015. Also, J. L. wrote this account of his endeavour to take photographs of the hostilities mounting in Italy: "When I left England my intention was to make a tour with the camera in Switzerland, but the exciting prospect of being able to get plates of battle-fields, sieges, and other incidental scenes, induced me to change my course, and, instead of remaining among the glaciers and ice-peaks, to make a journey to the sunny plains of Italy." J. L., above note 48, p. 183.

50 With crimes, disasters and calamities as its regular fare, Herbert Ingram's *Illustrated London News* focused on more serious issues than another British magazine, *Punch*, that launched around the same time. Concerned about readership, sales, social conventions and Victorian sensibilities, the *Illustrated London News* carefully meted out its news with a strong dose of excitement and entertainment. See Paul Hockings, "Disasters Drawn: *The Illustrated London News* in the Mid-19th Century", *Visual Anthropology*, Vol. 28, No.1, 2015, p. 22.

51 Walter Benjamin was one of the first to critically theorize photography as a social phenomenon, persuasively describing the power of the medium to shift perceptions. See Walter Benjamin, *The Work of Art in the Age of Its Technological Reproducibility and Other Writings on Media*, Belknap Press of Harvard University Press, Cambridge, MA, 2008.

exposure and disease than to enemy fire. During the unforgiving winter of 1854, William Russell of *The Times* of London shared narrative dispatches that painted "grim pictures" of the events.[52] Such reports differed greatly from government propaganda or from traditional historiographies that glorified warfare. Eventually, public criticism would become intense enough to lead to the resignation of the Aberdeen government in February 1855. It was in the spring and early summer of that year that Roger Fenton went to photograph the campaign. A matter of contention to this day, it is said that his supporters hoped to turn the tide of public sentiment with photographs that could command loyalty and build patriotism – essentially, propaganda.[53]

Among the earliest to photograph conflict, Fenton was a professional photographer commissioned by the English print-seller Thomas Agnew to photograph the Crimean War.[54] With his history of being a photographer to the Royal Family, Fenton also travelled under Royal Patronage and with the backing of the British government. Though not the only photographer taking pictures of the campaign, Fenton is credited as being the most prolific and is certainly the most recognized today.[55] The commission also entailed taking photographs to be used as source material by the renowned painter Thomas Jones Barker, which may explain Fenton's creative decisions.[56] Over the course of three months in the spring and early summer of 1855, Fenton succeeded in taking some 350 exposures, which were shared to broad public audiences through portfolio publications and travelling exhibitions.[57]

Fenton's photographs were always products of careful composition. With exposure times measured in seconds or minutes (as opposed to today's fractions of a second), heavy box-type large-format cameras, and the need to develop the plates immediately after being exposed, the technology at the time prevented him from taking candid exposures or action shots, and limitations on the subject matter may have also been imposed by his commission. The result is that his pictures have since been described as dull.[58] Compared to today's conflict photography made in the thick of battle, it is understandable that Fenton's images can come across as stale and outdated. Such an anachronistic perspective

52 T. J. Brady, "Roger Fenton and the Crimean War", *History Today*, Vol. 18, No. 2, 1968, p. 80.

53 *Ibid.*, p. 76. For a recent counter-argument to that which Brady references, see Sophie Gordon, *Shadows of War: Roger Fenton's Photographs of the Crimea, 1855*, Royal Collection Trust, London, 2017.

54 The publisher intended to turn a profit through the sale of postcards and portfolios, which were popular forms of circulating photographic prints at the time. See T. J. Brady, above note 52; Beaumont Newhall, *The History of Photography: From 1839 to the Present*, Museum of Modern Art, New York, 1982; J. Marwil, above note 2.

55 T. J. Brady mentioned several others who were commissioned to take photographs of the Crimean War: Richard Nicklin in 1854, and "two young officers – Ensigns Brandon and Dawson" in the spring of 1855. None of their photographs appear to have survived to the present. T. J. Brady, above note 52, p. 76.

56 S. Gordon, above note 53, p. 40.

57 Fenton travelled with his assistant by ship to Sevastopol, where they converted a four-horse-drawn wine cart into a mobile darkroom. Reproductions of Fenton's photographs were circulated later in 1855 among the British Royal Family and Napoleon III's court. An exhibition containing 312 prints was put on public display to thousands in both London and Paris. See T. J. Brady, above note 52, pp. 76, 83; S. Gordon, above note 53, p. 40.

58 J. Marwil, above note 2, p. 32.

led art critic Beaumont Newhall to conclude that Fenton had to "resolve [himself] to the still life in the aftermath of battle".[59] Newhall possibly made this statement with photographs like Fenton's most well-known photograph of the Crimean War in mind (see Figure 1).

The Valley of the Shadow of Death (1855) is a photograph of a dry and barren landscape transected by a roadway.[60] There are in fact two exposures made from the same tripod position. In one the road lays empty; in the other, cannonballs are strewn across it. Apparently, Fenton had cannonballs from the ditches repositioned to give the picture a different effect. While it was the exposure with the projectiles that was most widely circulated, the contemporary impact of these particular photographs has been overshadowed by subsequent, and still ongoing, debate about the integrity of the image.[61] Fenton did not, however, only photograph in the aftermath of battle. The majority of his war photographs were of soldiers and officers posed singularly or in groups in the British military camps.

Using the camera to its best advantage, Fenton meticulously posed his subjects in ways that accentuated their discipline, their camaraderie, their strength and their vitality. To be fair to Newhall, although technically superior to many of his peers, Fenton's conflict images were relatively staid compared to those of some of his contemporaries. Photographers practicing in conflict settings, then as now, had ample graphic subject matter on which to train their lenses. War photographs depicting the physical effects of conflict on human bodies were being taken, with the earliest known coming from the 1847 Mexican–American War depicting a battlefield amputation (see Figure 2). Fenton's sedate picture content was more likely a political-economic choice than due to failings of technology or its operator. Photographs of injury and death would not have suited his patron's goal.

Fenton's photographs of strapping soldiers and officers in full battle dress (see Figure 3) may have bolstered patriotic sentiments in the hearts of some of his viewers. Despite the anticipation of renewed support for the military engagement, the photographs also offered opportunities for perceiving the conflict in different ways. Following the 1855 London exhibition of the Crimea photographs, a journalist with *The Times* wrote that the pictures presented the "private soldiers [with] as good a likeness as the general".[62] To make such a remark may have simply been a statement of observable fact, but in the highly structured, hierarchical British society, to say that people of different social ranks were treated as equals – even if only in photographs – was quite an act of levelling.

59 B. Newhall, above note 54, p. 85.

60 The title was taken from the moniker that soldiers had given to another valley in Sevastopol. The nickname referenced both Psalm 23 and Tennyson's popular 1854 poem "Charge of the Light Brigade", based on the Battle of Balaclava that took place before Fenton arrived.

61 Michael Zhang, "Famous 'Valley of the Shadow of Death' Photo Was Almost Certainly Staged", *PetaPixel*, 1 October 2012, available at: https://petapixel.com/2012/10/01/famous-valley-of-the-shadow-of-death-photo-was-most-likely-staged/.

62 Quoted in B. Newhall, above note 54, p. 85.

Figure 1. *The Valley of the Shadow of Death*, by Roger Fenton, 1855. Library of Congress, LC-DIG-ppmsca-35546.

According to the philosopher Judith Butler,

> there are ways of framing that will bring the human into view in its frailty and precariousness, that will allow us to stand for the value and dignity of human life, to react with outrage when lives are degraded or eviscerated without regard for their value as lives. And there are frames that foreclose responsiveness.[63]

The framing Butler refers to is only partly to do with the way in which the photographer situates the subject within the camera's optical parameters. It also refers to situating the image within a set of political and ideological boundaries, in order to limit the surfeit of meaning that can accompany photographs.[64] It is unclear from the historical record what sort of "framing" Fenton intended with his images. As for his patrons, they may well have been framing his photographs in an effort to bolster public support for the Crimean campaign. However, once on public exhibition to thousands of everyday, common people, the photographs could not be guaranteed to remain bound to the ideals of the ruling class. Ordinary citizens whose sons were among the "private soldiers" referenced in

63 Judith Butler, *Frames of War: When Is Life Grievable?*, Verso, London and New York, 2009, p. 77.

64 *Ibid.* See also Stephen D. Reese, "Framing Public Life: A Bridging Model for Media Research", and James Tankard, "The Empirical Approach to the Study of Media Framing", in Stephen D. Reese, Oscar H. Gandy and August E. Grant (eds), *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*, Lawrence Erlbaum, Mahwah, NJ, 2003.

Figure 2. *Amputation, Mexican–American War, Cerro Gordo, 1847*, photographer unknown, 1847. National Photographic Archive, National Institute of Anthropology and History, Mexico City, Inventory No. 839971.

The Times may themselves have read something different than national pride in Fenton's images – particularly when placed in association with the "grim pictures" reported on the previous year, or alongside those reported in letters sent home from soldiers.[65] Spectators may not all have been conscious of the democratizing potential of Fenton's pictures, but the camera was making this more of a possibility. Once raised in profile, it was no great leap for photography to contribute to making the common soldier's life "grievable".[66] This can also be seen in the work and discourse of other eyewitnesses and photographers to European conflicts happening a few short years later. Several professional and amateur photographers have been identified as having taken pictures at the battles that were part of the Franco-Austrian War, of which Solferino was the final, decisive one.[67] Compared to Fenton's, the numbers of photographs that have survived are much fewer, and none have attained the status of his Crimean

65 Cambridge University Library, "The Crimean War Letters of Captain Blackett", *Cambridge University Library Special Collections*, 2012, available at: https://specialcollections-blog.lib.cam.ac.uk/?p=2308.

66 J. Butler, above note 63.

67 William Johnson, "Combat Photography during the Franco-Austrian War of 1859", 21 August 2017, available at: https://vintagephotosjohnson.com/2012/02/18/combat-photography-during-the-franco-austrian-war-of-1859/; see also J. Marwil, above note 2. The First Italian War of Independence (1848–49) was followed a decade later by the Second Italian War of Independence, also known as the

Figure 3. *Private Soldiers and Officers of the 3rd Regiment (The Buffs) Piling Arms*, by Roger Fenton, 1855. Library of Congress, LC-USZC4-9289.

pictures. Despite being the bloodiest battle of the nineteenth century, Solferino has not had the same lasting impact on popular memory as the Crimean War that preceded it or the American Civil War that would shortly follow. The dearth of Solferino photographs compared to the wealth of those produced in the other two battles is a likely contributing factor to Solferino's conflict being all but forgotten outside of Italy or humanitarian action circles. The contemporary comments about their creation, however, are telling: they reflect the impact of photography on changing perceptions of lives worth caring for.

A couple of sets of stereograph pictures exist from the Franco-Austrian War that in many respects share equivalences with Fenton's images.[68] Stereographs were a popular form of photography, particularly among landscape photographers; these dual-exposure pictures, when seen through a specially designed viewer, appeared three-dimensional.[69] The largest remaining collection

Franco-Austrian War (1859). The Battle of Solferino was the final battle in these wars. See C. Moorehead, above note 32.

68 W. Johnson, above note 67.

69 Stereographs fell out of favour at the end of the nineteenth century as the technology was not easily adaptable for commercial consumer cameras. As aid organizations today turn to techniques such as 360-degree photography and virtual reality, it would seem that 3-D technology is ripe for a comeback. The affective force of these technologies remains to be seen, but will likely emulate the pattern of historical technological innovations, including the stereograph.

from the Franco-Austrian War comes from the Gaudin Brothers, who were professional French photographers commissioned to take photographs of French soldiers and their allies in bivouac (see Figure 4). Like Fenton's, these too are clearly staged and carefully composed. Taking photographs of regular troops may have been a novel "means of memorializing military accomplishments", but it likewise presented a commercial opportunity.[70] Such pictures, which had few equivalents in earlier conventions of celebratory battle art, also communicated the scale and quality of lives put at risk.

Fewer in number are photographs attributed to Jules Couppier.[71] His images are of the battlefields and were made in the days following several of the conflicts that were part of the Franco-Austrian War, which lasted from 26 April to 11 July 1859. Unlike Fenton's pictures, these are panoramas taken at a distance. Fenton's *Valley* picture may have made the battlefield appear more palpable with its proximal composition, but Couppier and Fenton's battlefield scenes are equally devoid of human figures. Couppier, however, did take a certain type of photograph that Fenton did not, or could not: amidst the Couppier collection are a few photographs that contain images of the wounded and the dead. A stereograph of "a convoy of the wounded and survivors" has survived from Solferino.[72] It is taken at high angle, perhaps from a church bell tower, and shows carriages filled with injured soldiers, stretching as far back as the eye can see, being moved to makeshift medical facilities in the town of Brescia (see Figure 5). Dunant referred to similar scenes, describing a "long procession of Commissary carts" with "all ranks mixed up …[,] bleeding, exhausted, torn, and covered with dust".[73]

From the Battle of Magenta, which took place two weeks before Solferino, there exists a striking picture of a pile of corpses awaiting burial at the local cemetery (see Figure 6 and detail in Figure 7). The fact that this image was made into a stereograph enhances its affective potential – to see a photograph of a mass of bodies was itself shocking and new, but to see it in three-dimensional quality would have been harrowing. This is precisely what Oliver Wendell Holmes suggested in his account of encountering a stereograph (possibly this same image) of a "heap of dead lying unburied" at the cemetery at Melegnano in his friend's collection of pictures.[74] The American poet, physician, essayist, and co-founder of *The Atlantic Monthly* wrote:

70 W. Johnson, above note 67.

71 The photographs attributed to Jules Couppier and reproduced here have been researched by Janice Schimmelman, who has also researched Claude-Marie Ferrier, another contemporary photographer considered the possible creator of these stereographs. Based on a variety of factors including handwriting comparison, Schimmelman concludes that these images are by Couppier. The Couppier and Gaudin Brothers images are reproduced with permission from the personal holdings of photographic historian William G. Johnson. See John B. Cameron and Janice G. Schimmelman, *The Glass Stereoviews of Ferrier and Soulier, 1852–1908*, Collodion Press, Rochester, MI, 2016; Janice G. Schimmelman, *Jules Couppier: Glass Stereoviews, 1853–1860*, Collodion Press, Rochester, MI, 2018.

72 W. Johnson, above note 67.

73 H. Dunant, above note 4, pp. 53–54.

74 Oliver Wendell Holmes, "Sun-Painting and Sun-Sculpture: With a Stereoscopic Trip Across the Atlantic", *The Atlantic Monthly*, Vol. 8, No. 45, 1861, p. 27; see also J. Marwil, above note 2.

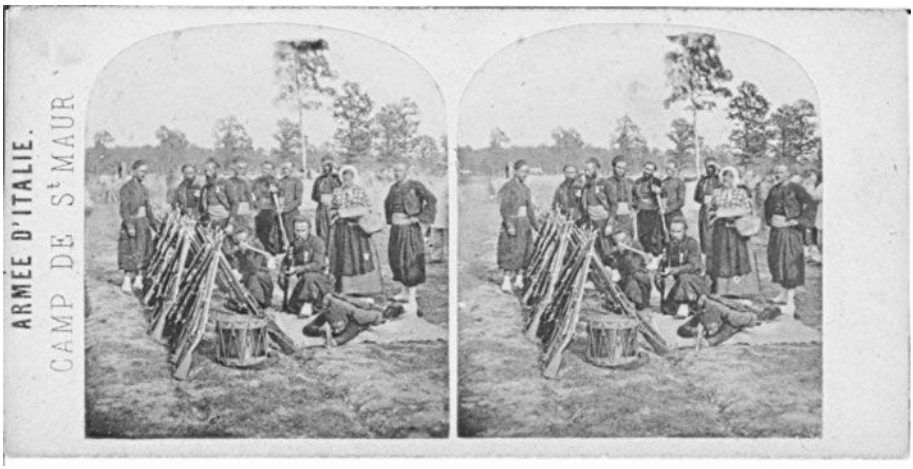


Figure 4. *French Army in Bivouac, 1859*, by the Gaudin Brothers, 1859. Stereograph. Courtesy of vintagephotosjohnson.com.

> Look away, young maiden and tender child, for this is what war leaves after it. Flung together, like sacks of grain, some terribly mutilated, some without mark of injury, all or almost all with a still, calm look on their faces. The two youths, before referred to, lie in the foreground, so simple-looking, so like boys who had been overworked and were lying down to sleep, that one can hardly see the picture for the tears these two fair striplings bring into the eyes.[75]

Holmes, whose own son would later go missing for a time during the Civil War, lamented the indignity done to the bodies of "simple" "boys" and the death of these "two fair striplings".[76] The lives of common troops – even those an ocean away – were generating passionate sentiments in powerful people. As an example of the discourse around international news that circulated among the influential literati, Holmes' reaction is also an exemplar of an aesthetic encounter that generated passionate responses in the author, who subsequently provided his audience with cues for developing humanitarian sentiment. Though not an eyewitness like Dunant, Holmes demonstrates the impact of photographs on his own moral progress. The way in which photography plays with time and space created a presence for Holmes akin to – though not equivalent to – the eyewitness experience, such that he came to consider the common soldiers as sons to be grieved.

The growing popular use of photography brought with it "the emergence of a new relation toward the visual".[77] It did so by seemingly collapsing distances in time and space, making it appear as though viewers were witnessing scenes

75 O. W. Holmes, above note 74, p. 27.
76 *Ibid.*
77 Ariella Azoulay, *Civil Imagination: A Political Ontology of Photography*, Verso, London, 2012, p. 65.

Figure 5. *Vue de l'Avenue do Brescia avec convoi de blesses et de Vivres* [*View of the Road to Brescia, with Convoy of the Wounded and Survivors*], attributed to Jules Couppier, 1859. Stereograph of the aftermath of the Battle of Solferino, c. 24–26 June 1859. Courtesy of vintagephotosjohnson.com.

Figure 6. *702. Vue du Cimetiere de Melegnano – le lendemain du Combat* [*View of the Cemetery at Melegnano – the Aftermath of Combat*], attributed to Claude-Marie Ferrier, 1859. Stereograph. Courtesy of vintagephotosjohnson.com.

first-hand. It also brought new subject matter into the homes and hands of spectators. Pictures from places and cultures all over the world were being circulated and displayed, opening people's eyes and minds to different possibilities and ways of life. Furthermore, the freezing action of the camera enabled viewing aspects of the physical world that were otherwise impossible to see with the unaided human eye. Photographs encouraged the lingering of ones'

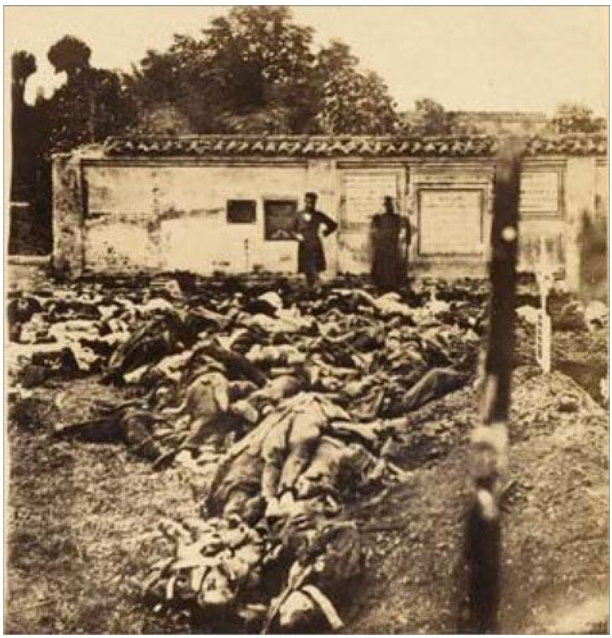


Figure 7. Detail of *702. Vue du Cimetiere de Melegnano – le lendemain du Combat*, attributed to Claude-Marie Ferrier, 1859. Courtesy of vintagephotosjohnson.com.

gaze on a scene, such as the young, dead soldiers in Melegnano, making a moment that had long since passed continue to exist and take on multivariate meaning. With the introduction of the camera into innumerable social arenas – including its presence on battlefields – the opportunities and limits of the medium contributed in unanticipated ways to a restructuring of sentiments.[78]

According to the cultural theorist Raymond Williams, "structures of feeling" refer, in essence, to the "ways of thinking and feeling" of a particular cultural context at a specific point in time.[79] As a form of communication that "outlives its bearers", it is in art that ideas and sentiments – the "actual living sense, the deep community" – of the past are contained and made accessible, which is of benefit to those working with history.[80] Art is also where deviations and disruptions in norms of sentiment initially appear. It is through artistic practice or engagement with artistic works that a

78 Raymond Williams, *The Long Revolution*, Encore Edition, Broadview Press, Peterborough, 1961.
79 *Ibid.*, p. 64.
80 *Ibid.*, p. 65. Art historian Peter Burke refers, similarly, to the distortions or artistic license of artistic creations themselves, which are of prime interest to historians; contained within them are the beliefs, ideologies and perceptions of an era. Burke goes on to explain that images can be understood as historical agents, in the way they influence perceptions of historical events. See Peter Burke, *Eyewitnessing: The Uses of Images as Historical Evidence*, University of Chicago Press, Chicago, IL, 2001, pp. 30, 145.

> new generation responds in its own ways to the unique world it is inheriting, taking up many continuities, that can be traced, and reproducing many aspects of the organization, which can be separately described, yet feeling its whole life in certain ways differently, and shaping its creative response into a new structure of feeling.[81]

It is in this sense that photography opens the possibility for new feelings to emerge and for sentiments to be restructured within a broader community. As such, it is in this way that photographs and the narratives associated with them can contribute to sentimental education towards a more caring and equitable world.

The photograph that Holmes referred to could not be circulated en masse at the time (half-tone presses still being a couple of decades away). The graphic language employed by Holmes to describe the picture reflected Dunant's and was an effort to replicate the eyewitness presence that photography seemed to offer. Likewise, such language was used by photography enthusiasts who followed the efforts of their colleagues through articles in popular newspapers and technical journals.[82] Though created with different purposes in mind, these writings yet add to the examples of photography's affective impact. In the months leading up to the Battle of Solferino, the *Photographic News* included the statement: "The example given by Mr. Fenton in the Crimea will not, therefore, want imitators in Italy."[83] Imitators, however, they got.

The reports from another photographer present at a battle that preceded Solferino provide additional context of a sentimental sort to Couppier's images. Known only as J. L., this photographer described himself as a British tourist who diverted his plans to take landscape photographs in the mountains of Switzerland and Italy when he heard of the impending battle at Palestro, which took place on 30–31 May 1859. By chance rather than design, he became a *de facto* correspondent for the *Photographic News*, a journal that predominantly featured articles on the technical aspects of the medium.

Over the course of three lengthy accounts, J. L. tells of the technical challenges of taking photographs during the conflict, including the intrusion of a "stupid Piedmontese soldier" who ruined two of his five exposures.[84] In fact, none of his exposures appear to have survived, but J. L. did provide a helpful description:

> I will send you proofs of these [surviving exposures] as soon as I have an opportunity of printing some. They will not be quite like what I hoped to send you. You will see many dead bodies scattered about among the trees, and many lying side by side ready to be thrown into the hole in which they

81 R. Williams, above note 80, p. 65.

82 The half-tone press, which would facilitate the reproduction of photographs onto paper, was still several decades away; Fenton's photographs were circulated through exhibitions and lantern lectures, and translated into woodblock prints for use in newspapers, a process which further editorialized and sanitized his pictures.

83 "Miscellaneous: Photographic Incidents", *Photographic News*, Vol. 2, No. 37, 20 May 1859, p. 129.

84 J. L., "Photography at the Seat of War", *Photographic News*, Vol. 2, No. 44, 8 July 1859, p. 208.

> will be interred as soon as it has been dug, but no bodies of men in actual conflict; I felt it would be absolutely impossible to get near enough to pitch my camera, though I was myself able to see the fight distinctly from beginning to end.[85]

Couppier's stereographs depict almost the same scene that J. L. described, with the two corroborating each other's pictures. J. L. continued to employ a photographic language to describe more of what he saw and what he could not photograph. Aside from mechanical matters, J. L. also revealed – reminiscent of Dunant's account in *Solferino* – the way in which witnessing the battle affected his own views on warfare. Watching the battle from atop a tree in the company of the priest who informed him of the fighting, J. L.'s account includes his transition from being someone who admired war to now being stunned by it:

> There is something wonderfully impressive in the sound of the marching of a body of armed men …. I afterwards saw bodies of men moving towards each other to engage in actual combat without any similar feeling …. [T]o describe what took place over the whole scene of the fighting is out of my power.[86]

Perhaps J. L. is bowdlerizing, or perhaps he was really taken aback. He describes a transition in his own thought, from seeing battles as conceivably heroic and exciting to warfare being a human tragedy. A little later, he goes on to graphically describe his experience in the moments immediately following the end of the fighting:

> Their [the wounded's] groans could have directed us to where they were lying, even if we I [*sic*] [had] not been able to see them. You can form no concept of the sickening sensation I felt when I found myself in the midst of pools of blood, which splashed about at every step spreading a sickening smell in the atmosphere. The bodies of the slain were lying pell-mell among the wounded, very few of whom were able to withdraw themselves from the horrible contact. We moved each in succession, and laid them gently on their backs – the dead, dying, and wounded side by side.[87]

Demonstrating a sort of kinship with Dunant's later text, J. L. was writing for the benefit of "you at home [who] have not a thorough conception of the horrors of warfare, or of the injustice and cruelty it involves".[88] Consistent with Dunant and to a great extent the combat photography of the era, J. L.'s accounts focus on the impact of warfare on the common soldier. J. L.'s description of dead and injured soldiers "lying pell-mell" is an image quite in contrast to conventional war art that, when showing wounded combatants, would more likely have focused on officers and generals surrounded by their supporters. Nineteenth-century combat

85 *Ibid.*
86 *Ibid.*
87 *Ibid.*
88 J. L., above note 48, p. 183.

photography brought to the fore the ways in which fighting brutalized the bodies of the average troops.

In many ways, J. L. and Dunant are similar. They are both self-described "tourists" of war, though J. L. was actually a tourist whose travels were diverted by the exciting pull of (and perhaps the challenge of photographing) battle. They both became eyewitnesses to the aftermath of battle and the impact of warfare on human bodies, with J. L. having witnessed actual fighting. Despite J. L. having been a photographer, both took to graphic writing to describe the similar scenes they saw. J. L. and Dunant are also dissimilar from each other, however. There is no indication in the historical record to suggest that J. L.'s life purpose emerged from this event, as it did for Dunant. While J. L. appears to have had his vision altered, he makes no call to intervene, as did Dunant, on the suffering he witnessed. He does not appear to have been inspired like Dunant to get immersed in caring for the wounded, nor to redirect his purpose to that of ameliorating battlefield suffering. J. L. is also very different to Fenton – though an early photographer of combat, his images (including his descriptions of them) have had far less reach than Fenton's. Yet, J. L.'s account provides an example of the role of aesthetic encounters in inducing passionate responses in himself. We get a strong sense that he is moved to feel and think differently about warfare, seeing it as less heroic than initially expected. As for his audience, he hardly provides guidance on emotional development or nurturance of a sentimental community, as Dunant and even Holmes provide in their accounts. J. L.'s responses, however, offer by their very existence an opportunity to view soldiers as lamentable, which has the potential of supporting the growth of a more caring community of humans.

Correspondences such as J. L.'s, about the battles that were part of the Franco-Austrian War, were not restricted to niche technical journals or to Europe.[89] With the telegraph and steam transport, news of the war travelled across to North America as well. Coincidentally, the first battle that the recently established *New York Times* newspaper covered was the same that inspired Henry Dunant.[90] The Crimean War was geographically too far away and did not have as much relevance to Americans as did the wars of Italian liberation. So in the spring of 1859, accompanied by two colleagues, the *New York Times*'s editor and co-founder, Henry Jarvis Raymond, went to Italy, driven by "that inexplicable perversity of human nature which pushes on towards scenes of carnage".[91] Like Dunant in *A Memory of Solferino*, Raymond wrote in a manner meant "to dissolve authorial distance and thereby enable readers to see in their minds what was not before their eyes".[92] Unlike Dunant, who arrived later, Raymond was

89 The *Journal de Genève* (1826–1991) did not reproduce prints to illustrate its newspaper; it did, however, reproduce photographic language. It was common practice to transcribe, with the use of the telegraph, accounts from other foreign newspapers. Dunant would have been reading the graphic accounts reproduced in this journal and would have been getting additional information about distant events from people within his transnational social network.

90 The *New York Times* was founded in 1851.

91 Jonathan Marwil, "The New York Times Goes to War", *History Today*, Vol. 55, No. 6, 2005, p. 47.

there at midday on 24 June, during the height of the battle. While J. L. had a clear vantage point from his perch in a tree, Raymond was unable to see much through the smoke from his position on a distant hillside. Nevertheless, from the periphery, he saw scenes of carnage echoing those that Dunant and J. L. would later write about.

Raymond did his best to describe these scenes for his readers, who would have been unfamiliar with such mass devastation. Similar to Dunant's "lamentable pictures", he wrote of musket and sabre wounds, jaws cut away and gaping holes in men's bodies, and tried to generate in his readers a sense of the scale of the assault by inviting them to imagine carts full of bloody soldiers in front of New York City Hall.[93] Again similar to Dunant, Raymond's writing, with its focus on suffering, takes on an appearance of nurturing humanitarian sentiment. However, unlike Dunant, but in the vein of news reporting, little guidance is provided on how to mobilize that sentiment. The graphic language that Raymond used might not have been enough for readers with no reference pictures in their personal or collective imaginations to draw from in order to visualize the grotesque scenes emerging from this battle.[94] In a few short years, however, there would be a massive collection of war photographs made during the American Civil War that would be have lasting impact.

The photographs that Mathew Brady and his company made during the American Civil War did include the type of photographs that Fenton was taking and circulating;[95] they also included more harrowing images like those from Solferino. However, Brady and his colleagues' pictures differed from those of the other two battles in a critical way. Brady and his team had secured access to the battlefields and taken photographs before the bodies of the dead soldiers were removed (see Figure 8). The common practice of staging living combatants in their camps or rearranging items in a "still life" scene was extended to the repositioning of corpses for added effect. Today's preoccupation with the fidelity and integrity of images was not a concern in 1863 – truthfulness was more important than accuracy, and the truth in these photographs was that the men and boys depicted in them were damaged and dead as a result of the war.

Fenton and his peers' cameras may not have been used with the intent of changing public perception, particularly in terms of transforming warfare into an anti-heroic event. Likewise, they may not have been popularly interpreted that way, but inherent in the photographs are changing attitudes towards the value of

92 *Ibid.*, p. 48.

93 *Ibid.*, p. 52.

94 In forming mental images, Dunant's European readers would also have been referencing the many illustrated newspapers and a visual arts tradition that made use of photography (including several popular panorama paintings of the Battle of Sevastopol, such as Jean-Charles Langlois' two *Bataille de Sébastopol* paintings of 1855 and 1856), or even their own first-hand memories of recent battles in Europe. See John Hannavy, "Crimea in the Round", *History Today*, Vol. 54, No. 9, 2004.

95 Mathew Brady was a commercial photographer in New York at the time that the Civil War began. He hired upwards of twenty photographers, or camera operators, to make visual records of various aspects of the war's battles. Among the most well known of these were Alexander Gardner and Timothy H. O'Sullivan.

Figure 8. *Incidents of the War: A Harvest of Death*, by Timothy H. O'Sullivan, 1863. Library of Congress, LC-B8184-7964-A.

human lives and the question of whose lives matter.[96] The ways in which the photographers operated at Crimea, around Solferino and in the American Civil War amount to a pattern. Included in each collection were pictures of soldiers and officers, always carefully posed. Not all were for propaganda, as is how Fenton's pictures have predominantly been viewed; some were mementos commissioned by individual soldiers.[97] Each collection of photographs from these three battles also included images of the sites where the battles took place. Although none of the photographers included images of the fighting itself, for obvious technical reasons, all had created what amounted to before-and-after pictures. All had created narratives in which the vital, the strong, the living combatants were cut down, killed and ultimately extinguished from the scenes of battle. Comparing all three sets, Fenton's barren battlefield landscapes may have presented a more allegorical narrative. Couppier's Solferino pictures were certainly more harrowing, but anonymous. And those from the Civil War made the story all the more graphic, with human figures that were recognizable and identifiable, as this quote from an 1863 exhibition review that appeared in the *New York Times* reveals:

96 P. Burke, above note 82.

97 The *Photographic News* reported: "[W]e know that most of the subaltern officers figure largely in the collections of portraits which have been made. It is the fashion to have one's portrait taken in camp." See "Miscellaneous", above note 85, p. 129.

> [W]e could scarce choose to be in the gallery when one of the women bending over them should recognize a husband, a son or a brother in the still, lifeless lines of bodies that lie ready for the gaping trenches.[98]

With each conflict, the intimacy between the spectator and the dead increased, yet the soldiers continued to remain nameless masses. The indiscriminate suffering of soldiers and the gruesomeness of war were made palpable with the aid of the camera. Photography's democratic treatment of soldiers and generals also invited consideration of more uniform and egalitarian medical treatment across all ranks.

A result of being at the forefront of "convey[ing] the human face of war", nineteenth-century combat photography made it possible that soldiers were increasingly being "regarded as fellow citizens – sons, brothers, fathers".[99] Riding a rising wave that seemed to accompany the creation and circulation – through exhibitions, textual descriptions or prints – of combat photographs, passionate and empathetic responses made way for an organized political response. People were feeling differently towards warfare. Such emotions benefited from the concrete actions that Dunant appealed for; he was able to give that emotion an action. It was within a landscape in which soldiers were included in the broadening terms of humans worthy of attention that Dunant's visions in *A Memory of Solferino* could come into focus and spark an international movement.[100]

Conclusion

Despite the scale and number of conflicts being a sign of apparent moral decay, the mid-nineteenth-century moment discussed in this article is considered one of humanitarian progress. Dunant's skill in uniting people to a cause can be credited for his success, but the spark emanating from his book may not have ignited had it not been for the combustible material, so to speak, that circulated at unprecedented speed and in unsurpassed detail with the help of the camera. *A*

98 Quoted in Susan Moeller, "Photography, Civil War", available at: www.encyclopedia.com/defense/energy-government-and-defense-magazines/photography-civil-war. Cara Finnegan provides invaluable insight into the rising importance of photography in the everyday lives of a growing audience of visual spectators in this mid-nineteenth-century moment, which she characterizes as "a period when photography became a dominant medium of cultural life". She also points out the difficulty of locating reactions and commentary from this audience, as these were not recorded or valued as historically relevant at the time. See Cara A. Finnegan, *Making Photography Matter: A Viewer's History from the Civil War to the Great Depression*, University of Illinois Press, Urbana, IL, 2017.

99 J. Marwil, above note 2, p. 35.

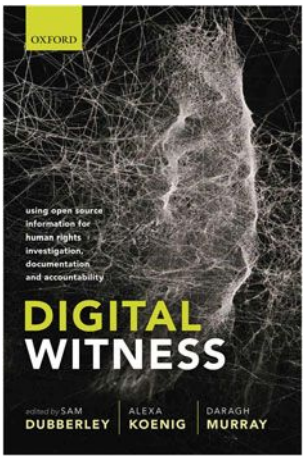
100 From its earliest years, the Red Cross movement mobilized photography, but it was not an adopter of atrocity photographs and did not participate in what would later be characterized as a trade in the visuals of suffering in the way other aid groups or industries might. It has been argued that institutions like the Red Cross did not have to display such images as supporters could "imagine" them when referenced circuitously. Indeed, images of suffering, be they photographs or prints of another sort, were prominent in the surrounding contemporary context, including evangelical pamphlets and lantern lectures, the commercial "yellow press", and pulp fiction. As noted by Rozario, "phantom spectacles of suffering [can be] conjured up imaginatively even as they are renounced rhetorically". K. Rozario, above note 22, p. 443.

Memory of Solferino appeared at a time in which new forms of international travel and communication were accelerating changes in sentiment and in social and political culture. Prime among the new technologies was the photographic camera. The popularity of photography from the late 1830s onward brought with it new capacities for viewing – in essence, heralding a new visual culture. The technology of the camera extended the gaze temporally and geographically, thus enabling a reconfiguration of existing ways of seeing and feeling. The result was an ability to draw attention to previously overlooked and vulnerable bodies – in Dunant's case, that of the common soldier. Operating within his social and political milieux, Dunant found support from influential individuals who were receptive to the ideas at the root of his appeal. His may have been a modest call by some standards – an appeal to care for wounded soldiers, rather than to bring an end to warfare altogether – but the result was the bringing into focus of a new form of humanitarianism – third-party, neutral and impartial – that is the basis of modern humanitarian action today. For humanitarian actors and agencies working with visual narratives in the modern world, this historical example offers a valuable perspective on the role of photography and spectators' engagement with it in expanding a shared vision of lives worth caring for.

International Review of the Red Cross (2020), **102** (913), 445–449.
Digital technologies and war
doi:10.1017/S1816383120000405

IRRC

BOOK REVIEW



Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability

Edited by Sam Dubberley, Alexa Koenig and Daragh Murray*

Book review by Emma Irving, consultant and academic in the field of technology and international justice specializing in international criminal law and international human rights law.[1]

To gain a full picture of the field of open-source human rights investigation, documentation and accountability, it is necessary to leave the comfort of one's own professional or academic discipline and to delve into the whole range of specialties that are called into action when human rights work is undertaken online. The book *Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability*, edited by Sam Dubberley, Alexa Koenig and Daragh Murray, allows readers to do precisely this. The first of its kind, it is a long-awaited resource that brings together the many elements that make up this field, and is essential reading for anyone interested in open-source

* Published by Oxford University Press, Oxford, 2020.

human rights work. From readers who operate daily in this space to those pursuing a nascent curiosity, *Digital Witness* will prove valuable across the board.

Digital Witness brings together practitioners and academics who specialize in different areas of open-source human rights work. Their contributions are divided across the four sections of the book. The first section sets out the contextual landscape of open source human rights work with chapters on the theory, history and sociology of the field.[2] The second section populates this landscape with technical know-how, offering chapters dedicated to particular practical skill sets – from how to find open-source material relevant to human rights work, to how to preserve and archive it.[3] The third section could be termed the "conscience of the book", as it deals with what open-source investigators owe to others and to themselves. From the perspectives of ethics and security, the chapters in this section discuss the risks that open-source human rights work poses to those about whom information is collected and processed, and how these risks can be mitigated.[4] Importantly, the book includes a discussion of the risks facing open-source investigators themselves, both psychologically and in terms of security. In the fourth and final section, the two concluding chapters turn to the future of open-source human rights work, setting out the groundwork for how the field can move forward in a constructive way.[5]

In addition to being a rich source of information in a fragmented field, perhaps *Digital Witness*'s biggest success is the way in which it bridges the gap between the practical and the academic, between the personal and the technical, and between different disciplinary fields. The range of knowledge and specialities required to operate effectively in this field has long been a challenge for those working within it, who are required to assume various roles, sometimes simultaneously.

Beginning with the practical and the academic, this traditionally entrenched gap is bridged in the book by the bringing together of information types that would normally be placed at opposite ends of a library. Take, for example, the explanation of how search syntax works in Chapter 6.[6] In his chapter on how to discover relevant open-source material online, Paul Myers takes the reader through the ways in which the results given by a search engine can be narrowed down and tailored using particular search terms (also known as Boolean operators). In an academic book, one might expect such an explanation to foreshadow a theoretical or academic argument (about the overwhelming nature of the online information environment, perhaps). However, here the purpose of the explanation is simpler: the aim is to instruct the reader in how to

1 Dr Emma Irving has held the position of Assistant Professor of Public International Law at the Grotius Centre for International Legal Studies at Leiden Law School, and has degrees from the Universities of Amsterdam (PhD), Leiden (LLM) and Cambridge (MA).
2 *Digital Witness*, Part I, Chaps 1–5.
3 *Ibid.*, Part II, Chaps 6–10.
4 *Ibid.*, Part III, Chaps 11–13.
5 *Ibid.*, Part IV, Chaps 14, 15.
6 Paul Meyers, "How to Conduct Discovery Using Open Source Methods", in *Digital Witness*, pp. 109–111.

conduct a search. Sitting comfortably alongside practical chapters of this type are the more traditionally academic chapters. Chapter 1, for example, tracks the history of using open-source investigations for human rights reporting, grouping the developments together into different categories, and using case studies to support the authors' description of events.[7] A natural but interesting consequence of this bridging of the academic and the practical is that the edited volume as a whole draws on a great diversity of sources. In *Digital Witness*, a reference to an academic article is as much at home as a screenshot of a Google search or a reference to an interview with a practitioner.

Turning to the bridge between the personal and the technical, this is achieved – at least in part – through the amalgam of writing styles featured in *Digital Witness*. As the different chapters oscillate between narrative, theoretical and instructional styles, the book does justice to the human side of open-source human rights work while still engaging with the technical necessities. In Chapter 2, for example, Alexa Koenig adopts a storytelling style to take the reader on the journey of the evolution of open-source human rights work in legal accountability.[8] Told through the stories of those working in the field, this aptly named "Modern Social History" chapter tracks developments in international and domestic accountability proceedings. The reader comes to know the names of these individuals, along with their struggles and successes. This narrative style contrasts with the theory-rich Chapter 4, in which Ella McPherson, Isabel Guenette Thornton and Matt Mahmoudi frame the rise of open-source investigations as a "knowledge controversy" that is shifting the standard practices of human rights fact-finding.[9] The technical chapters in Part II of the book are different in style once again, with the adoption of an almost step-by-step guide approach. Even within these instructive chapters, the human side of the field is kept in focus; in Chapter 10, for instance, Micah Farfour illustrates the value of satellite imagery analysis for human rights investigations by taking the reader through a number of case studies.[10] Alongside discussion of spatial and spectral resolution are examples of human rights abuses affecting particular communities, with descriptions of how these abuses are visible from above.

Last but not least, the driving force behind the two bridges just described is the book's multidisciplinarity. A legal scholar would easily recognize Lindsay Freeman's overview of the use of open-source evidence in international criminal prosecutions in Chapter 3; an ethics scholar would be familiar with Zara Rahman and Gabriela Ivens' discussion in Chapter 11 of the ethics involved in open-source human rights work. In order for the book's bridging of different

7 Christopher Koettl, Daragh Murray and Sam Dubberley, "Open Source Investigation for Human Rights Reporting: A Brief History", in *Digital Witness*, pp. 12–32.
8 Alexa Koenig, "Open Source Evidence and Human Rights Cases: A Modern Social History", in *Digital Witness*, pp. 32–48.
9 Ella McPherson, Isabel Guenette and Matt Mahmoudi, "Open Source Investigations and the Technology Driven Knowledge Controversy in Human Rights Fact Finding", in *Digital Witness*, pp. 68–87.
10 Micah Farfour, "The Role and Use of Satellite Imagery for Human Rights Investigations", in *Digital Witness*, pp. 228–249.

disciplines to work, the traditional divides between the academic, practical, personal and technical must necessarily be set aside. The success of doing so in an effective way rests on how accessible information is to those approaching it from another field. While some parts of the text are (perhaps inevitably) dense and complex, the book achieves the accessibility and readability required to reach a broad audience.

As with any work that crosses disciplinary boundaries, readers seeking specialized information from a particular disciplinary perspective will not necessarily find the depth they need for an expert understanding. A lawyer, for example, may wish for greater attention to be paid to the legal dimensions of the field, and for legal discussion such as that featured in Chapter 3 to be found extensively elsewhere in the book too. This, however, is not a weakness of the book, but a consequence of its positioning within the field – it should be seen as a central point in the information environment from which specialized strands of knowledge can grow, rather than the pinnacle of the information pyramid.

An important caveat present throughout the book is that open-source techniques are not the silver bullet that will solve all challenges in the investigation and documentation of human rights abuses, nor in holding perpetrators accountable for such abuses. Several chapters mention that open-source information can be valuable to the extent that it corroborates other forms of evidence, such as witness testimonies, and stresses that open-source information must always be corroborated. The introduction to the book relates what could be called the quintessential example of the importance of open-source material, as it takes the reader through an analysis of a social media video showing extrajudicial killings by the Cameroonian military.[11] Even here, where there is a highly compelling video, the writers make clear the importance of speaking to witnesses and of travelling to the site of the killings in order to be sure of their findings. Chapter 14 is largely dedicated to this message, as it describes open-source information as one piece of the investigation puzzle: to leave it out risks an incomplete investigation, but to rely on it too much risks a biased one.[12]

An unfortunate drawback of writing in any technology-related field is that the pace of developments quickly renders the material outdated. Those reading *Digital Witness* in 2020 will experience it at its most relevant, whereas readers who come to it later may need to cross-reference with newer writings to ensure their knowledge is current. The book format is particularly susceptible to this problem, and edited volumes perhaps even more so. The time required for the editors to coordinate a diverse range of authors into updating and resubmitting their chapters for a new edition, combined with the processing time of the publisher, makes for a cumbersome and slow task. This challenge will affect some

11 Sam Dubberley, Alexa Koenig and Daragh Murray, "Introduction: The Emergence of Digital Witnesses", in *Digital Witness*, p. 3.

12 Fred Abrahams and Daragh Murray, "Open Source Information: Part of the Puzzle", in *Digital Witness*, pp. 317–331.

parts of the book more than others. McPherson, Thornton and Mahmoudi's theory-based chapter on how human rights work is living through a knowledge controversy will likely remain relevant for some time to come, as will Alexa Koenig and Lindsay Freeman's suggested guiding principles for the future of open-source human rights work in Chapter 15.[13] This contrasts with the practical chapters in Part II of the book, a number of which reference tools and resources that may come and go with time.

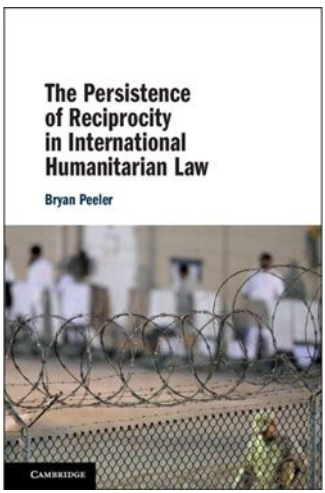
Digital Witness fills an important gap in the information ecosystem surrounding open-source human rights work. It renders the field accessible in a way that it has not been to date, by opening up all aspects of it – the academic, the practical, the personal, the technical – to actors who may otherwise have only been exposed to one or two of them. It is to be hoped that it is the first of many editions to come, and that it will act as a springboard for further research in the field.

13 Alexa Koenig and Lindsay Freeman, "Open Source Investigations for Legal Accountability: Challenges and Best Practices", in *Digital Witness*, pp. 331–323.

International Review of the Red Cross (2020), **102** (913), 451–455.
Digital technologies and war
doi:10.1017/S1816383120000442

IRRC

BOOK REVIEW



The Persistence of Reciprocity in International Humanitarian Law

Bryan Peeler*

Book review by Matthias Vanhullebusch, Chenxing Associate Professor at the KoGuan School of Law, Shanghai Jiao Tong University.[1]

Compliance has been one of the most perennial questions of international humanitarian law (IHL), and theoretical and empirical studies have sought to comprehend how IHL functions respectively in the legal and policy debates and on the battlefield. Falling into the former category, Bryan Peeler's recent book on *The Persistence of Reciprocity in International Humanitarian Law* develops a new theory on compliance from the perspective of the State in general and, accordingly, illustrates this exclusively with reference to US State practice. Peeler contends that both the development and the interpretation of IHL are inherently informed by a logic of consequences: that is to say, respect for IHL is driven by reciprocal commitments on behalf of States in times of war as well as in peacetime.[2] His thesis draws from realist[3] and neoliberal institutionalist[4] readings of international relations. As Peeler applies these to the study of compliance with IHL obligations, he juxtaposes them with the "humanization of IHL" thesis. This latter perspective instead proclaims a logic of appropriateness when explaining normative conduct on the international plane, whereby the decision-making

* Published by Cambridge University Press, Cambridge, 2019.

process privileges a perception "of normatively correct behaviour" above a mere cost-benefit analysis on such conduct.[5] Such logic is advanced by various other strands of international relations, including liberalism[6] and constructivism,[7] as well as international law schools of thought, including international legal process. Following the theory of the realist Keohane, who distinguishes between specific[8] and diffuse reciprocity[9] – embodying the logic of consequences – Peeler's new theory on compliance is less concerned with diffuse reciprocity; instead, it further refines specific reciprocity and, accordingly, differentiates between legal and strategic reciprocity in his study on compliance with IHL obligations (Chapter 2).[10]

In this respect, regarding legal reciprocity, Peeler argues that, from its inception onwards, IHL has incorporated reciprocity into its language. He finds support for this in the negotiating history of the Geneva Conventions and their Additional Protocols (Chapter 3),[11] which stipulate conditions under which belligerent reprisals are possible under the rules governing the conduct of hostilities and afford standards of protection to those who comply with IHL obligations.[12] Strategic reciprocity, on the other hand, has – as a matter of policy – informed decision-making processes regardless of the (il)legality of a country's conduct on the battlefield. Here, Peeler tests his model with reference to the United States' usage of positive and negative strategic reciprocity in respect of its treatment of enemy fighters respectively during the Vietnam War (Chapter 4)[13] and the Global War on Terror, including aspects of the use of

1 Dr Matthias Vanhullebusch (PhD, SOAS) is a Visiting Professor on various training programmes organized by the International Committee of the Red Cross (ICRC) across Asia. He has been a Visiting Scholar at the Geneva Academy of International Humanitarian Law and Human Rights, the Oxford Institute for Ethics, Law and Armed Conflict, and the Asia-Pacific Centre for Military Law.

2 *The Persistence of Reciprocity in International Humanitarian Law*, pp. 39, 42.

3 According to Peeler, realist theories "deny that IHL can have … an independent constraining effect on the behaviour of states". *Ibid*., p. 23.

4 In order to ensure compliance with IHL obligations, Peeler argues, "neoliberal institutionalism suggests that states include measures implementing the strategy of TFT [tit-for-tat] into the law". *Ibid*., p. 23.

5 *Ibid*., p. 40.

6 Peeler considers that liberal international relations theories responsibilize the agent to promote "a culture of IHL compliance". *Ibid*., p. 46.

7 Constructivists, from the perspective of Peeler, believe that compliance with international norms is possible given States' capacity towards socialization – or internalization – of international law. *Ibid*., p. 48.

8 According to Keohane, specific reciprocity occurs when "specified partners exchange items of equivalent value in a strictly delimitated sequence. If any obligations exist, they are clearly specified in terms of rights and duties of particular actors." See Robert Keohane, "Reciprocity in International Relations", *International Organization*, Vol. 40, No. 1, 1989, p. 4.

9 What distinguishes specific from diffuse reciprocity is, according to Keohane, the involvement of multiple actors concerned with the behaviour of other actors with whom they do not entertain a specific relationship, which affects in turn how they will perform certain obligations under international law vis-à-vis each other in different situations and across different timeframes. Here, according to Keohane, "the definition of equivalence is less precise, one's partners may be viewed as a group rather than as particular actors, and the sequence of events is less narrowly bounded. Obligations are important." *Ibid*., p. 4.

10 *The Persistence of Reciprocity in International Humanitarian Law*, pp. 12–58.

11 *Ibid*., pp. 59–94.

12 *Ibid*., p. 93.

13 *Ibid*., pp. 95–127.

enhanced interrogation techniques against such persons (Chapter 5).[14] Like Hart, he argues that the application of primary obligations of the fighting parties under IHL is inherently conditioned by the secondary rules on reciprocity.[15] Peeler richly explores this understanding through US case studies in which a multiplicity of contested arguments across the spectrum of authorities was present throughout the planning and execution of the Vietnam War and the Global War on Terror.

By drawing from archival resources, memos and Supreme Court judgments, Peeler offers a unique insight into those US legal and policy debates pertaining to the treatment and detention of opponents in the Vietnam War and the Global War on Terror. Interviews which he conducted in February and March 2014 with key (legal) advisers to the US State Department, such as John Bellinger and William Taft, and officials at the US Defense Department add another layer to the rich exposé of conflicting interpretations of the rules governing the treatment, detention and interrogation of enemy persons by the United States. Indeed, for (legal) historians and political scientists, this is definitely one of the most engaging and informative accounts of the different voices within the relevant US administrations – White House officials, Judges Advocate General, members of the Houses, etc. – in crucial decision-making processes on these matters. The logic of consequences, according to Peeler, was ultimately followed by those in power and the White House officials who were capable throughout these decision-making processes of silencing their domestic (political) opponents. The latter favoured a logic of appropriateness for the sake of the interests of the United States' national image and the security of its soldiers in those conflicts and beyond.[16] Peeler's new theory sheds a different light on these case studies compared to the previous work by Mark Osiel in *The End of Reciprocity*. Contrary to Osiel, who argues that "[i]t is immaterial, when one's own violations are judged, that one's military opponent committed the same breaches",[17] Peeler finds that the United States' decision-making processes explicitly engage with different discourses on reciprocity in order to justify certain strategic options regarding the treatment of enemy fighters in detention and their respective interrogation.

Peeler argues that during the Vietnam War, despite the ambiguities on the application of Geneva Convention III relative to the Treatment of Prisoners of War (GC III), the Johnson administration advanced a positive strategic reciprocity strategy. In this respect, by providing humane treatment to captured enemy fighters, the United States expected that American prisoners of war would equally deserve decent treatment when falling into the hands of the Democratic Republic

14 *Ibid*., pp. 128–168.

15 Peeler refers to Hart's distinction of primary and secondary rules, explaining that "the primary rules of a legal system are those rules that either forbid or require certain actions and generate duties or obligations. The secondary rules, on the other hand, are rules that describe the manner in which we recognize, change and adjudicate violations of primary rules." *Ibid*., p. 42. See also Herbert L. A. Hart, *The Concept of Law*, Oxford University Press, Oxford, 1961, p. 87.

16 *The Persistence of Reciprocity in International Humanitarian Law*, pp. 145–146.

17 Mark Osiel, *The End of Reciprocity*, Cambridge University Press, Cambridge, 2009, p. 73.

of Vietnam and the Viet Cong – despite the latter's defection from IHL obligations. Perennial defectors, however, would not enjoy equal decent treatment.[18] Positive reciprocity was also advanced through the United States' investigation into its own commission of war crimes in order to bring the communists to the negotiating table.[19] Conversely, the Bush administration introduced a departure from the specific positive reciprocity and general focus on IHL compliance on behalf of the US military that had existed since the Vietnam War and the conflict in Yugoslavia.[20] The perspectives of the Judges Advocate General were ignored throughout the decision-making processes on the drafting of the Military Commissions Act (2006/09)[21] and the standards of conduct regarding enhanced interrogation techniques that reflected the White House's prevailing negative strategic reciprocity argument.[22]

According to Peeler, this sovereigntist move in the Global War on Terror predates the early 1980s, when it was justified for the US to opt out of international law in general and GC III in particular if done in the national interest.[23] Therefore, the White House pursued the use of military commissions to try unlawful combatants, deny the procedural rights of detainees under the Uniform Code of Military Justice, and inhumanely treat suspected terrorists, despite rulings of the US Supreme Court (*Hamdan v. Rumsfeld* in 2006,[24] *Boumediene v. Bush* in 2008[25]), potential adverse security threats against US interests, and the country's deteriorating international image beyond the specific conflicts at hand.[26] Moreover, the deliberate rhetoric antagonizing Islam in general and Islamic fighters in particular was put at the service of the negative strategic reciprocity argument, justifying the waiver of combatant privileges to terrorists because they did not respect the laws of war.[27] As a result, inducing compliance – pursuant to a logic of positive strategic reciprocity – was less an issue during the Global War on Terror given the rare cases of US soldiers being detained by terrorists. The CIA rendition programmes, however, have been left untouched in this study. Parallels could have been drawn to the earlier discussion on the Vietnam War, where the United States transferred Viet Cong detainees to the Army of the Republic of Vietnam, which failed to treat them humanely despite US training.[28]

While Peeler's alternative theory on compliance has offered these new insights, the reader is left wondering whether it can have a wider personal scope

18 *The Persistence of Reciprocity in International Humanitarian Law*, pp. 95, 103, 113.
19 *Ibid.*, pp. 115, 117.
20 *Ibid.*, pp. 124, 134.
21 Military Commission Act, Pub. L. 109-366, 120 Stat. 2600, 17 October 2006; Military Commission Act, Pub. L. 111-84, 123 Stat. 2574, 28 October 2009.
22 *The Persistence of Reciprocity in International Humanitarian Law*, pp. 130, 141.
23 *Ibid.*, p. 132.
24 Supreme Court of the United States, *Hamdan v. Rumsfeld*, 548 US 557 (2006), 2006.
25 Supreme Court of the United States, *Boumediene v. Bush*, 128 US 2229 (2008), 2008.
26 *The Persistence of Reciprocity in International Humanitarian Law*, p. 153.
27 *Ibid.*, p. 152. Needless to say, once the conditions for its applicability have been satisfied, IHL governs the conduct of all fighting parties irrespective of their causes of warfare.
28 *Ibid.*, pp. 110–112.

of application. That is to say, can his study on the intersection of international relations and international law, naturally focusing on the conduct of States waging war, also be applied to the normative conduct of non-State armed groups? After all, most contemporary armed conflicts – the majority of which are of a non-international character – involve non-State armed groups. Applying Peeler's theory to such conflicts would require further empirical research to gain access to the policy debates amongst those non-State armed groups, similar to the interviews that Peeler carried out to understand how the legal arguments of the White House in the respective conflicts have been construed.

Fortunately, such privileged insights into the conduct of non-State armed groups have already been well documented in psycho-sociological studies carried out by international humanitarian organizations such as the International Committee of the Red Cross (ICRC).[29] Peeler's book could have borrowed from such earlier studies in order to offer a more holistic analysis on the role of reciprocity in the normative conduct of all fighting parties in today's non-international armed conflicts. The seeds of such analysis are already present in Peeler's approach across the entire book as he verifies the impact of dehumanization discourses on the enunciation of positive and negative strategic reciprocity arguments – at least from the perspective of the US government. Humanitarian legal professionals and front-line humanitarian negotiators could have benefited even more from this valuable political science study if it had engaged with their particular legal methodologies and sources.

Irrespective of the reservations of humanitarian lawyers on the limited personal and material scope of application of Peeler's alternative theory on compliance, political scientists and historians alike will find this book very informative and innovative in its approach and presentation as it revisits the role of reciprocity in policy debates before US administrations both past and present.[30]

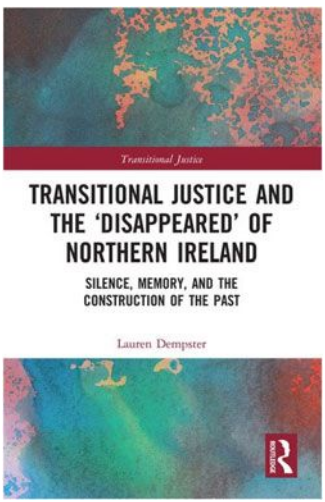
29 Daniel Muñoz-Rojas and Jean-Jacques Frésard, *The Roots of Behaviour in War: Understanding and Preventing IHL Violations*, ICRC, Geneva, 2004. This study identifies the principal cause of violations of IHL as the moral disengagement of the fighters in the course of armed conflicts. In other words, parties to the conflict justify their violations of IHL by having recourse to their superior moral and/or religious cause of warfare and by dehumanizing their opponents – both fighters and civilian populations.

30 In his concluding chapter, Peeler briefly touches on the challenges presented by the Trump administration regarding its (lack of) respect for international norms, including IHL. See *The Persistence of Reciprocity in International Humanitarian Law*, pp. 176–184.

International Review of the Red Cross (2020), **102** (913), 457–461.
Digital technologies and war
doi:10.1017/S1816383120000430

IRRC

LIBRARIAN'S PICK



Transitional Justice and the "Disappeared" of Northern Ireland: Silence, Memory, and the Construction of the Past

Lauren Dempster*

Book review by Charlotte Mohr, ICRC Reference Librarian for the collections on the ICRC's history and activities.

In Northern Ireland, as in many other post-conflict contexts, the "disappeared" have become a central issue in the discourse on conflict legacy and reconciliation. Between 1972 and 1985, sixteen people are presumed to have been secretly executed and buried by Republican paramilitary groups.[1] A most insidious form of conflict violence, disappearances are the cause of long-term psychological and social harm. In a post-conflict setting, they represent both an open wound and a remnant of war – a thorny issue for transitional justice to address, at the intersection of the personal and political. In 1998, the signature of the Good Friday Agreement brought an end to "the Troubles", the three-decade period of

* Published by Routledge, New York, 2019. Dr Lauren Dempster is a lecturer in the School of Law of Queen's University Belfast. She is currently a co-investigator on a pilot project exploring the representation of victims and victimhood at sites of so-called "dark tourism".

ICRC Library

The "Librarian's Pick" is a new section of the International Review of the Red Cross, replacing "New Publications in International Humanitarian Law and on the International Committee of the Red Cross". In this section, one of the International Committee of the Red Cross's (ICRC) librarians picks and writes about their favourite new book relating to international humanitarian law, policy or action, which they recommend to the readers of the journal.

The ICRC Library welcomes researchers interested in international humanitarian law (IHL) and the institution's work throughout the years. Its online catalogue is the gateway to the most recent scholarship on the subject, documents of diplomatic and international conferences, all ICRC publications, rare documents published between the founding of the ICRC and the end of the First World War, and a unique collection of military manuals. The Library Team also publishes research guides in order to help researchers access the full texts of the most relevant and reliable sources in the field of IHL and the ICRC, as well as a comprehensive IHL Bibliography, with three issues every year.[2]

conflict between nationalists and unionists in Northern Ireland. No provision for an overarching truth-finding or reconciliation commission was included. Twenty-two years later, a consensus has yet to be reached on how to address the legacy of those years of violence.[3] What has repeatedly been described as a piecemeal approach to transitional justice in Northern Ireland has, however, given birth to the Independent Commission for the Location of Victims' Remains (ICLVR). This independent, non-judicial investigative body for the disappeared was established via an intergovernmental treaty between the United Kingdom and Ireland in 1999. To date, the remains of thirteen out of the sixteen victims have been recovered.

Twenty years after the creation of the ICLVR, Lauren Dempster provides a first extensive qualitative study of the Commission's history and impact in her book

1 The Provisional Irish Republican Army (IRA) has admitted responsibility for the disappearance of thirteen of the sixteen victims, mostly in one main statement dated 1999, and the Irish National Liberation Army for one additional victim. The Republican discourse on the issue has mainly remained consistent, if less vehement with time: the disappearances have been presented as the result of internal disciplinary measures carried out against informants or individuals accused of stealing IRA weapons.

2 The online catalogue is available at: https://library.icrc.org/library/. For the most recent publications, see: https://library.icrc.org/library/search/date. For more information on the research guides, see: blogs.icrc.org/cross-files/category/research-guide. To subscribe to the IHL Bibliography, email library@icrc.org with "IHL Bibliography subscription" in the subject line.

3 The 2014 Stormont House Agreement between the British and Irish governments included a series of measures to deal with legacy issues of the Northern Irish conflict: the creation of a Historical Investigation Unit overseen by the Northern Ireland Policing Board, potentially opening the door to prosecutions for crimes committed during the Troubles; an Independent Commission on Information Retrieval; an Implementation and Reconciliation Group; and an Oral History Archive project. A lack of political consensus has delayed the implementation of these measures of the Agreement since its adoption.

Transitional Justice and the "Disappeared" of Northern Ireland. In its preface, the book purports to "[explore] the response to the 'disappearances' that were perpetrated during the Northern Ireland conflict through a transitional justice lens". Arguably, it is also an exploration of concepts central to the field of transitional justice – such as victimhood, "breaking silence", and the construction of a common history post-conflict – through the case study of the disappeared of Northern Ireland. The author's qualitative approach is particularly well suited to this aim. The relatively small number of victims and limited scope of the ICLVR allow for a careful, detailed examination of the Commission's record. The book is based both on archival research, mostly in news media archives, and on a series of twenty semi-structured interviews with relatives of the disappeared, victim advocates, ex-combatants from the Republican movement, staff members of the ICLVR, academics, journalists, politicians and others.[4] This methodology is complemented by an interdisciplinary approach, drawing on concepts from transitional justice, socio-legal studies, anthropology, criminology, memory studies and human geography. The back-and-forth between the interrogation of academic concepts and the voices of the people interviewed, regularly quoted in the text, allows the reader not to lose sight of the human stakes of the issue. Most importantly, it guarantees that the book's examination of key transitional justice notions such as silence, collective memory and reconciliation is rooted in the lived-in reality of a post-conflict situation.

The working process of the ICLVR is touched on, but is not at the centre of the book. Dempster focuses instead on the Commission as a response to the issue of the disappeared, and on the way key actors have responded to the Commission in return. The result is an investigation of the dialogue between a mechanism of transitional justice and the communities it serves. As such, the book is a welcome addition to the literature on truth commissions and other non-judicial forms of transitional justice, which has traditionally relied on empirical fieldwork with a more quantitative approach to evaluate their record. It also fills a gap in the mostly State-centric literature, as a case study on the response to disappearances perpetrated by non-State armed groups.

In post-conflict settings, the issue of the disappeared and the truth commissions aiming to establish their fate has regularly become tied to the tension between the need for justice and the need for peace. The introduction of limited immunity or amnesty-like measures to incentivize participation is a well-known source of controversy.[5] In Northern Ireland, the ICLVR is bound by guarantees of confidentiality in order to encourage the provision of information by ex-members of paramilitary groups believed to have committed the disappearances. Information given to the Commission cannot be used in criminal

4 A list of interviewees is provided at the end of the book, except for four individuals who wished to remain anonymous.

5 Some scholars have challenged the common assumption of truth commissions as paving the way to impunity, however. See, for example, Priscilla B. Hayner's highly regarded empirical review of truth commissions, Unspeakable Truth: Facing the Challenge of Truth Commissions, New York: Routledge, 2011.

proceedings; forensic testing on remains is limited to establishing a victim's identity. There is one exception to this confidential approach: the ICLVR is allowed, when it receives information about a potential location of a victim's remains, to share that location with the victim's family. Though it is an oversimplification, the justice versus peace opposition nevertheless touches on the diverse and sometimes conflicting needs of victims, families and communities post-conflict – a reality that mechanisms of transitional justice have to negotiate. Dempster points out that there has traditionally been a gulf between theory and practice in the way that such mechanisms have responded to victims' needs. Looking at the proven track record of the ICLVR, she argues that it represents an example of a truly victim-centred form of transitional justice, without hiding the "enormous compromise" at its core.[6] If the term "success" can seem inappropriate in light of this compromise, the ICLVR is nevertheless argued to have avoided the major pitfalls of truth commissions: becoming politicized, being doomed by a lack of trust in the process by crucial participants, being unable to manage expectations, and alienating or re-traumatizing victims.

Building on the value of "quiet diplomacy" to build trust and relationships, Dempster convincingly attributes this positive outcome to the ICLVR being an example of a form of "quiet transitional justice". Her analysis points to the confidentiality of the ICLVR's investigations as a key factor in its ability to deliver on its mandate and meet a specific need expressed by the victims' families: the return of the remains of the disappeared for burial. The association between "quiet" and "victim-centred" may appear counterintuitive, when silence is both a common motive behind the disappearances (as a means to silence an alleged informer) and the cause of much of the families' suffering. This is where the multiplicity of voices included in the book and the author's interrogation of classic tropes of transitional justice and post-conflict reconciliation strengthen its central argument. Dempster dives into the motives and rationales of victims' families, communities, political actors and perpetrators to keep or break their silence. She points out how current power relations frame who speaks out and the way collective memory on the issue of the disappeared is shaped. Finally, she highlights how the families' powerful public campaign and the ICLVR's confidentiality guarantees have provided the right incentive structure to spur participation and trust in the process. Here lies part of the value to be found in the case study of the ICLVR: it depicts how a "successful" mechanism of transitional justice addressing the issue of the disappeared integrates the current motivations and needs of the different actors involved.

The book's conclusion turns quite logically to the implications of the ICLVR for a broader effort to address conflict legacy issues in Northern Ireland. The Commission's restricted mandate is both a crucial factor in its success and its most obvious limit, leaving needs beyond the recovery and repatriation of

6 The central implication of this "enormous compromise", as the author calls it on p. 81, is that the information uncovered by the ICLVR will lead neither to the public identification of perpetrators and of the motives for their crimes, nor to their criminal prosecution.

remains unmet. Dempster's answer is hopeful but granular. She highlights the potential of the ICLVR's experience to be replicated to address other issues, but also uncovers factors that might impede such efforts, the grey areas – notably the role of larger communities and institutions in conflict violence – that will emerge. In a perhaps predictable but most appropriate conclusion, the book ends with one last quote from the author's interview with Eugene McVeigh, brother of Columba McVeigh, one of the disappeared: "I see the movement towards peace – albeit fragile and difficult and all of that – as worthwhile and worthy …[;] we have a stake in it because we made a sacrifice."[7] This final note highlights the importance of affected families and communities' agency and sense of purpose in "doing transitional justice" and negotiating the "tensions between principle and pragmatism" at its core.[8]

7 *Transitional Justice and the "Disappeared" of Northern Ireland*, p. 235.
8 As synthetized by Kieran McEvoy and Louise Mallinder in "Amnesties in Transition: Punishment, Restoration and the Governance of Mercy", *Journal of Law and Society*, Vol. 39, No. 3, 2012, p. 412. Quoted in *Transitional Justice and the "Disappeared" of Northern Ireland*, p. 183.

International Review of the Red Cross (2020), **102** (913), 463–479.
Digital technologies and war
doi:10.1017/S1816383120000454

IRRC

REPORTS AND DOCUMENTS

Artificial intelligence and machine learning in armed conflict: A human-centred approach

Note: This is an edited version of a paper published by the ICRC in June 2019.

1. Introduction

At a time of increasing conflict and rapid technological change, the International Committee of the Red Cross (ICRC) needs both to understand the impact of new technologies on people affected by armed conflict and to design humanitarian solutions that address the needs of the most vulnerable.

The ICRC, like many organizations across different sectors and regions, is grappling with the implications of **artificial intelligence** (AI) and **machine learning** for its work. AI is the use of computer systems to carry out tasks – often associated with human intelligence – that require cognition, planning, reasoning or learning; and machine learning systems are AI systems that are "trained" on and "learn" from data, which ultimately define the way they function. Since these are software tools, or algorithms, that could be applied to many different tasks, the potential implications may be far-reaching and yet to be fully understood.

There are two broad – and distinct – areas of application of AI and machine learning in which the ICRC has a particular interest: first, its **use in the conduct of warfare** or in other situations of violence;[1] and second, its **use in humanitarian action** to assist and protect the victims of armed conflict.[2] This paper sets out the

ICRC's perspective on the use of AI and machine learning in armed conflict, the potential humanitarian consequences, and the associated legal obligations and ethical considerations that should govern its development and use. It also makes reference to the use of AI tools for humanitarian action, including by the ICRC.

2. The ICRC's approach to new technologies of warfare

The ICRC has a long tradition of assessing the implications of contemporary and near-future developments in armed conflict. This includes considering new means and methods of warfare; specifically, in terms of their compatibility with the rules of international humanitarian law (also known as the law of armed conflict, or the law of war) and the risks of adverse humanitarian consequences for protected persons.

The ICRC is not opposed to new technologies of warfare *per se*. Certain military technologies – such as those enabling greater precision in attacks – may assist conflict parties in minimizing the humanitarian consequences of war, in particular on civilians, and in ensuring respect for the rules of war. However, as with any new technology of warfare, precision technologies are not beneficial in themselves, and humanitarian consequences on the ground will depend on the way new weapons are used in practice. It is essential, therefore, to have a realistic assessment of new technologies that is informed by their technical characteristics *and* the way they are used, or are intended to be used.

Any new technology of warfare must be used, and must be capable of being used, in compliance with existing rules of international humanitarian law. This is a minimum requirement.[3] However, the unique characteristics of new technologies of warfare, the intended and expected circumstances of their use, and their foreseeable humanitarian consequences may raise questions of whether existing rules are sufficient or need to be clarified or supplemented, in light of the new technologies' foreseeable impact.[4] What is clear is that military applications of new and emerging technologies are not inevitable. They are choices made by

1 ICRC, "Expert Views on the Frontiers of Artificial Intelligence and Conflict", *ICRC Humanitarian Law and Policy Blog*, 19 March 2019, available at: https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict.

2 ICRC, *Summary Document for UN Secretary-General's High-Level Panel on Digital Cooperation*, January 2019, available at: https://digitalcooperation.org/wp-content/uploads/2019/02/ICRC-Submission-UN-Panel-Digital-Cooperation.pdf.

3 States party to Additional Protocol I to the Geneva Conventions have an obligation to conduct legal reviews of new weapons during their development and acquisition, and prior to their use in armed conflict. For other States, legal reviews are a common-sense measure to help ensure that the State's armed forces can conduct hostilities in accordance with their international obligations.

4 ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, report for the 33rd International Conference of the Red Cross and Red Crescent, Geneva, October 2019 (ICRC Challenges Report 2019), pp. 18–29, available at: www.icrc.org/en/publication/4427-international-humanitarian-law-and-challenges-contemporary-armed-conflicts; ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, report for the 32nd International Conference of the Red Cross and Red Crescent, Geneva, October 2015 (ICRC Challenges Report 2015), pp. 38–47, available at: www.icrc.org/en/document/international-humanitarian-law-and-challenges-contemporary-armed-conflicts.

States which must be within the bounds of existing rules and must take into account potential humanitarian consequences for civilians and for combatants no longer taking part in hostilities, as well as broader considerations of "humanity" and "public conscience".[5]

3. Use of AI and machine learning by conflict parties

The ways in which parties to armed conflict – whether States or non-State armed groups – might use AI and machine learning in the conduct of warfare, and their potential implications, are not yet fully known. Nevertheless, there are at least **three overlapping areas that are relevant from a humanitarian perspective**, including for compliance with international humanitarian law.

3.1 Increasing autonomy in physical robotic systems, including weapons

One significant application is the use of digital **AI and machine learning tools to control physical military hardware**, in particular the increasing number of unmanned robotic systems – in the air, on land and at sea – with a wide range of sizes and functions. AI and machine learning may enable increasing autonomy in these robotic platforms, whether armed or unarmed, and whether controlling the whole system or specific functions such as flight, navigation, surveillance or targeting.

For the ICRC, **autonomous weapon systems** – weapon systems with autonomy in their "critical functions" of selecting and attacking targets – are an immediate concern from a humanitarian, legal and ethical perspective, given the risk of loss of human control over weapons and the use of force.[6] This loss of control raises risks for civilians, because of unpredictable consequences; legal questions,[7] because combatants must make context-specific judgements in carrying out attacks under international humanitarian law; and ethical concerns,[8] because human agency in decisions to use force is necessary to uphold moral

5 The "principles of humanity" and the "dictates of public conscience" are mentioned in Article 1(2) of Additional Protocol I and in the preamble of Additional Protocol II to the Geneva Conventions, referred to as the Martens Clause, which is part of customary international humanitarian law.

6 ICRC, Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 25–29 March 2019, available at: https://tinyurl.com/yyeadno3.

7 ICRC Challenges Report 2019, above note 4, pp. 29–31; Neil Davison, "Autonomous Weapon Systems under International Humanitarian Law", Perspectives on Lethal Autonomous Weapon Systems, United Nations Office for Disarmament Affairs Occasional Paper No. 30, November 2017, available at: www.icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law.

8 ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, report of an expert meeting, Geneva, 3 April 2018, available at: www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control.

responsibility and human dignity. For these reasons, the ICRC has proposed practical elements of human control as the basis for internationally agreed limits on autonomy in weapon systems with a focus on the following:[9]

- **Controls on weapon parameters**, which can inform limits on types of autonomous weapon systems including the targets they are used against, as well as limits on their duration and geographical scope of operation, and requirements for deactivation and fail-safe mechanisms;
- **Controls on the environment**, which can inform limits on the situations and locations in which autonomous weapon systems may be used, notably in terms of the presence and density of civilians and civilian objects; and
- **Controls through human–machine interaction**, which can inform requirements for human supervision and ability to intervene and deactivate autonomous weapon systems, and requirements for predictable and transparent functioning.

It is important to recognize that **not all autonomous weapons incorporate AI and machine learning**; existing weapons with autonomy in their critical functions, such as air-defence systems with autonomous modes, generally use simple, rule-based control software to select and attack targets. However, **AI and machine learning software** – specifically of the type developed for "automatic target recognition" – **could form the basis of future autonomous weapon systems, bringing a new dimension of unpredictability to these weapons**, as well as concerns about lack of explainability and bias (see Section 5.2).[10] The same type of software might also be used in "decision support" applications for targeting, rather than directly to control a weapon system (see Section 3.3).

Conversely, not all military robotic systems using AI and machine learning are autonomous weapons, since the software might be used for control functions other than targeting, such as surveillance, navigation or flight. While, from the ICRC's perspective, autonomy in weapon systems – including AI-enabled systems – raises the most urgent questions, the use of AI and machine learning to increase autonomy in military hardware in general – such as in unmanned aircraft, land vehicles and sea vessels – may also raise questions of human–machine interaction and safety. Discussions in the civil sector about ensuring safety of autonomous vehicles – such as self-driving cars or drones – may hold lessons for their use in armed conflict (see also Section 3.3).

9 ICRC, *ICRC Commentary on the "Guiding Principles" of the CCW GGE on "Lethal Autonomous Weapons Systems"*, Geneva, July 2020, available at: https://documents.unoda.org/wp-content/uploads/2020/07/20200716-ICRC.pdf; Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*, ICRC and Stockholm International Peace Research Institute, June 2020, available at: www.icrc.org/en/document/limits-autonomous-weapons; ICRC, "The Element of Human Control", UN Doc. CCW/MSP/2018/WP.3, working paper, CCW Meeting of High Contracting Parties, 20 November 2018, available at: https://tinyurl.com/y3c96aa6.

10 ICRC, Statement to the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems under Agenda Item 6(b), Geneva, 27–31 August 2018, available at: https://tinyurl.com/y4cql4to.

3.2 New means of cyber and information warfare

The application of **AI and machine learning to the development of cyber weapons or capabilities** is another important area. Not all cyber capabilities incorporate AI and machine learning. However, these technologies are expected to **change the nature of both capabilities to defend against cyber attacks and capabilities to attack**. For example, AI and machine learning-enabled cyber capabilities could automatically search for vulnerabilities to exploit, or defend against cyber attacks while simultaneously automatically launching counter-attacks. These types of developments could increase the scale, and change the nature and perhaps the severity, of attacks.[11] Some of these systems might even be described as "digital autonomous weapons", potentially raising questions about human control similar to those that apply to physical autonomous weapons.[12]

The ICRC's focus with respect to cyber warfare remains on ensuring that existing international humanitarian law rules are upheld in any cyber attacks in armed conflict, and that the particular challenges in ensuring the protection of civilian infrastructure and services are addressed by those carrying out or defending against such attacks,[13] in order to minimize the human cost.[14]

A related application of AI and machine learning in the digital sphere is the **use of these tools for information warfare**, in particular the creation and spreading of false information with intent to deceive – i.e., **disinformation** – as well as the spreading of false information without such intent – i.e., **misinformation**. Not all involve AI and machine learning, but these technologies seem set to change the nature and scale of the manipulation of information in warfare as well as the potential consequences. AI-enabled systems have been widely used to produce fake information – whether text, audio, photos or video – which is increasingly difficult to distinguish from real information. Use of these systems by conflict parties to amplify age-old methods of propaganda in order to manipulate opinion and influence decisions could have significant implications on the ground.[15] For the ICRC, there are concerns that civilians might, as a result of digital disinformation or misinformation, be subject to arrest or ill-treatment, discrimination or denial of access to essential services, or attacks on their person or property.[16]

11 Miles Brundage *et al.*, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, Oxford, February 2018.

12 United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, 2017.

13 By asserting that international humanitarian law applies to cyber operations, the ICRC is in no way condoning cyber warfare, nor is it condoning the militarization of cyberspace: ICRC Challenges Report 2015, above note 4, pp. 38–44.

14 ICRC, *The Potential Human Cost of Cyber Operations*, report of an expert meeting, Geneva, May 2019, available at: www.icrc.org/en/document/potential-human-cost-cyber-operations.

15 Steven Hill and Nadia Marsan, "Artificial Intelligence and Accountability: A Multinational Legal Perspective", in *Big Data and Artificial Intelligence for Military Decision Making*, STO Meeting Proceedings STO-MP-IST-160, NATO, 2018.

16 ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict*, March 2019, p. 9, available at: www.icrc.org/en/event/digital-risks-symposium.

3.3 Changing nature of decision-making in armed conflict

Perhaps the broadest and most far-reaching application is the use of **AI and machine learning for decision-making**, enabling widespread collection and analysis of data sources in order to identify people or objects, assess patterns of life or behaviour, make recommendations for military strategy or operations, or make predictions about future actions or situations.

These **"decision support"** or **"automated decision-making" systems are effectively an expansion of intelligence, surveillance and reconnaissance tools**, using AI and machine learning to automate the analysis of large data sets in order to provide "advice" to humans in making particular decisions, or to automate both the analysis and the subsequent initiation of a decision or action by the system. Relevant AI and machine learning applications include pattern recognition, natural language processing, image recognition, facial recognition and behaviour recognition. The **possible use of these systems is extremely broad**,[17] from decisions about who – or what – to attack and when,[18] to decisions about who to detain and for how long,[19] to decisions about military strategy – even on use of nuclear weapons[20] – and specific operations, including attempts to predict or pre-empt adversaries.[21] Depending on their use or misuse – and the capabilities and limitations of the technology – these decision-making applications could lead to increased risks for civilian populations.

AI and machine learning-based **decision support systems** may enable better decisions by humans in conducting hostilities in compliance with international humanitarian law and minimizing risks for civilians by facilitating quicker and more widespread collection and analysis of available information. However, over-reliance on the same algorithmically generated analyses, or predictions, might also facilitate worse decisions or violations of international humanitarian law and exacerbate risks for civilians, especially given the current limitations of the technology, such as unpredictability, lack of explainability and bias (see Section 5.2).

From a humanitarian perspective, a **very wide range of different AI-mediated – or AI-influenced – decisions by conflict parties could be relevant**, especially where they pose risks of injury or death to persons or destruction of objects, and where the decisions are governed by specific rules of international humanitarian law. For example, the use of AI and machine learning for **targeting decisions in armed conflict**, where there are serious consequences for life, will require specific considerations to ensure humans remain in a position to make

17 Dustin A. Lewis, Gabriella Blum and Naz K. Modirzadeh, *War-Algorithm Accountability*, Harvard Law School Program on International Law and Armed Conflict, August 2016.

18 United States, "Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems", working paper, CCW Group of Governmental Experts, March 2019.

19 Ashley Deeks, "Predicting Enemies", Virginia Public Law and Legal Theory Research Paper No. 2018-21, March 2018, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3152385.

20 Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, Vol. 1: *Euro-Atlantic Perspectives*, Stockholm International Peace Research Institute, Stockholm, May 2019.

21 S. Hill and N. Marsan, above note 15.

the context-based judgements required for compliance with the legal rules on the conduct of hostilities (see Section 5). An AI system used to directly initiate an attack (rather than producing an analysis, or "advice", for human decision-makers) would effectively be considered an autonomous weapon system, raising similar issues (see Section 3.1).

The use of decision support and automated decision-making systems may also raise **legal and ethical questions for other applications, such as decisions on detention in armed conflict**, which also have serious consequences for people's lives and are governed by specific rules of international humanitarian law. Here there are parallels with discussions in the civil sector about the role of human judgement, and issues of bias and inaccuracy, in risk-assessment algorithms used by the police in decisions on arrest, and in the criminal justice system for decisions on sentencing and bail.[22]

More broadly, these types of AI and machine learning tools might lead to an increasing **personalization of warfare** (with parallels to the personalization of services in the civilian world), with digital systems bringing together personally identifiable information from multiple sources – including sensors, communications, databases, social media and biometric data – to form an algorithmically generated determination about a person, their status and their targetability, or to predict their future actions.

In general, potential humanitarian consequences – **digital risks** – for civilian populations from misuse of AI-enabled **digital surveillance, monitoring and intrusion** technologies could include being targeted, arrested, facing ill-treatment, having their identity stolen and being denied access to services, having assets stolen or suffering from psychological effects from the fear of being under surveillance.[23]

4. Use of AI and machine learning for humanitarian action

The ways in which AI and machine learning might be used for humanitarian action, including by the ICRC, are also likely to be very broad. These tools are being explored by humanitarian organizations for environment scanning, monitoring and analysis of public sources of data in specific operational contexts – applications that could help **inform assessments of humanitarian needs**, such as the type of assistance needed (food, water, shelter, economic, health) and where it is needed.

Similar AI-enabled data aggregation and analysis tools might be used to help **understand humanitarian consequences** on the ground, including civilian protection needs – for example, tools for image, video or other pattern analysis to assess damage

22 Lorna McGregor, "The Need for Clear Governance Frameworks on Predictive Algorithms in Military Settings", *ICRC Humanitarian Law and Policy Blog*, 28 March 2019, available at: https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings; AI Now Institute, *AI Now Report 2018*, New York University, December 2018, pp. 18–22.

23 ICRC, above note 16, p. 8.

to civilian infrastructure, patterns of population displacement, viability of food crops, or the degree of weapon contamination (unexploded ordnance). These systems might also be used to analyze images and videos in order to detect and assess the conduct of hostilities, and the resulting humanitarian consequences.

The ICRC, for example, has developed **environment scanning dashboards** using AI and machine learning to capture and analyze large volumes of data to inform and support its humanitarian work in specific operational contexts, including using predictive analytics to help determine humanitarian needs.

A wide range of humanitarian services might benefit from the application of AI and machine learning tools for specific tasks. For example, there is interest in technologies that could **improve identification of missing persons**, such as AI-based facial recognition and natural language processing for name matching; the ICRC has been exploring the use of these technologies to support the work of its Central Tracing Agency in reuniting family members separated by conflict. It is also exploring the use of AI and machine learning-based **image analysis and pattern recognition for satellite imagery**, whether to map population density in support of infrastructure assistance projects in urban areas or to complement its documentation of respect for international humanitarian law as part of its civilian protection work.

These **applications for humanitarian action also bring potential risks**, as well as legal and ethical questions, in particular with respect to data protection, privacy, human rights, accountability and ensuring human involvement in decisions with significant consequences for people's lives and livelihoods. Any applications for humanitarian action must be designed and used under the principle of **"do no harm"** in the digital environment, and respect the right to privacy, including as it relates to personal data protection.

The ICRC will also ensure that the **core principles and values of neutral, independent and impartial humanitarian action** are reflected in the design and use of AI and machine learning applications it employs, taking into account a realistic assessment of the capabilities and limitations of the technology (see Section 5.2). The ICRC has led – with the Brussels Privacy Hub – an initiative on data protection in humanitarian action to develop guidance on the use of new technologies, including AI and machine learning, in the humanitarian sector in a way that maximizes the benefits without losing sight of these core considerations. The second edition of the ICRC/Brussels Privacy Hub *Handbook on Data Protection in Humanitarian Action* was published in May 2020.[24]

5. A human-centred approach

As a humanitarian organization working to protect and assist people affected by armed conflict and other situations of violence, deriving its mandate from international humanitarian law and guided by the Fundamental Principle of

24 ICRC and Brussels Privacy Hub, *Handbook on Data Protection in Humanitarian Action*, 2nd ed., Geneva, May 2020, available at: www.icrc.org/en/data-protection-humanitarian-action-handbook.

humanity,[25] the **ICRC believes it is critical to ensure a genuinely human-centred approach to the development and use of AI and machine learning**. This starts with consideration of the obligations and responsibilities of humans and what is required to ensure that the use of these technologies is compatible with international law, as well as societal and ethical values.

5.1 Ensuring human control and judgement

The ICRC believes it is **essential to preserve human control over tasks and human judgement in decisions that may have serious consequences** for people's lives in armed conflict, especially where these tasks and decisions pose risks to life, and where they are governed by specific rules of international humanitarian law. **AI and machine learning systems must be used to serve human actors, and as tools to augment human decision-makers, not replace them**. Given that these technologies are being developed to perform tasks that would ordinarily be carried out by humans, there is an inherent tension between the pursuit of AI and machine learning applications and the centrality of the human being in armed conflict, which will need continued attention.

Human control and judgement will be particularly important for tasks and decisions that can lead to injury or loss of life, or damage to, or destruction of, civilian infrastructure. These will likely raise the most serious legal and ethical questions, and may demand policy responses, such as new rules and regulations. **Most significant are decisions on the use of force, determining who and what is targeted and attacked in armed conflict**. However, a much wider range of tasks and decisions to which AI might be applied could also have serious consequences for those affected by armed conflict, such as decisions on arrest and detention. In considering the use of AI for sensitive tasks and decisions, there may be lessons from broader discussions in the civil sector about the governance of "safety-critical" AI applications – those whose failure can lead to injury or loss of life, or serious damage to property or the environment.[26]

Another area of tension is the **discrepancy between humans and machines in the speed at which they carry out different tasks**. Since humans are the legal – and moral – agents in armed conflict, the technologies and tools they use to conduct warfare must be designed and used in a way that enables combatants to fulfil their legal and ethical obligations and responsibilities. This may have significant implications for AI and machine learning systems that are used in decision-

25 ICRC and International Federation of Red Cross and Red Crescent Societies, *The Fundamental Principles of the International Red Cross and Red Crescent Movement: Ethics and Tools for Humanitarian Action*, Geneva, November 2015, available at: https://shop.icrc.org/les-principes-fondamentaux-de-la-croix-rouge-et-du-croissant-rouge-2757.html.

26 See, for example, the Partnership on AI's focus on the safety of AI and machine learning technologies as "an urgent short-term question, with applications in medicine, transportation, engineering, computer security, and other domains hinging on the ability to make AI systems behave safely despite uncertain, unanticipated, and potentially adversarial environments". Partnership on AI, "Safety-Critical AI: Charter", 2018, available at: www.partnershiponai.org/working-group-charters-guiding-our-exploration-of-ais-hard-questions.

making; in order to preserve human judgement, systems may need to be designed and used to inform decision-making at "human speed", rather than accelerating decisions to "machine speed" and beyond human intervention.

Legal basis for human control in armed conflict

For conflict parties, **human control over AI and machine learning applications employed as means and methods of warfare is required to ensure compliance with the law**. The 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. In particular, combatants have a unique obligation to make the judgements required of them by the international humanitarian law rules governing the conduct of hostilities, and this responsibility cannot be transferred to a machine, a piece of software or an algorithm.

These rules require context-specific judgements to be taken by those who plan, decide upon and carry out attacks, in order to: ensure **distinction** – between military objectives, which may lawfully be attacked, and civilians or civilian objects, which must not be attacked; ensure **proportionality** – in terms of ensuring that the incidental civilian harm expected from an attack will not be excessive in relation to the concrete and direct military advantage anticipated; and enable **precautions in attack** – so that risks to civilians can be further minimized.

Where AI systems are used in attacks – whether as part of physical or cyber-weapon systems, or in decision support systems – **their design and use must enable combatants to make these judgements**.[27] With respect to autonomous weapon systems, the States party to the Convention on Certain Conventional Weapons (CCW), have recognized that "human responsibility" for the use of weapon systems and the use of force "must be retained",[28] and many States, international organizations – including the ICRC – and civil society organizations have stressed the requirement for human control to ensure compliance with international humanitarian law and compatibility with ethical values.[29]

Beyond the use of force and targeting, the potential use of AI systems for other decisions governed by specific rules of international humanitarian law will likely require careful consideration of necessary human control, and judgement, such as in detention.[30]

27 ICRC, above note 6.

28 United Nations, *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2018/3, 23 October 2018, Sections III.A.26(b), III.C.28(f), available at: http://undocs.org/en/CCW/GGE.1/2018/3.

29 See, for example, the statements delivered at the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 25–29 March 2019, available at: https://tinyurl.com/yyeadno3.

30 Tess Bridgeman, "The Viability of Data-Reliant Predictive Systems in Armed Conflict Detention", *ICRC Humanitarian Law and Policy Blog*, 8 April 2019, available at: https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention.

Ethical basis for human control

Emerging applications of AI and machine learning have also brought ethical questions to the forefront of public debate. **A common aspect of general "AI principles"** developed and agreed by governments, scientists, ethicists, research institutes and technology companies **is the importance of the human element** to ensure legal compliance and ethical acceptability.

For example, the 2017 Asilomar AI Principles emphasize alignment with human values, compatibility with "human dignity, rights, freedoms and cultural diversity", and human control; "humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives".[31] The European Commission's High-Level Expert Group on Artificial Intelligence stressed the importance of "human agency and oversight", such that AI systems should "support human autonomy and decision-making", and of ensuring human oversight through human-in-the-loop, human-on-the-loop or human-in-command approaches.[32] The Organisation for Economic Co-operation and Development (OECD) Principles on Artificial Intelligence – adopted in May 2019 by all thirty-six member States, together with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania – highlight the importance of "human-centred values and fairness", specifying that users of AI "should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art".[33] The Beijing AI Principles, adopted in May 2019 by a group of leading Chinese research institutes and technology companies, state that "continuous efforts should be made to improve the maturity, robustness, reliability, and controllability of AI systems" and encourage "explorations on Human-AI coordination … that would give full play to human advantages and characteristics".[34] A number of individual technology companies have also published AI principles highlighting the importance of human control,[35] especially for sensitive applications presenting the risk of harm,[36] and emphasizing that the "purpose of AI … is to augment – not replace – human intelligence".[37]

31 Future of Life Institute, "Asilomar AI Principles", 2017, available at: https://futureoflife.org/ai-principles.
32 European Commission, *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, 8 April 2019, pp. 15–16, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
33 OECD, "Recommendation of the Council on Artificial Intelligence", OECD/LEGAL/0449, 22 May 2019, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
34 Beijing Academy of Artificial Intelligence, "Beijing AI Principles", 28 May 2019, available at: https://baip.baai.ac.cn/en.
35 Google, "AI at Google: Our Principles", *The Keyword*, 7 June 2018, available at: www.blog.google/technology/ai/ai-principles. "We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control."
36 Microsoft, "Microsoft AI Principles", 2019, available at: www.microsoft.com/en-us/ai/our-approach-to-ai; Rich Sauer, "Six Principles to Guide Microsoft's Facial Recognition Work", *Microsoft Blog*, 17 December 2018, available at: https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work.
37 IBM, "IBM's Principles for Trust and Transparency", *THINKPolicy Blog*, 30 May 2018 available at: www.ibm.com/blogs/policy/trust-principles.

Some governments are also developing AI principles for the military. For example, the US Department of Defense (DoD), which called for the "human-centered" adoption of AI in its 2018 AI Strategy,[38] tasked its Defense Innovation Board with providing recommendations. Foremost among them was that "[h]uman beings should exercise appropriate levels of judgment and remain responsible" for any use of AI.[39] This informed the first of five DoD principles adopted in early 2020, which states that AI must be "[r]esponsible. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities".[40] In France, the Ministry of Defence has committed to the use of AI in line with three guiding principles – compliance with international law, maintaining sufficient human control, and ensuring permanent command responsibility – and has established a Ministerial Ethics Committee to address emerging technologies.[41]

In the ICRC's view, preserving **human control** over tasks and **human judgement** in decisions that have serious consequences for people's lives will also be **essential to preserve a measure of humanity in warfare**. **The ICRC has stressed the need to retain human agency over decisions to use force in armed conflict**,[42] a view which derives from broader ethical considerations of humanity, moral responsibility, human dignity and the dictates of public conscience.[43]

However, ethical considerations of human agency may have broader applicability to other uses of AI and machine learning in armed conflict and other situations of violence. There are perhaps **lessons from wider societal discussions about sensitive applications of dual-use AI and machine learning technologies**, especially for safety-critical applications, and associated proposals for governance by scientists and developers in the private sector. Google, for example, has said that there may be "sensitive contexts where society will want a human to make the final decision, no matter how accurate an AI system", and that fully delegating high-stakes decisions to machines – such as legal judgements of criminality or life-altering decisions about medical treatment – "may fairly be seen as an affront to human dignity".[44] Microsoft, in considering AI-based facial recognition, has emphasized ensuring "an appropriate level of human control for uses that may affect people in consequential ways", requiring a "human in the loop" or "meaningful human review" for sensitive uses such as those involving

38 DoD, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, 2019.

39 DoD, Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, 31 October 2019.

40 DoD, "DOD Adopts Ethical Principles for Artificial Intelligence", news release, 24 February 2020, available at: www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

41 French Ministry of Defence, "Florence Parly Wants High-Performance, Robust and Properly Controlled Artificial Intelligence", *Actualités*, 10 April 2019, available at: www.defense.gouv.fr/english/actualites/articles/florence-parly-souhaite-une-intelligence-artificielle-performante-robuste-et-maitrisee.

42 ICRC, *ICRC Strategy 2019–2022*, Geneva, 2018, p. 15, available at: www.icrc.org/en/publication/4354-icrc-strategy-2019-2022.

43 ICRC, above note 8, p. 22.

44 Google, *Perspectives on Issues in AI Governance*, January 2019, pp. 23–24, available at: http://ai.google/perspectives-on-issues-in-AI-governance.

"risk of bodily or emotional harm to an individual, where an individual's employment prospects or ability to access financial services may be adversely affected, where there may be implications on human rights, or where an individual's personal freedom may be impinged".[45] Since applications in armed conflict are likely to be among the most sensitive, these broader discussions may hold insights for necessary constraints on AI applications.

Preserving **human control and judgement will be an essential component** for ensuring legal compliance and mitigating ethical concerns raised by certain applications of AI and machine learning. **But it will not, in itself, be sufficient to guard against potential risks** without proper consideration of human–machine interaction issues such as: **situational awareness** (knowledge of the state of the system at the time of human intervention); **time available** for effective human intervention; **automation bias** (risk of human overtrust in the system); and the **moral buffer** (risk of humans transferring responsibility to the system).[46] Further, ensuring meaningful and effective human control and judgement will require careful consideration of both the capabilities and the limitations of AI and machine learning technologies.

5.2 Understanding the technical limitations of AI and machine learning

While much is made of the new capabilities offered by AI and machine learning, a **realistic assessment of the capabilities and limitations of these technologies is needed**, especially if they are to be used for applications in armed conflict. This should start with an acknowledgement that in using AI and machine learning for certain tasks or decisions, we are not replacing like with like. It requires an **understanding of the fundamental differences in the way humans and machines do things, as well as their different strengths and weaknesses**; humans and machines do things differently, and they do different things. We must be clear that, as inanimate objects and tools for use by humans, "machines will never be able to bring a genuine humanity to their interactions, no matter how good they get at faking it".[47]

With this in mind, there are several technical issues that demand caution in considering applications in armed conflict (and indeed for humanitarian action). **AI, and especially machine learning, brings concerns about unpredictability and unreliability** (or safety),[48] **lack of transparency** (or explainability), **and bias**.[49]

45 R. Sauer, above note 36: "We will encourage and help our customers to deploy facial recognition technology in a manner that ensures an appropriate level of human control for uses that may affect people in consequential ways."
46 ICRC, above note 8, p. 13.
47 Google, above note 44, p. 22.
48 Dario Amodei *et al.*, *Concrete Problems in AI Safety*, Cornell University, Ithaca, NY, 2016, available at: https://arxiv.org/abs/1606.06565.
49 ICRC, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, report of an expert meeting, Geneva, August 2019, available at: www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

Rather than following a pre-programmed sequence of instructions, **machine learning systems build their own rules based on the data they are exposed to** – whether training data or through trial-and-error interaction with their environment. **As a result, they are much more unpredictable** than pre-programmed systems in terms of how they will function (reach their output) in a given situation (with specific inputs), and their functioning is highly dependent on the quantity and quality of available data for a specific task. For the developer it is difficult to know when the training is complete, or even what the system has learned. The same machine learning system may respond differently even when exposed to the same situation, and some systems may lead to unforeseen solutions to a particular task.[50] These core problems are exacerbated when the system continues to "learn" and change its model after deployment for a specific task. The unpredictable nature of machine learning systems, which can be an advantage in solving tasks, may not be a problem for benign tasks, such as playing a board game,[51] but it may be a significant concern for applications in armed conflict, such as autonomous weapon systems, cyber warfare and decision support systems (see Sections 3.1–3.3).

Complicating matters further, many machine learning systems are **not transparent**; **they produce outputs that are not explainable**. This "black box" nature makes it difficult – and, in many cases, currently impossible – for the user to understand *how* and *why* the system reaches its output from a given input; in other words, there is a lack of explainability and interpretability.

These issues of unpredictability and lack of explainability make **establishing trust in AI and machine learning systems a significant challenge**. An additional problem for trust is **bias**, which can have many facets, whether reinforcing existing human biases or introducing new ones in the design and/or use of the system. A common form is bias from training data, where limits in the quantity, quality and nature of available data to train an algorithm for a specific task can introduce bias into the functioning of the system relative to its task. This will likely be a significant issue for applications in armed conflict, where high-quality, representative data for specific tasks is scarce. However, other forms of bias can derive from the weighting given to different elements of data by the system, or to its interaction with the environment during a task.[52]

Concerns about unpredictability, lack of transparency or explainability, and bias have been documented in various applications of AI and machine learning, for example in image recognition,[53] facial recognition[54] and automated

50 Joel Lehman *et al.*, *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, Cornell University, Ithaca, NY, 2018, available at: https://arxiv.org/abs/1803.03453.

51 David Silver *et al.*, "Mastering the Game of Go without Human Knowledge", *Nature*, Vol. 550, No. 7676, 19 October 2017.

52 UNIDIR, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer*, 2018.

53 Matthew Hutson, "A Turtle – or a Rifle? Hackers Easily Fool AIs into Seeing the Wrong Thing", *Science*, 19 July 2018, available at: www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing.

54 AI Now Institute, above note 22, pp. 15–17.

decision-making systems.[55] Another fundamental issue with applications of AI and machine learning, such as computer vision, is **the semantic gap**, which shows that humans and machines carry out tasks very differently.[56] A computer-vision algorithm trained on images of particular subjects may be able to identify and classify those subjects in a new image. However, the algorithm has no understanding of the *meaning* or *concept* of that subject, which means it can make mistakes that a human never would, such as classifying an object as something completely different and unrelated. This would obviously raise serious concerns in certain applications in armed conflict, such as in autonomous weapon systems or decision support systems for targeting (see Sections 3.1 and 3.3).

The use of AI and machine learning in armed conflict will likely be even more difficult to trust in situations where it can be assumed that adversaries will apply countermeasures such as trying to trick or spoof each other's systems. **Machine learning systems are particularly vulnerable to adversarial conditions**, whether modifications to the environment designed to fool the system or the use of another machine learning system to produce adversarial images or conditions (a generative adversarial network, or GAN). In a well-known example, researchers tricked an image classification algorithm into identifying a 3D-printed turtle as a "rifle", and a 3D-printed baseball as an "espresso".[57] The risks of this type of problem are also clear should an AI-based image recognition system be used in weapon systems or for targeting decisions.

6. Conclusions and recommendations

AI and machine learning systems could have **profound implications for the role of humans in armed conflict**, especially in relation to: increasing autonomy of weapon systems and other unmanned systems; new forms of cyber and information warfare; and, more broadly, the nature of decision-making. In the view of the ICRC, governments, militaries and other relevant actors in armed conflict must pursue a genuinely **human-centred approach to the use of AI and machine learning systems based on legal obligations and ethical responsibilities**. The use of AI in weapon systems must be approached with great caution.

As a general principle, it is **essential to preserve human control and judgement in applications of AI and machine learning for tasks and in decisions that may have serious consequences for people's lives**, especially where these tasks and decisions pose risks to life, and where they are governed by specific rules of international humanitarian law. **AI and machine learning systems remain tools that must be used to serve human actors, and augment human decision-makers, not replace them**.

55 *Ibid.*, pp. 18–22.
56 Arnold W. M. Smeulders *et al.*, "Content-Based Image Retrieval at the End of the Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, 2000.
57 M. Hutson, above note 53.

Ensuring human control and judgement in AI-enabled physical and digital systems that present such risks will be **needed for compliance with international humanitarian law and, from an ethical perspective, to preserve a measure of humanity in armed conflict**. In order for humans to meaningfully play their role, these systems may need to be designed and used to **inform decision-making at human speed, rather than accelerating decisions to machine speed** and beyond human intervention. These considerations may ultimately lead to constraints in the design and use of AI and machine learning systems to allow for meaningful and effective human control and judgement, based on legal obligations and ethical responsibilities.

An overall principle of human control and judgement is an essential component, but it is not sufficient in itself to guard against the potential risks of AI and machine learning in armed conflict. **Other related aspects to consider** will be ensuring: **predictability** and **reliability** – or safety – in the operation of the system and the consequences that result; **transparency** – or **explainability** – in how the system functions and why it reaches a particular output; and **lack of bias** – or fairness – in the design and use of the system. These issues will need to be addressed in order to **build trust** in the use of a given system, including through **rigorous testing in realistic environments** before being put into operation.[58]

The nature of human–AI interaction required will likely depend on ethical considerations and the particular rules of international humanitarian law and other applicable law that apply in the circumstances. Therefore, **general principles may need to be supplemented by specific principles, guidelines or rules on the use of AI and machine learning for specific applications and in particular circumstances**.

In the ICRC's view, one of the most pressing concerns is the relationship between humans and machines in decisions to kill, injure, damage or destroy, and the **critical importance of ensuring human control over weapon systems and the use of force** in armed conflict. With increasingly autonomous weapon systems, whether AI-enabled or not, there is a risk of effectively leaving these decisions to sensors and algorithms, a prospect that raises legal and ethical concerns which must be addressed with some urgency.

The ICRC has proposed key elements of human control necessary to comply with international humanitarian law and satisfy ethical concerns as a basis for internationally agreed limits on autonomy in weapon systems, including controls on weapon parameters, controls on the environment and

58 Netta Goussac, "Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-fighting", *ICRC Humanitarian Law and Policy Blog*, 18 April 2019, available at: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting; Dustin A. Lewis, "Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider", *ICRC Humanitarian Law and Policy Blog*, 21 March 2019, available at: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider.

controls through human–machine interaction.[59] It is clear to the ICRC that limits are needed on the types of autonomous weapons used and the situations in which they are used.[60]

This **human control-based approach** to autonomous weapon systems **would also be pertinent to broader applications of AI and machine learning in decision-making in armed conflict**, in particular where there are significant risks for human life and specific rules of international humanitarian law that apply, such as the use of decision support systems for targeting and detention.

59 ICRC, *Commentary on the "Guiding Principles"*, above note 9; ICRC, "The Element of Human Control", above note 9; V. Boulanin *et al.*, above note 9.

60 ICRC, Statement to the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 21–25 September 2020, available at: https://documents.unoda.org/wp-content/uploads/2020/09/20200921-ICRC-General-statement-CCW-GGE-LAWS-Sep-2020.pdf.

International Review of the Red Cross (2020), **102** (913), 481–492.
Digital technologies and war
doi:10.1017/S1816383120000478

IRRC

REPORTS AND DOCUMENTS

International humanitarian law and cyber operations during armed conflicts

ICRC position paper submitted to the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security and the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, November 2019

Executive summary

- **Cyber operations have become a reality in contemporary armed conflict**. The International Committee of the Red Cross (ICRC) is concerned by **the potential human cost** arising from the increasing use of cyber operations during armed conflicts.

- **In the ICRC's view, international humanitarian law (IHL) limits cyber operations during armed conflicts** just as it limits the use of any other weapon, means or method of warfare in an armed conflict, whether new or old.
- Affirming the applicability of IHL does not legitimize cyber warfare, just as it does not legitimize any other form of warfare. **Any use of force by States – cyber or kinetic – remains governed by the Charter of the United Nations and the relevant rules of customary international law**, in particular the prohibition against the use of force. International disputes must be settled by peaceful means, in cyberspace as in all other domains.
- It is now critical **for the international community to affirm the applicability of international humanitarian law** to the use of cyber operations during armed conflicts. The ICRC also calls for discussions among governmental and other experts on *how* existing IHL rules apply and whether the existing law is adequate and sufficient. In this respect, the **ICRC welcomes the intergovernmental discussions** currently taking place in the framework of two United Nations General Assembly mandated processes.
- Events of recent years have shown that cyber operations, whether during or outside armed conflict, can disrupt the operation of critical civilian infrastructure and hamper the delivery of essential services to the population. **In the context of armed conflicts, civilian infrastructure is protected against cyber attacks by existing IHL principles and rules,** in particular the principles of distinction, proportionality and precautions in attack. IHL also affords special protection to hospitals and objects indispensable to the survival of the civilian population, among others.
- **During armed conflicts, the employment of cyber tools that spread and cause damage indiscriminately is prohibited.** From a technological perspective, some cyber tools can be designed and used to target and harm only specific objects and to not spread or cause harm indiscriminately. However, the interconnectivity that characterizes cyberspace means that whatever has an interface with the Internet can be targeted from anywhere in the world and that a cyber attack on a specific system may have repercussions on various other systems. As a result, there is a real risk that cyber tools are not designed or used – either deliberately or by mistake – in compliance with IHL.
- **States' interpretation of existing IHL rules will determine the extent to which IHL protects against the effects of cyber operations**. In particular, States should take clear positions about their commitment to interpret IHL so as to preserve civilian infrastructure from significant disruption and to protect civilian data. The availability of such positions will also influence the assessment of whether the existing rules are adequate or whether new rules may be needed. If States see a need to develop new rules, they should **build on and strengthen the existing legal framework – including IHL**.

1. Introduction

The use of cyber operations during armed conflicts is a reality.[1] While only a few States have publicly acknowledged using such operations, an increasing number of States are developing military cyber capabilities, and their use is likely to increase in future.

Moreover, there have been significant technological advances in offensive cyber capabilities: in recent years, cyber operations have shown that they can seriously affect civilian infrastructure and might result in human harm.

In line with its mission and mandate, the International Committee of the Red Cross (ICRC) is primarily concerned with cyber operations used as means and methods of warfare during an armed conflict and the protection that international humanitarian law (IHL) affords against their effects.

The ICRC welcomes the intergovernmental discussions currently taking place in the framework of the two United Nations General Assembly mandated processes, namely the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security and the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security. Both groups are mandated to study "how international law applies to the use of information and communications technologies by States".[2] The ICRC submits this position paper to both groups to support States' deliberation on this matter.

This position paper is limited to legal and humanitarian questions arising from the use of cyber operations during armed conflict. It does not address questions relating to the legal framework applicable to cyber operations unrelated to armed conflict.

2. The potential human cost of cyber operations

During armed conflict, cyber operations have been used in support of or alongside kinetic operations. The use of cyber operations may offer alternatives that other means or methods of warfare do not, but it also carries risks. On the one hand, cyber operations have the potential to enable parties to armed conflicts to achieve their military aims without harming civilians or causing physical damage to civilian infrastructure. On the other hand, recent cyber operations – which have been mostly conducted outside the context of armed conflict – show that

1 In this position paper, the term "cyber operations during armed conflicts" is used to describe operations against a computer, a computer system or network, or another connected device, through a data stream, when used as a means or method of warfare in the context of an armed conflict. Cyber operations rely on information and communication technologies.

2 UNGA Res. 73/27, "Developments in the Field of Information and Telecommunications in the Context of International Security", UN Doc. A/RES/73/27, 5 December 2018, op. para. 5; UNGA Res. 73/266, "Advancing Responsible State Behaviour in Cyberspace in the Context of International Security", UN Doc. A/RES/73/266, 22 December 2018, op. para. 3.

sophisticated actors have developed the capability to disrupt the provision of essential services to the civilian population.

By means of cyber operations, it is possible for belligerents to infiltrate a system and collect, exfiltrate, modify, encrypt or destroy data. It is also possible to trigger, alter or otherwise manipulate processes controlled by a compromised computer system. A variety of "targets" in the real world can be disrupted, altered or damaged, such as industries, infrastructures, telecommunications, transport, or governmental and financial systems. Based on discussions with experts from all parts of the world and its own research, the ICRC is particularly concerned about the potential human cost of cyber operations on critical civilian infrastructure, including health infrastructure.[3]

In recent years, cyber attacks have exposed the vulnerability of essential services. They are reportedly becoming more frequent and their severity is increasing more rapidly than experts had anticipated. Moreover, much is unknown with respect to the most sophisticated cyber capabilities and tools that have been or are being developed, how technology may evolve, and the extent to which the use of cyber operations during armed conflicts might be different from the trends observed so far.

Moreover, the characteristics of cyberspace raise specific concerns. For example, cyber operations entail a risk for escalation and related human harm for the simple reason that it may be difficult for the targeted party to know whether the attacker's aim is intelligence collection or more harmful effects. The target may thereby react with greater force than necessary out of anticipation of a worst-case scenario.

Cyber tools also proliferate in a unique manner. Once used, they can be repurposed and widely used by actors other than the one that developed or used them initially.

3. The application of IHL to cyber operations during armed conflicts

For the ICRC, there is no question that IHL applies to, and therefore limits, cyber operations during armed conflict – just as it regulates the use of any other weapon, means or method of warfare in an armed conflict, whether new or old.[4]

3 See ICRC, *The Potential Human Cost of Cyber Operations*, Geneva, 2019, available at: www.icrc.org/en/download/file/96008/the-potential-human-cost-of-cyber-operations.pdf.

4 ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, 31IC/11/5.1.2, Geneva, 2011 (ICRC Challenges Report 2011), pp. 36–37, available at: www.icrc.org/en/doc/assets/files/red-cross-crescent-movement/31st-international-conference/31-int-conference-ihl-challenges-report-11-5-1-2-en.pdf; ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, 32IC/15/11, Geneva, 2015 (ICRC Challenges Report 2015), p. 40, available at: www.icrc.org/en/download/file/15061/32ic-report-on-ihl-and-challenges-of-armed-conflicts.pdf; ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, 33IC/19/9.7, Geneva, 2019 (ICRC Challenges Report 2019), p. 18, available at: https://rcrcconference.org/app/uploads/2019/10/33IC-IHL-Challenges-report_EN.pdf.

This holds true whether cyberspace is considered as a new domain of warfare similar to air, land, sea or outer space; a different type of domain because it is man-made while the former are natural; or not a domain as such.

When States adopt IHL treaties, they do so to regulate present and future conflicts. States have included rules that anticipate the development of new means and methods of warfare in IHL treaties, presuming that IHL will apply to them. For instance, if IHL did not apply to future means and methods of warfare, it would not be necessary to review their lawfulness under existing IHL, as required by Article 36 of the 1977 Additional Protocol I.

This conclusion finds strong support in the International Court of Justice's Advisory Opinion on the *Legality of the Threat or Use of Nuclear Weapons*: the Court recalled that the established principles and rules of IHL applicable in armed conflict apply "to all forms of warfare and to all kinds of weapons", including "those of the future".[5] In the ICRC's view, this finding applies to the use of cyber operations during armed conflict.

The ICRC welcomes that an increasing number of States and international organizations have affirmed that IHL applies to cyber operations during armed conflicts and welcomes discussion on how IHL applies.

States may also decide to impose additional limits to those found in existing law and to develop complementary rules, in particular in order to strengthen the protection of civilians and civilian infrastructure against the effects of cyber operations. In the ICRC's view, any new rules need to build on and strengthen the existing legal framework, including IHL.

In cases not covered by existing rules of IHL, civilians and combatants remain protected by the so-called "Martens Clause", meaning that they remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience.[6]

It is important to underline that affirming the application of IHL to cyber operations during armed conflict does not legitimize cyber warfare or encourage the militarization of cyberspace. In fact, IHL imposes some limits to the militarization of cyberspace by prohibiting the development of military cyber capabilities that would violate IHL.[7] Moreover, any use of force by States – cyber or kinetic – remains governed by the Charter of the United Nations and the relevant rules of

5 International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, 8 July 1996 (Nuclear Weapons Advisory Opinion), para. 86.

6 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) (AP I), Art. 1(2); Hague Convention (II) with respect to the Laws and Customs of War on Land and Its Annex: Regulations concerning the Laws and Customs of War on Land, The Hague, 29 July 1899 (entered into force 4 September 1900), preambular para. 9; Hague Convention (IV) respecting the Laws and Customs of War on Land and Its Annex: Regulations concerning the Laws and Customs of War on Land, The Hague, 18 October 1907 (entered into force 26 January 1910), preambular para. 8.

7 See, among others, Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, Vol. 1: *Rules*, Cambridge University Press, Cambridge, 2005 (ICRC Customary Law Study), Rules 70, 71, available at: https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1. See also AP I, Art. 36.

customary international law, in particular, the prohibition against the use of force. International disputes must be settled by peaceful means, in cyberspace as in all other domains.

4. The protection afforded by existing IHL

Existing IHL treaties and customary law provide rules on a number of issues during armed conflict. In cyberspace, the rules on the conduct of hostilities are particularly relevant. These rules aim to protect the civilian population against the effects of hostilities. They are based on the cardinal principle of distinction, which requires that belligerents distinguish at all times between the civilian population and combatants and between civilian objects and military objectives, and direct their operations only against military objectives.[8]

Notwithstanding the interconnectivity that characterizes cyberspace, a careful examination of the functioning of cyber tools shows that they are not necessarily indiscriminate. Many of the recent cyber attacks that have been reported in public sources appear to have been rather "discriminate" from a technical point of view: they have been designed and actually used to target and harm only specific objects and have not spread or caused harm indiscriminately. Ensuring that cyber operations affect only the targeted object may, however, be technically challenging and require careful planning in their design and use. Moreover, it must be noted that a cyber operation that is technically discriminate is not necessarily lawful, whether during or outside of an armed conflict.

This being said, some known cyber tools have been designed to self-propagate and indiscriminately affect widely used computer systems. They have not done so by chance: the ability to self-propagate needs to be specifically included in the design of such tools. The interconnectivity that characterizes cyberspace means that whatever has an interface with the Internet can be targeted from anywhere in the world. Moreover, an attack on a specific system may have repercussions on various other systems and cause indiscriminate effects. As a result, there is a real risk that cyber tools are not designed or used – either deliberately or by mistake – in compliance with IHL.

Affirming that IHL – including the principles of distinction, proportionality, and precaution – applies to cyber operations during armed conflicts means that under existing law, among many other rules:

- cyber capabilities that qualify as weapons and are by nature indiscriminate are prohibited;[9]
- direct attacks against civilians and civilian objects are prohibited, including when using cyber means or methods of warfare;[10]

8 AP I, Art. 48; ICRC Customary Law Study, above note 7, Rules 1, 7; Nuclear Weapons Advisory Opinion, above note 5, para. 78.
9 ICRC Customary Law Study, above note 7, Rule 71.
10 AP I, Arts 48, 51, 52; ICRC Customary Law Study, above note 7, Rules 1, 7.

- acts or threats of violence the primary purpose of which is to spread terror among the civilian population are prohibited, including when carried out through cyber means or methods of warfare;[11]
- indiscriminate attacks, namely those of a nature to strike military objectives and civilians or civilian objects without distinction, are prohibited, including when using cyber means or methods of warfare;[12]
- disproportionate attacks are prohibited, including when using cyber means or methods of warfare. Disproportionate attacks are those which may be 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.[13]
- during military operations, including when using cyber means or methods of warfare, constant care must be taken to spare the civilian population and civilian objects; all feasible precautions must be taken to avoid or at least minimize incidental civilian harm when carrying out attacks, including through cyber means and methods of warfare;[14]
- attacking, destroying, removing or rendering useless objects indispensable to the survival of the population is prohibited, including through cyber means and methods of warfare;[15]
- medical services must be protected and respected, including when carrying out cyber operations during armed conflicts.[16]

In addition, all feasible precautions must also be taken to protect civilians and civilian objects against the effects of attacks conducted through cyber means and methods of warfare, which is an obligation that States must already implement in peacetime.[17] Measures that could be considered include, among others: segregating military from civilian cyber infrastructure and networks; segregating computer systems on which essential civilian infrastructure depends from the

11 AP I, Art. 51(2); ICRC Customary Law Study, above note 7, Rule 2.

12 AP I, Art. 51(4); ICRC Customary Law Study, above note 7, Rules 11, 12. Indiscriminate attacks are those: (a) which are not directed at a specific military objective; (b) which employ a method or means of combat which cannot be directed at a specific military objective; or (c) which employ a method or means of combat the effects of which cannot be limited as required by international humanitarian law; and consequently, in each such case, are of a nature to strike military objectives and civilians or civilian objects without distinction.

13 AP I, Arts 51(5)(b), 57; ICRC Customary Law Study, above note 7, Rule 14.

14 AP I, Art. 57; ICRC Customary Law Study, above note 7, Rules 15–21.

15 AP I, Art. 54; Protocol Additional (II) to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts, 1125 UNTS 609, 8 June 1977 (entered into force 7 December 1978) (AP II), Art. 14; ICRC Customary Law Study, above note 7, Rule 54.

16 See, for instance, Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed forces in the Field of 12 August 1949, 75 UNTS 31 (entered into force 21 October 1950), Art. 19; Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of the Armed Forces at Sea of 12 August 1949, 75 UNTS 85 (entered into force 21 October 1950), Art. 12; Geneva Convention (IV) relative to the Protection of Civilian Persons in Time of War of 12 August 1949, 75 UNTS 287 (entered into force 21 October 1950), Art. 18; AP I, Art. 12; AP II, Art. 11; ICRC Customary Law Study, above note 7, Rules 25, 28, 29.

17 AP I, Art. 58; ICRC Customary Law Study, above note 7, Rules 22, 24.

Internet; and working on the identification in cyberspace of the cyber infrastructure and networks serving specially protected objects like hospitals.[18]

5. The need to discuss how IHL applies

Affirming that IHL applies to cyber operations in armed conflict is an essential first step to avoiding or minimizing the potential human suffering that cyber operations might cause. However, the ICRC also encourages States to work towards a common understanding of *how* IHL principles and rules apply to cyber operations. This is necessary because the interconnected nature of cyberspace and its largely digital character pose challenges for the interpretation of key IHL principles and concepts on the conduct of hostilities.

Among the various issues, in this position paper the ICRC emphasizes three.

The military use of cyberspace and the effect on its civilian character

Except for some specific military networks, cyberspace is predominantly used for civilian purposes. However, civilian and military networks may be interconnected. Furthermore, military networks may rely on civilian cyber infrastructure, such as undersea fibre-optic cables, satellites, routers or nodes. Conversely, civilian vehicles, shipping and air traffic controls increasingly rely on navigation satellite systems that may also be used by the military. Civilian logistical supply chains and essential civilian services use the same web and communication networks through which some military communications pass.

Not every use for military purposes renders a civilian object a military objective under IHL.[19] If it does, however, the object is no longer protected by the prohibition against direct attacks on civilian objects. It would be a matter of serious concern if the military use of cyberspace led to the conclusion that many objects forming part thereof would no longer be protected as civilian objects. This could lead to large-scale disruption of the ever-increasingly important civilian usage of cyberspace.

This being said, even if certain parts of the cyberspace infrastructure were no longer protected as civilian objects during armed conflicts, any attack would remain governed by the prohibition on indiscriminate attacks and the rules of proportionality and precautions in attack. Precisely because civilian and military networks are so interconnected, assessing the expected incidental civilian harm of

18 ICRC Challenges Report 2015, above note 4, p. 43.

19 See AP I, Art. 52(2); ICRC Customary Law Study, above note 7, Rule 8: "In so far as objects are concerned, military objectives are limited to those objects which by their nature, location, purpose or use make an effective contribution to military action and whose partial or total destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage." For more details on the limits to cyber infrastructure becoming a military objective under IHL, see ICRC Challenges Report 2015, above note 4, p. 42.

any cyber operation is critical to ensuring that the civilian population is protected against its effects.[20]

The notion of "attack" under IHL and cyber operations

Critical civilian infrastructure enabling the provision of essential services increasingly relies on digitalized systems. Safeguarding such infrastructure and services against cyber attacks or incidental damage is essential to protect the civilian population.

IHL provides specific protection for certain infrastructure, such as medical services and objects indispensable to the survival of the population, regardless of the type of harmful operation.[21] However, most rules stemming from the principles of distinction, proportionality and precaution – which provide general protection for civilians and civilian objects – apply only to military operations that qualify as "attacks" as defined in IHL.[22] Article 49 of Additional Protocol I defines attacks as "acts of violence against the adversary, whether in offence or in defence".[23] The question of how widely or narrowly the notion of "attack" is interpreted with regard to cyber operations is therefore essential for the applicability of these rules and the protection they afford to civilians and civilian infrastructure.

It is widely accepted that cyber operations expected to cause death, injury or physical damage constitute attacks under IHL. In the ICRC's view, this includes harm due to the foreseeable direct and indirect (or reverberating) effects of an attack, for example the death of patients in intensive care units caused by a cyber operation on an electricity network that results in cutting off a hospital's electricity supply.

Beyond this, attacks that significantly disrupt essential services without necessarily causing physical damage constitute one of the most important risks for civilians. Diverging views exist, however, on whether a cyber operation that results in a loss of functionality without causing physical damage qualifies as an attack as defined in IHL. In the ICRC's view, during an armed conflict an operation designed to disable a computer or a computer network constitutes an attack under IHL, whether the object is disabled through kinetic or cyber means.[24] If the notion of attack is interpreted as only referring to operations that cause death, injury or physical damage, a cyber operation that is directed at making a civilian network

20 See ICRC, *The Principle of Proportionality in the Rules Governing the Conduct of Hostilities under International Humanitarian Law*, Geneva, 2018, pp. 37–40, available at: www.icrc.org/en/download/file/79184/4358_002_expert_meeting_report_web_1.pdf.

21 See text in relation to above notes 16 and 15. With regard to the latter, they must not be attacked, destroyed, removed or rendered useless.

22 The notion of attack under IHL, defined in Article 49 of AP I, is different from and should not be confused with the notion of "armed attack" under Article 51 of the UN Charter, which belongs to the realm of *jus ad bellum*. To affirm that a specific cyber operation, or a type of cyber operations, amounts to an attack under IHL does not necessarily mean that it would qualify as an armed attack under the UN Charter.

23 For rules that apply specifically to attacks, see text in relation to above notes 10–14.

24 See ICRC Challenges Report 2011, above note 4, p. 37; ICRC Challenges Report 2015, above note 4, pp. 41–42.

(such as electricity, banking or communications) dysfunctional, or is expected to cause such effect incidentally, might not be covered by essential IHL rules protecting the civilian population and civilian objects. Such an overly restrictive understanding of the notion of attack would be difficult to reconcile with the object and purpose of the IHL rules on the conduct of hostilities. It is therefore essential that States find a common understanding in order to adequately protect the civilian population against the effects of cyber operations.

Civilian data and the notion of "civilian objects"

Essential civilian data – such as medical data, biometric data, social security data, tax records, bank accounts, companies' client files or election lists and records – are an essential component of digitalized societies. Such data are key to the functioning of most aspects of civilian life, be it at the individual or societal level. There is increasing concern about safeguarding such essential civilian data.

Some of the specific protection afforded by IHL extends to essential data, such as data belonging to medical units, which are encompassed in the obligation to respect and protect such units.[25]

More generally, the main IHL principles and rules governing the conduct of hostilities protect civilians and civilian objects.[26] It would therefore be important for States to agree on an understanding that civilian data is protected by these rules.

Deleting or tampering with essential civilian data can quickly bring government services and private businesses to a complete standstill. Such operations could cause more harm to civilians than the destruction of physical objects. While the question of whether and to what extent civilian data constitute civilian objects remains unresolved, in the ICRC's view the assertion that deleting or tampering with such essential civilian data would not be prohibited by IHL in today's data-reliant world seems difficult to reconcile with the object and purpose of IHL. The replacement of paper files and documents with digital files in the form of data should not decrease the protection that IHL affords to them.[27] Excluding essential civilian data from the protection afforded by IHL to civilian objects would result in an important protection gap.

6. Attribution of conduct in cyberspace for the purposes of State responsibility

Cyberspace provides various technical possibilities for actors to hide or falsify their identity, which increases the complexity of attribution by other actors. This creates major difficulties. For example, even during armed conflict, IHL only applies to

25 See above note 16.
26 See text in relation to above notes 10–15.
27 ICRC Challenges Report 2015, above note 4, p. 43; ICRC Challenges Report 2019, above note 4, p. 21.

operations that are linked to the conflict. If the author of a cyber operation – and thus the link of the operation to an armed conflict – cannot be identified, it may be difficult to determine whether IHL is even applicable to the operation. Attribution of cyber operations is also important to ensure that actors who violate international law, including IHL, can be held accountable. The perception that it will be easier to deny responsibility for such attacks may also weaken the taboo against their use – and may make actors less scrupulous about using them in violation of international law.[28]

This being said, attribution is not a problem from the perspective of the actors who conduct, direct or control cyber operations: they have all the facts at hand to determine under which international legal framework they are operating and which obligations they must respect.

Under international law, a State is responsible for conduct attributable to it, including possible violations of IHL. This includes:

- conduct by organs of the State, including its armed forces or intelligence services;
- conduct by persons or entities, such as private companies, that the State has empowered to exercise elements of governmental authority;
- conduct by persons or groups, such as militias or groups of hackers, acting in fact on the State's instructions, or under its direction or control; and
- conduct by private persons or groups which the State acknowledges and adopts as its own conduct.[29]
- These principles apply whether the conduct is carried out by cyber or any other means.

7. Conclusion

The use of cyber operations as means or methods or warfare in an armed conflict poses a real risk of harm to civilians. For the protection of the civilian population and civilian infrastructure, it is critical to recognize that such operations do not occur in a legal vacuum. The ICRC urges all States to affirm that IHL applies to cyber operations during armed conflicts, on the understanding that such affirmation neither encourages the militarization of cyberspace nor legitimizes cyber warfare.

At the same time, the ICRC believes that further discussion – especially among States – is needed on how IHL should be interpreted and applied in cyberspace. There is a pressing need for such discussion because States that decide to develop or acquire cyber capabilities that qualify as weapons, means or methods of warfare – whether for offensive or defensive purposes – must ensure that these capabilities can be used in accordance with their obligations under

28 ICRC Challenges Report 2011, above note 4, p. 37; ICRC Challenges Report 2019, above note 4, p. 20.
29 See ICRC Customary Law Study, above note 7, Rule 149. See also International Law Commission, *Responsibility of States for Internationally Wrongful Acts*, 2001, in particular Arts 4–11.

IHL.[30] Discussion should be informed by an in-depth understanding of the development of military cyber capabilities, their potential human cost, and the protection afforded by existing law. States need to determine whether existing law is adequate and sufficient to address the challenges posed by the interconnected and largely digital character of cyberspace, or whether it needs adaptation to the specific characteristics of cyberspace. If new rules are to be developed to protect civilians against the effects of cyber operations or for other reasons, they should build on and strengthen the existing legal framework – including IHL.

The ICRC welcomes the intergovernmental discussions currently taking place in the framework of two United Nations General Assembly mandated processes and it is grateful for the opportunity to share its views with the participating States. The ICRC also stands ready to lend its expertise to such discussions, as States deem appropriate.

30 See ICRC Challenges Report 2019, above note 4, pp. 28–29; 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*, Geneva, 2006, p. 4; AP I, Art. 36.

International Review of the Red Cross (2020), **102** (913), 495–509.
Digital technologies and war
doi:10.1017/S181638312000048X

IRRC

REPORTS AND DOCUMENTS

Reports

This section of the Review *provides a summary of new ICRC-affiliated reports relating to this issue's theme of "Digital Technologies and War", including the executive summaries of three such reports. For access to the full reports, please follow the links provided.*

: : : : : : :

Symposium Report: Digital Risks in Armed Conflicts, October 2019

This report summarizes the key findings and action points coming out of a symposium on digital risks in armed conflicts and other situations of violence, held in December 2018 by the International Committee of the Red Cross (ICRC). The two-day event brought together representatives from humanitarian organizations, academia, tech companies and governments, as well as donor representatives. Discussions focused on how the use of digital technologies, including by parties to conflict and private companies, but also by the humanitarian sector as part of a humanitarian response, could put crisis-affected people at risk and make them more vulnerable both on- and offline.

Available at: https://shop.icrc.org/symposium-report-digital-risks-in-armed-conflicts-print-en.

The Humanitarian Metadata Problem: "Doing No Harm" in the Digital Era, October 2018

New technologies continue to present great risks and opportunities for humanitarian action. To ensure that their use does not result in any harm, humanitarian organizations must develop and implement appropriate data protection standards, including robust risk assessments. However, this requires a good understanding of what these technologies are, what risks are associated with

their use, and how we can try to avoid or mitigate those risks. This joint report by Privacy International and the ICRC aims to provide people who work in the humanitarian sphere with the knowledge they need to understand the risks involved in the use of certain new technologies. The report also discusses the "do no harm" principle and how it applies in a digital environment.

Available at: www.icrc.org/en/download/file/85089/the_humanitarian_metadata_problem_-_icrc_and_privacy_international.pdf.

Handbook on Data Protection in Humanitarian Action, Second Edition, May 2020

This handbook was published as part of the Brussels Privacy Hub and ICRC's Data Protection in Humanitarian Action project. It is aimed at the staff of humanitarian organizations involved in processing personal data as part of humanitarian operations, particularly those in charge of advising on and applying data protection standards. The handbook builds on existing guidelines, working procedures and practices established in humanitarian action in the most volatile environments and for the benefit of the most vulnerable victims of humanitarian emergencies. It seeks to help humanitarian organizations comply with personal data protection standards, by raising awareness and providing specific guidance on the interpretation of data protection principles in the context of humanitarian action, particularly when new technologies are employed.

Available at: https://shop.icrc.org/handbook-on-data-protection-in-humanitarian-action-print-en.

The Potential Human Cost of Cyber Operations, May 2019

Executive summary

Cyber operations during armed conflicts: Assessing the challenges for international humanitarian law

The use of cyber operations during armed conflicts is a reality. While only a few States so far have publicly acknowledged that they use them, cyber operations are a known feature of present-day military operations and the use of them is likely to increase in the future.

This new reality has triggered a debate regarding the rules of international law that apply to such operations. In this debate, the ICRC has recalled that during armed conflict, cyber operations are subject to the rules of international humanitarian law (IHL).[1] It is nevertheless clear that cyberspace and these new

1 See, in particular, ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, Geneva, 2015, pp. 39–44. The restrictions imposed by IHL do not legitimize the use of force in cyberspace, which remains governed by the United Nations Charter.

military operations raise a number of questions as to precisely how certain rules of IHL – which were drafted primarily with the kinetic realm in mind – apply to cyber operations.

Assessing these questions requires an understanding of the expected use and military potential of cyber technology. What aims may belligerents want to achieve by using new tools at the strategic, operational or tactical levels during conflicts? How does this new technology compare to other, existing means of warfare?

Furthermore, to assess how IHL protects civilians in armed conflict, and whether further regulation is needed, lawyers and policy-makers require an understanding of the actual or potential human cost of cyber technologies. Indeed, one of the main aims of IHL is to protect civilians from the effects of military operations.

Purpose and scope of the meeting on IHL and cyber operations

As part of its mandate to work for the clarification of IHL and, if necessary, prepare any development thereof, the ICRC monitors the development of new technologies that are, or could be, used as means and methods of warfare during armed conflicts. This approach is based on legal, technical, military and humanitarian considerations, which are interrelated.

To develop a realistic assessment of cyber capabilities and their potential humanitarian consequences in light of their technical characteristics, the ICRC brought together scientific and cyber security experts from all over the world to share their knowledge about the technical possibilities, expected use and potential effects of cyber operations. The three-day meeting drew on the expertise of participants working for global IT companies, cyber threat intelligence companies, computer emergency response teams, a national cyber security agency, participants with expertise in cyber security (including that of hospitals, electrical grids and other services), participants with expertise in the development and use of military cyber operations, lawyers and academics.

States and militaries remain reluctant to disclose their cyber capabilities, including the details of cyber operations conducted in the context of armed conflicts, and little is known about the few acknowledged cases. Therefore, the experts discussed a number of the most sophisticated known cyber operations, regardless of whether they occurred in the context of an armed conflict or in peacetime. Examining the technical features of these attacks and the specific vulnerabilities of the respective targets provides a powerful evidence base for what is technically possible during armed conflict. The meeting focused in particular on the risk that cyber operations might cause death, injury or physical damage, affect the delivery of essential services to the population, or affect the reliability of internet services. It looked at the specific characteristics of cyber tools, how cyber threats have evolved, and the cyber security landscape.

Approaching the subject from a humanitarian law and humanitarian action perspective, the ICRC seeks a sober and – to the greatest extent possible – evidence-

based understanding of the risks of cyber attacks[2] for the civilian population. The meeting allowed the ICRC to confirm much of its own research, and to supplement it with highly valuable additional expert knowledge. The meeting was extremely useful in that it contributed to a nuanced picture of cyber operations, demystifying some of the assumptions that often surround discussions on cyber warfare.

Areas of concern

Discussions helped to put the spotlight on four areas of particular concern in terms of the potential human cost of cyber operations:

a. the specific vulnerabilities of certain types of infrastructure;
b. the risk of overreaction due to potential misunderstanding of the intended purpose of hostile cyber operations;
c. the unique manner in which cyber tools may proliferate;
d. the obstacles that the difficulty of attributing cyber attacks create for ensuring compliance with international law.

a. Specific vulnerabilities of certain types of infrastructure: Cyber attacks that may affect the delivery of health care, industrial control systems, or the reliability or availability of core internet services

Apart from causing substantial economic loss, cyber operations can harm infrastructure in at least two ways. First, they can affect the delivery of essential services to civilians, as has been shown with cyber attacks against electrical grids and the health-care sector. Second, they can cause physical damage, as was the case with the Stuxnet attack against a nuclear enrichment facility in Iran in 2010, and an attack on a German steel mill in 2014.

Cyber attacks that may affect the delivery of health care

The health-care sector is moving towards increased digitization and interconnectivity. For example, hospital medical devices are normally connected to the hospital's information technology (IT) system to enable automatic electronic filing. Connected biomedical devices, such as pacemakers and insulin pumps, make it possible to remotely monitor individual patients' health as well as the functioning of the medical devices themselves.

This increased digital dependency, combined with an increased 'attack surface', has not been matched by a corresponding improvement in cyber security. Consequently, this infrastructure is particularly vulnerable, with potentially serious consequences for health and life.

2 The terms "cyber attacks" and "cyber operations" are used throughout the report in a technical (mainstream or colloquial) sense and not as they may be understood under IHL, unless specifically stated.

Cyber attacks against industrial control systems, including those used in critical civilian infrastructure

Industrial control systems are protected by complex safety mechanisms and often have built-in redundancy to guarantee safety and reliability. For example, electrical networks are grids with multiple power sources to avoid widespread effects when one of their parts is affected. Nonetheless, attacks on specific nodes might still cause a significant impact, such as if a critical system (like a hospital) depends on a specific sub-system or node, or because they have cascading harmful consequences.

Carrying out a cyber attack against an industrial control system requires a certain expertise and sophistication, and often, custom-made malware. Such attacks have been less frequent so far than other types of cyber operations. Nonetheless, their frequency is reportedly increasing, and the severity of the threat has evolved more rapidly than was anticipated only a few years ago. There is a risk that tools developed by the best-resourced actors may be repurposed or purchased by other actors who lack the expertise required to develop them from scratch. Moreover, there is a possibility that a number of undetected actors are capable of attacking industrial control systems.

Cyber attacks that may affect the reliability or availability of internet services

Cyber attacks that disrupt core internet services – such as the domain name system (DNS), which supports communications on the Internet – or disrupt the functioning of major cloud services may impact all services that rely on them. However, the risk of seriously compromising these core internet services was assessed by the experts as unlikely at the present moment thanks to the high degree of redundancy in the DNS and because major cloud providers tend to offer high security standards. If, however, such disruption were to occur, it could have widespread and potentially serious consequences, for example when life-saving services such as ambulances rely on the cloud.

Finally, distributed denial-of-service (DDoS) attacks have been used against services provided by governments for the population. Such attacks are carried out through increasingly large botnets. The arrival of the "Internet of things" will further increase the number of connected devices that could be used in such attacks. Furthermore, DDoS attacks might have a wider impact than expected by their author, in particular when information about the targeted network is incomplete.

b. Risk of overreaction due to the potential misunderstanding of the intended purpose of hostile cyber operations

Cyber operations can be broadly divided into two categories, depending on their purpose:

- activity encompassing reconnaissance, surveillance and the exfiltration of data and information, for example for espionage, often referred to as computer network exploitation (CNE), or "access operations";

- activity aimed at generating effects on a targeted system or device, such as tampering with data integrity (deletion, modification), affecting availability (disabling, including for prolonged periods of time), or causing physical effects, such as damaging the system, often referred to as a computer network attack (CNA), or "effects operations".

The distinction is primarily one of purpose. From a technical perspective, the initial steps of a CNE and a CNA to gain and maintain persistent access to the target may be identical. CNEs can then be turned into CNAs relatively simply, mostly through the use of specific payloads of a different nature. While the initial steps of the attacks may be tracked, it is often difficult to fully assess the attacker's purpose until the effect on the end target is actually achieved.

When the target does not know the actual purpose of the operation, its reaction may be to consider the potential worst-case impact that the attacker could achieve though a CNA and react in a stronger manner than it would have if it had known that the intended purpose of the attack was CNE. This escalation risk factor may give rise to a potentially harmful overreaction.

c. *Proliferation of cyber tools*

A third concern is the proliferation of cyber tools – an issue that in some respects raises concerns similar to those that may exist with regard to weapons proliferation or the proliferation of dual-use technology, although the specific nature of cyber tools must be taken into account.

Cyber tools and methods can proliferate in a unique manner that is difficult to control. First, cyberspace is a global domain: provided that the attacker can overcome the cyber security and defence measures in place, any network node and information residing on the network can be accessed from anywhere in the world. At the same time, cyber tools can be repurposed or re-engineered. The combination of these two characteristics means that when cyber tools are used, stolen, leaked or otherwise become available, actors other than those who developed them might be able to find them, reverse engineer them, and reuse them for their own purposes.

Finally, the fact that cyber tools and methods can be repurposed and reused is one of the factors making rapid and reliable technical attribution of cyber attacks a challenging process.

d. *Attribution of attacks*

While not a primary focus of the meeting, the discussions also touched upon the anonymity of attacks and the difficulty of attributing them to a specific actor, which is a fourth area of concern.

Cyberspace is a complex domain where multiple actors operate: individual hackers; criminal groups, potentially motivated by financial gain; States; non-State armed groups; and other non-State actors. Actors may also cooperate: for

example, States may buy cyber tools or have an operation performed on their behalf against a target they have identified.

Digital forensics and the capabilities of attribution of malicious cyber activity appear to be improving. Nonetheless, the ability of threat actors to obscure or effectively hide the origin of their operations on the Internet, compounded by the ability to buy, repurpose or re-engineer cyber tools developed or used by other actors, continues to make it difficult to rapidly and reliably attribute cyber attacks to a specific actor. This hampers the possibility of identifying actors who violate IHL in cyberspace and holding them responsible. This is a concern because to hold such actors responsible is one way to ensure compliance with IHL. It may also lower the threshold of using cyber attacks and of using them in violation of international law, because attackers can deny responsibility.

Cyber operations during armed conflicts: Implications for international humanitarian law

It is well-established that international law applies to cyber operations. More specifically, IHL and its principles of distinction, proportionality, precaution, military necessity and humanity restrict the use of cyber means and methods during armed conflict. Further discussions may however be needed to clarify how IHL applies and whether it is adequate and sufficient or requires further development, building on existing law.

The meeting helped to clarify which areas of humanitarian concern should be the focus of attention. In brief, based on the detailed knowledge available of cyber operations during peacetime, and the somewhat lesser knowledge of cyber operations in times of armed conflict, the following picture emerges.

Distinction in cyber space

First, cyber attacks are not necessarily indiscriminate. As the report illustrates in more detail, cyber tools can be designed to self-propagate or not. Even if they self-propagate and cause cyber security concerns for all those infected, they can be designed to only cause damage to a specific target. While some self-propagating malware that caused indiscriminate harmful effects has made headlines, many cyber operations have in fact been rather discriminate from a technical perspective (which does not mean they were lawful).

Furthermore, certain types of cyber attacks, such as those that would aim to cause physical damage to industrial control systems, require custom-made cyber tools. In many cases this would also effectively hamper the ability to carry such attacks out in a large-scale, indiscriminate manner.

This is important from an IHL perspective, because contrary to the assumption often heard that the principle of distinction might have become meaningless in cyberspace because of the interconnectivity that characterizes it, not all offensive cyber tools are inherently indiscriminate. On the contrary, they may well be very precisely tailored and create effects on specific targets only.

Highlighting the potential human cost

Secondly, and of equal importance, it is nonetheless clear that cyber tools can cause substantial damage and can be – and have sometimes been – indiscriminate, and that certain systems are particularly at risk, first and foremost, perhaps, health-care systems. Moreover, the threats that can be observed have been evolving faster than anticipated, in particular regarding attacks against industrial systems. Finally, much is still unknown in terms of the rapid evolution of the technology, the capabilities and the tools developed by the most sophisticated actors, and the extent to which the increased use of cyber operations during armed conflicts might be different from the trends observed so far. In other words, while the risk of human cost based on current observations does not appear extremely high, especially considering the destruction and suffering that conflicts always cause, the evolution of cyber operations still merits close attention due to existing uncertainties and the rapid pace of change.

Legal protection through IHL

Many of the attacks described in the report targeted or indiscriminately affected civilian infrastructure. In the view of the ICRC, if carried out in times of armed conflict, such attacks would be prohibited. First of all, direct attacks against civilian infrastructure and indiscriminate attacks would be prohibited. Secondly, even if the infrastructure or some parts of it had become military objectives (such as a part of an electricity grid), IHL would require that only this part be attacked, and that there be no excessive damage to the remaining civilian parts. Thirdly, IHL would require parties to the conflict to take all feasible precautions to avoid or at least minimize incidental harm to civilians and civilian objects. Finally, even when they do not amount to attacks under IHL,[3] such operations might be prohibited by the specific protection afforded by IHL to medical facilities or objects indispensable to the survival of the population. These are powerful protections that remain entirely relevant in view of the technical characteristics of cyber operations. For IHL to truly provide legal protection to civilians against the effects of cyber warfare, however, States must commit to its applicability and to an interpretation of its rules that is effective for the protection of civilians and civilian infrastructure. In particular, it would require a clear recognition that cyber operations which impair the functionality of civilian infrastructure are subject to the rules governing attacks under IHL.[4] This report will hopefully help to illustrate the need for such an interpretation in order to ensure that civilian infrastructure is protected.

3 Under IHL, "attack" has a specific meaning which would not encompass all cyber operations that are referred to as cyber attacks in a colloquial sense.

4 See ICRC, above note 1, p. 41.

Avenues that could be explored to reduce the potential human cost of cyber operations

Cyber security measures

Beyond the restraints imposed by IHL upon those carrying out cyber operations, it is critical to enhance the cyber security posture and resilience of the actors potentially affected. While cyber security and defence are constantly improving, older systems with outdated or even nonexistent cyber security are particularly vulnerable to cyber attacks and will remain a concern in the years to come. Both the public and private sectors have a role to play through industry standards and legal regulation.

In the health-care sector, for instance, the regulatory environment should be adapted to the increased risk, such as through standardization requirements, with a view to ensuring resilience in the event of a cyber attack. Cyber security needs to be taken into account in the design and development of medical devices and updated throughout their lifetime, no matter how long they last. Similarly, for industrial control systems, industry standards, whether imposed or self-imposed, are critical. This includes reporting incidents and sharing information between trusted partners.

In terms of IHL, parties to armed conflicts must take all feasible precautions to protect civilians and civilian objects under their control against the effects of attack. This is one of the few IHL obligations that States must already implement in peacetime.

Disclosing vulnerabilities

The preferred option for enhancing the safety of cyberspace should be disclosing vulnerabilities to the appropriate software developer so that those vulnerabilities can be fixed. Some States have therefore put in place equity processes to balance competing interests and risks and decide whether to disclose the vulnerabilities they identify.

Measures to prevent proliferation

Those who develop cyber weapons should consider creating obstacles in order to make repurposing difficult and expensive. While it is hardly possible from a technical standpoint to guarantee that malware cannot be repurposed, methods like encrypting its payload and including obstacles in different components of the code, for example, could raise the bar in terms of the expertise required to re-engineer malicious tools. While there is currently no express obligation under IHL to create obstacles to the repurposing of cyber tools, this could prevent at least some actors from doing so and therefore reduce the risk of subsequent misuse that their proliferation entails. The unique way in which cyber tools proliferate also raises the question of whether existing law is adequate or sufficient to address this phenomenon.

Marking of certain civilian infrastructure

Another avenue, which builds on existing international law, could be to create a "digital watermark" to identify certain actors or infrastructure in cyberspace that must be protected (such as objects that enjoy specific protection under IHL). The aim would be to help their identification and prevent them from being targeted during armed conflicts. The potentially positive effects in terms of protection against unintended harm by law-abiding actors would however need to be balanced against the risk of disclosing information on critical infrastructure to potential adversaries, including criminals. The prospects of positive effects might depend in part on attribution becoming easier.

Improving attribution and accountability

Finally, enhanced attribution capacities would help ensure that actors who violate international law in cyberspace can be held accountable, which is a means to strengthen compliance with the law and more generally encourage responsible behaviour in cyberspace.

Way forward

The use of cyber operations in armed conflict is likely to continue and might remain shrouded in secrecy. Analyzing its consequences is a complex and long-term endeavour that requires multidisciplinary expertise and interaction with a wide variety of stakeholders.

Building upon the conclusions reached at the expert meeting, the ICRC would like to pursue dialogue with governments, experts and the IT sector. It looks forward to the feedback to this report in order to continue to follow the evolution of cyber operations, in particular during armed conflicts, and their potential human cost, to explore avenues that could reduce them, and to work towards a consensus on the interpretation of existing IHL rules, and potentially the development of complementary rules that afford effective protection to civilians.

Available at: www.icrc.org/en/document/potential-human-cost-cyber-operations.

***Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, August 2019**

Executive summary

The ICRC has emphasized the need to maintain human control over weapon systems and the use of force, to ensure compliance with international law and to satisfy ethical concerns. This approach has informed the ICRC's analysis of the legal, ethical, technical and operational questions raised by autonomous weapon systems.

In June 2018, the ICRC convened a round-table meeting with independent experts in autonomy, artificial intelligence (AI) and robotics to gain a better understanding of the technical aspects of human control, drawing on experience with civilian autonomous systems. This report combines a summary of the discussions at that meeting with additional research, and highlights the ICRC's main conclusions, which do not necessarily reflect the views of the participants. Experience in the civilian sector yields insights that can inform efforts to ensure meaningful, effective and appropriate human control over weapon systems and the use of force.

Autonomous (robotic) systems operate without human intervention, based on interaction with their environment. These systems raise such questions as "How can one ensure effective human control of their functioning?" and "How can one foresee the consequences of using them?" The greater the complexity of the environment and the task, the greater the need for direct human control and the less one can tolerate autonomy, especially for tasks and in environments that involve risk of death and injury to people or damage to property – in other words, safety-critical tasks.

Humans can exert some control over autonomous systems – or specific functions – through supervisory control, meaning "human-on-the-loop" supervision and the ability to intervene and deactivate. This requires the operator to have:

- situational awareness;
- enough time to intervene;
- a mechanism through which to intervene (a communication link or physical controls) in order to take back control, or to deactivate the system should circumstances require.

However, human-on-the-loop control is not a panacea, because of such human–machine interaction problems as automation bias, lack of operator situational awareness and the moral buffer.

Predictability and reliability are at the heart of discussions about autonomy in weapon systems, since they are essential to achieving compliance with IHL and avoiding adverse consequences for civilians. They are also essential for military command and control.

It is important to distinguish between reliability – a measure of how often a system fails; and predictability – a measure of how the system will perform in a particular circumstance. Reliability is a concern in all types of complex system, whereas predictability is a particular problem with autonomous systems. There is a further distinction between predictability in a narrow sense of knowing the process by which the system functions and carries out a task, and predictability in a broad sense of knowing the outcome that will result.

It is difficult to ensure and verify the predictability and reliability of an autonomous (robotic) system. Both factors depend not only on technical design but also on the nature of the environment, the interaction of the system with that environment, and the complexity of the task. However, setting boundaries or

imposing constraints on the operation of an autonomous system – in particular on the task, the environment, the time frame of operation and the scope of operation over an area – can render the consequences of using such a system more predictable.

In a broad sense, all autonomous systems are unpredictable to a degree because they are triggered by their environment. However, developments in the complexity of software control systems – especially those based on AI and machine learning – add unpredictability in the narrow sense that the process by which the system functions is unpredictable.

The "black box" manner in which many machine learning systems function makes it difficult – and in many cases impossible – for the user to know how the system reaches its output. Not only are such algorithms unpredictable but they are also subject to bias, whether by design or in use. Furthermore, they do not provide explanations for their outputs, which seriously complicates establishing trust in their use and exacerbates the already significant challenges of testing and verifying the performance of autonomous systems. And the vulnerability of AI and machine learning systems to adversarial tricking or spoofing amplifies the core problems of predictability and reliability.

Computer vision and image recognition are important applications of machine learning. These applications use deep neural networks (deep learning), of which the functioning is neither predictable nor explainable, and such networks can be subject to bias. More fundamentally, machines do not see like humans. They have no understanding of meaning or context, which means they make mistakes that a human never would.

It is significant that industry standards for civilian safety-critical autonomous robotic systems – such as industrial robots, aircraft autopilot systems and self-driving cars – set stringent requirements regarding human supervision, intervention and deactivation, or fail-safe; predictability and reliability; and operational constraints. Leading developers of AI and machine learning have stressed the need to ensure human control and judgement in sensitive applications – and to address safety and bias – especially where applications can have serious consequences for people's lives.

Civilian experience with autonomous systems reinforces and expands some of the ICRC's viewpoints and concerns regarding autonomy in the critical functions of weapon systems. The consequences of using autonomous weapon systems are unpredictable because of uncertainty for the user regarding the specific target, and the timing and location of any resulting attack. These problems become more pronounced as the environment or the task become more complex, or freedom of action in time and space increases. Human-on-the-loop supervision and intervention and the ability to deactivate are absolute minimum requirements for countering this risk, but the system must be designed to allow for meaningful, timely, human intervention – and even that is no panacea.

All autonomous weapon systems will always display a degree of unpredictability stemming from their interaction with the environment. It might be possible to mitigate this to some extent by imposing operational constraints on the task, the time frame of operation, the scope of operation over an area and the

environment. However, the use of software control based on AI – and especially machine learning, including applications in image recognition – brings with it the risk of inherent unpredictability, lack of explainability and bias. This heightens the ICRC's concerns regarding the consequences of using AI and machine learning to control the critical functions of a weapon system and raises questions about their use in decision support systems for targeting.

This review of technical issues highlights the difficulty of exerting human control over autonomous (weapon) systems and shows how AI and machine learning could exacerbate this problem exponentially. Ultimately it confirms the need for States to work urgently to establish limits on autonomy in weapon systems.

Further, this review reinforces the ICRC's view that States should agree on the type and degree of human control required to ensure compliance with international law and to satisfy ethical concerns, while also underlining its doubts that autonomous weapon systems could be used in compliance with IHL in all but the narrowest of scenarios and the simplest of environments.

Available at: www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control, June 2020

ICRC and SIPRI, by Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson

Executive summary[5]

The challenges posed by autonomous weapon systems (AWS) are the focus of an intergovernmental discussion under the framework of the United Nations Convention on Certain Conventional Weapons (CCW). Despite enduring disagreements on whether additional regulation is needed, and in what form, there is emerging consensus among States that autonomy in weapon systems cannot be unlimited: humans must "retain" and "exercise" responsibility for the use of weapon systems and the use of force in armed conflict. This report explores the difficult question of how that principle must be applied in practice. It offers an in-depth discussion on the type and degree of control that humans need to exercise over AWS, in light of legal requirements, ethical concerns and operational considerations. It provides policy-makers with practical guidance on how measures for human control should form the basis of internationally agreed limits on AWS, whether rules, standards or best practices.

The report is the result of a joint project of the ICRC and the Stockholm International Peace Research Institute (SIPRI). Chapter 1 introduces the context and conceptual approach. Chapter 2 explores the legal, ethical and operational

5 This executive summary © SIPRI 2020, reproduced with permission.

perspectives on human control. Chapter 3 provides practical guidance on the type, degree and combination of control measures needed for compliance with IHL and to address ethical concerns, while taking into account military operational considerations. Chapter 4 presents the key findings and recommendations for policy-makers.

A core problem with AWS is that they are triggered by the environment, meaning that the user does not know, or choose, the specific target, timing and/or location of the resulting application of force. This process by which AWS function and the associated unpredictability in the consequences of their use can raise serious risks for civilians and challenges for compliance with IHL, as well as fundamental ethical concerns about the role of humans in life-and-death decisions, and challenges for military command and control.

A key question, therefore, is what limits are needed on AWS to address these challenges. An examination of the legal, ethical and operational requirements for human control indicates the need for a combination of three types of control measures:

1. *Controls on the weapon system's parameters of use*, including measures that restrict the type of target and the task the AWS is used for; place temporal and spatial limits on its operation; constrain the effects of the AWS; and allow for deactivation and fail-safe mechanisms.
2. *Controls on the environment*, namely, measures that control or structure the environment in which the AWS is used (e.g. using the AWS only in environments where civilians and civilian objects are not present, or excluding their presence for the duration of the operation).
3. *Controls through human–machine interaction*, such as measures that allow the user to supervise the AWS and to intervene in its operation where necessary.

These control measures can help to reduce or at least compensate for the unpredictability inherent in the use of AWS and to mitigate the risks involved, in particular for civilians. From a legal perspective, a user must exercise sufficient control to have reasonable certainty about the effects of an AWS when used in an attack and to be able to limit them as required by IHL. Ethical considerations may demand additional constraints, especially given concerns with AWS designed or used against persons.

The report concludes with five recommendations. First, States should focus their work on determining how measures needed for human control apply in practice. Since these three types of control measures are not tied to specific technologies, they provide a robust normative basis applicable to the regulation of both current and future AWS.

Second, measures for human control should inform any development of internationally agreed limits on AWS, whether new rules, standards or best practices. This work must be guided by the legal, ethical and operational requirements for human control. Any normative development should also focus on human obligations and responsibilities, not on technological fixes, so as to remain relevant and practical, and adaptable to future technological developments.

Third, States should clarify where IHL rules already set constraints on the development and use of AWS, and where new rules, standards and best practice guidance may be needed.

Fourth, any new rules, standards and best practices must build on existing limits on autonomy under IHL, and should draw on existing practice. It is likely that new rules, standards and best practice guidance can be most effectively articulated in terms of limits on specific types of AWS and on the manner and circumstances of their use, and requirements for human supervision and intervention.

Fifth, human control criteria should be considered in the study, research and development, and acquisition of new weapon systems.

Available at: www.icrc.org/en/document/limits-autonomous-weapons.

International
Review
of the Red Cross

The Review is a peer-reviewed journal printed in English and is published three times a year.

Annual selections of articles are also published in Arabic, Chinese, French, Russian and Spanish.

Published in association with Cambridge University Press.

Submission of manuscripts

The *International Review of the Red Cross* invites submissions of manuscripts on subjects relating to international humanitarian law, policy and action. Issues focus on particular topics, decided by the Editorial Board, which can be consulted under the heading 'Call for Papers' on the website of the *Review*. Submissions related to these themes are particularly welcome.

Articles may be submitted in Arabic, Chinese, English, French, Russian or Spanish. Selected submissions are translated into English if necessary.

Submissions must not have been published, submitted or accepted elsewhere. Articles are subjected to a peer-review process. The final decision on publication is taken by the Editor-in-Chief. The Review reserves the right to edit articles.

Manuscripts may be sent by e-mail to:
review@icrc.org

Manuscript requirements
Articles should be 7,000 to 10,000 words in length. Shorter contributions can be published as comments or opinion notes. Articles on themes other than the main theme of an edition may be published under the heading 'Selected articles on IHL and humanitarian action'.

For further information, please consult the website of the Review:
https://international-review.icrc.org/.

Subscriptions

The subscription price is available online at:
https://www.cambridge.org/core/journals/international-review-of-the-red-cross/subscribe

Requests for subscriptions can be made to the following address:

Cambridge University Press, University Printing House, Shaftesbury Road, Cambridge CB2 8BS; or in the USA, Canada and Mexico, email journals@cambridge.org: Cambridge University Press, 1 Liberty Plaza, Floor 20, New York, NY 10006, USA, email journals_subscriptions@cup.org.

The *International Review of the Red Cross* is indexed in the Thomson Reuters Journal Citation Reports/Social Sciences Edition and has an Impact Factor. The Review is available on LexisNexis.

This journal issue has been printed on FSC™-certified paper and cover board. FSC is an independent, nongovernmental, not-for-profit organization established to promote the responsible management of the world's forests. Please see www.fsc.org for information.

Printed in the UK by Bell & Bain Ltd.

Digital technologies and war

ICRC
ISSN 1816-3831

https://international-review.icrc.org/

Cambridge Core
For further information about this journal please go to the journal web site at:
cambridge.org/irc

Volume 102 Number 913

International **Review** of the Red Cross

CAMBRIDGE UNIVERSITY PRESS