‘Handle with Care’: Due Diligence Obligations in the Employment of AI Technologies

Accepted for publication in Robin Geiß and Henning Lahmann (eds.) 2024. Research Handbook on Warfare and Artificial Intelligence. Edward Elgar.

In R. Geiss, H. Lahmann, *Handbook on Warfare and Artificial Intelligence* (Edward Elgar, forthcoming 2023). Chapter written in July 2021

'HANDLE WITH CARE':
DUE DILIGENCE OBLIGATIONS IN THE EMPLOYMENT OF AI TECHNOLOGIES

Antonio Coco and Talita Dias

1. Introduction: Managing the Risks of AI Technology under International Law

Defined as the study and employment of computers to mimic certain human skills, such as perception, association, prediction, planning and motor control,[1] artificial intelligence (AI) technology has been hailed by many as a unique opportunity to address some of the world's most pressing challenges, from ending poverty and hunger to curing diseases. The immense potential of AI technology has already found application in numerous settings, both civilian and military. Civilian applications include, among others, law enforcement — such as facial recognition,[2] investigative support,[3] and other policing activities[4] — criminal justice,[5] online dispute resolution,[6] cybersecurity enhancement,[7] or even simple tasks like language translation. When thinking of military uses, the international lawyers' mind instinctively goes to autonomous weapons systems (AWS). Other instances, however, include computer sensors, navigational software, demining robots, digital content authentication tools, data and threat analysis, natural-language processing,[8] intelligence, surveillance and reconnaissance (ISR),[9] and cyber defence.[10]

Yet, with these opportunities also come risks, of which scientists, ethicists and lawyers have warned.[11] In particular, insufficient quality or quantity of — often biased — data used to design and train AI algorithms may lead to inaccurate or discriminatory outcomes. To give but one example, several machine learning image detection systems, including some manufactured by the biggest IT companies — IBM, Microsoft, Facebook, and Amazon — have been found to be

---

[1] Margaret A. Boden, *AI: Its Nature and Future* (OUP 2016) 1-3; Melanie Mitchell, *Artificial Intelligence: A Guide for Thinking Humans* (Pelican Books 2019) 7; 'Human Rights in the Age of Artificial Intelligence' (*Access Now*, 8 November 2018) <www.accessnow.org/human-rights-in-the-age-of-AI> accessed 10 June 2021, 8.

[2] Daragh Murray, 'Using Human Rights Law to Inform States' Decisions to Deploy AI' (2020) 114 AJIL Unbound 158, 158.

[3] eg Lindsay Freeman, 'Weapons of War, Tools of Justice: Using Artificial Intelligence to Investigate International Crimes' (2021) 19 Journal of International Criminal Justice <https://doi.org/10.1093/jicj/mqab013> accessed 6 September 2021.

[4] eg Marion Oswald et al, 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality' (2018) 27 Information & Communications Technology Law 223.

[5] eg the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm used by some US criminal jurisdictions.

[6] Ashley Deeks, 'High-Tech International Law' (2020) 88 George Washington Law Review 574, 587.

[7] Al Perlman, 'The Growing Role of Machine Learning in Cybersecurity' (*Security Roundtable)* <www.securityroundtable.org/the-growing-role-of-machine-learning-in-cybersecurity/> accessed 10 June 2021; Gordon Gottsegen, 'Machine Learning Cybersecurity: How it Works and Companies to Know' (*Built In*, 30 June 2019) <https://builtin.com/artificial-intelligence/machine-learning-cybersecurity> accessed 10 June 2021.

[8] National Security Commission on Artificial Intelligence (US), *Final Report* (2021) 68, 81, 114; 'Maintaining the Intelligence Edge: Reimagining and Reinventing Intelligence Through Innovation' (*CSIS*, January 2021) 8-22 <csis-website-prod.s3.amazonaws.com/s3fs-public/publication/210113_Intelligence_Edge.pdf> accessed 10 June 2021; 'Drones, Infrared Cameras and AI Join the Search for Mines' (*Inspired*, 16 June 2020) <https://blogs.icrc.org/inspired/2020/06/16/drones-infrared-cameras-mines/> accessed 10 June 2021.

[9] Ashley Deeks, Noam Lubell and Daragh Murray, 'Machine Learning, Artificial Intelligence, and the Use of Force by States' (2019) 10 Journal of National Security Law and Policy 1, 6; Steven Hill, 'AI's Impact on Multilateral Military Cooperation: Experience from NATO' (2020) 114 AJIL Unbound 147, 150.

[10] Hill (n 9) 150; Deeks, Lubell and Murray (n 9) 8.

[11] eg the four risks identified by Deeks, 'High-Tech International Law' (n 6) 641–643. Similarly, Lorna McGregor, Daragh Murray and Vivian Ng, 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68 International and Comparative Law Quarterly 309, 310.

significantly worse at recognising faces of female and non-white subjects.[12] This is because virtually all existing AI applications are narrow,[13] meaning they perform specific tasks whilst relying on statistical analysis of historical data, which in turn reproduces and amplifies societal biases.[14] This also means that current AI systems cannot be trained to deal with all possible scenarios they might be confronted with in the future. Coupled with the technology's lack of contextual and common-sense knowledge, this may lead to errors that a human being could easily avoid, such as a self-driving car's confusion between de-icing salt lines and lane markings.[15]

Furthermore, AI decisions are often unpredictable or inexplainable.[16] Not only may this raise doubts as to whether the relevant parameters are being respected, [17] but humans may also suffer from 'automation bias' and yield to the machine's decision.[18] In a pernicious example, teacher evaluation systems using AI and machine-learning have used inexplicable calculations based on predicted student test results to poorly rate and fire good teachers.[19] Effective development and employment of AI technology also requires the collection of a vast amount of potentially confidential, sensitive, or otherwise protected information — which could be misused, falsified, or manipulated[20] and comes at the cost of individual privacy regardless.[21]

With respect to military applications, despite AI's promise of superhuman accuracy, precision, and reduced casualties, side-lining human judgment in key decisions, such as target selection and engagement, may have disastrous human consequences. Many modern battlefields, such as urban areas, are dynamic, congested, or complex.[22] This is especially so in the case of non-international armed conflict, where confrontations permeate highly populated civilian areas, and combatants may not be easily distinguishable.[23] Add to this the fact that virtually all ongoing armed conflicts are happening in the developing world, which means that most combatants and civilians are not the same white people with which machine learning algorithms have been trained. As in peacetime, social media and messaging platforms can be used to spur chaos and unrest among civilians and combatants alike.[24] And how will self-driving tanks deal with unforeseen objects in their way? As others have noted, even if humans exercise supervisory control over AI-powered autonomous

---

[12] See Gender Shades <http://gendershades.org/overview.html> accessed 10 June 2021; Russell Brandom, 'Amazon's facial recognition matched 28 members of Congress to criminal mugshots' (*The Verge*, 26 July 2018) <www.theverge.com/2018/7/26/17615634/amazon-rekognition-aclu-mug-shot-congress-facial-recognition> accessed 10 June 2021; Access Now (n 1) 24. Likewise, human biases may have been encoded into the system algorithms. See Mitchell (n 1) 123-124; Access Now (n 1) 12; Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Penguin Books 2016) 154–155.

[13] Though some AI companies claim to be close to developing systems employing artificial general intelligence (AGI). See, e.g., Anthony Cuthbertson, "The Game is Over': Google's DeepMind says it is on verge of achieving human-level AI', *The Independent* (London, 23 May 2022) <https://www.independent.co.uk/tech/ai-deepmind-artificial-general-intelligence-b2080740.html.> accessed 01 February 2023.

[14] See also Access Now (n 1) 12

[15] Mitchell (n 1) 117-119.

[16] Also known as the 'black box' problem: see, e.g., Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015); Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI* (UNIDIR 2020).

[17] Especially for 'neural nets': Deeks, Lubell and Murray (n 9) 19.

[18] ibid 17. Also known as 'judgmental atrophy', since the human loses the habit of making decisions.

[19] O'Neil (n 12) 137–138.

[20] Deeks, 'High-Tech International Law' (n 6) 643.

[21] Access Now (n 1) 15-16, 20.

[22] ICRC, 'ICRC Position on Autonomous Weapon Systems' (2021) 9.

[23] ibid.

[24] ICRC, 'Misinformation, Disinformation and Hate Speech in Armed Conflict and Other Situations of Violence: A Practical Guide for Field Teams' (2021) 8, 11-12; Mercy Corps, 'The Weaponization of Social Media: How social media can spark violence and what can be done about it' (2019) 7, 16-19.

systems, what has been referred to as 'human-on/in-the-loop', they may lack the necessary time and situational awareness to intervene and regain control.[25]

Like more rudimentary automated weapons, such as anti-personal mines and anti-radar loitering weapons, an AWS might be predictable and reliable in the sense that we know, *in general or abstract terms*, what its triggers and results are.[26] However, precisely because of such generalised decision-making, these weapons are unpredictable when it comes to *specific* targets and circumstances.[27] Outside of ideal conditions, we do not know what objects or persons in the real world can trigger these weapons, which means that they can have significantly indiscriminate effects.[28] In addition, the 'deeper' the neural networks employed by these and other military AI systems, i.e. the more data they analyze and the more complex the probabilistic calculations they perform, the harder it will be for humans to understand how they operate in order to predict and rectify errors.[29] Unlike computer systems used in airplanes, such as the autopilot function, machine learning's obscure algorithms cannot be verified mathematically for accuracy and completeness, a problem that is compounded when it interacts with an unstable environment.[30] Furthermore, even autopilot systems have stringent requirements, such as constant human supervision, intervention, and deactivation mechanisms.[31]

Existing international law, despite its undeniable imperfections, is not blind to these risks. To the extent that they are 'technology-neutral', international obligations apply to all tools or technologies developed and employed by states over time, including weapons.[32] Thus, no one has seriously questioned that international law applies to AI technologies and their vast civilian and military applications. This is particularly the case with lethal AWS, which the states parties to the Convention on Certain Conventional Weapons (CCW) agreed are governed by international law, including international humanitarian law (IHL).[33]

In addition to *negative* duties to *refrain* from causing harm by employing AI technologies, states also have *positive* obligations to (pro)actively *protect* other states and individuals from any such harm. As in the case of environmental, health, kinetic, or cyber hazards, regardless of who causes it,[34] international law imposes several obligations on states to exercise due diligence in preventing, stopping, or redressing AI-generated harm. Whether or not individuals can be held accountable for the consequences of AI errors or malfunctions, states may be liable for failing to prevent, halt, or redress the resulting harm. At its core, due diligence is about the prevention and management of these and other risks or harms.

---

[25] ICRC, 'Autonomy, artificial intelligence and robotics: Technical aspects of human control' (2019) 9-10.
[26] ibid 10.
[27] ICRC, 'Position' (n 22) 5.
[28] ICRC, 'Autonomy' (n 25) 11.
[29] ibid 12.
[30] ibid.
[31] ibid 3.
[32] e.g., *Legality of the Threat or Use of Nuclear Weapons* (Advisory Opinion) [8 July 1996] ICJ Rep 226, paras 39 and 85-86; International Law Commission (ILC), 'Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with commentaries' in Report of the International Law Commission on the work of its fifty-third session (23 April–1 June and 2 July–10 August 2001) UN Doc A/56/10 at 154-155, Commentary to Draft Article 3, paras 11 and 14.
[33] 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 13 December 2019 CCW/MSP/2019/9 Annex III, 'Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System' preamble and paras (a), (c)-(e), and (h).
[34] See Antonio Coco and Talita de Souza Dias, 'Prevent, Respond, Cooperate: States? Due Diligence Duties Vis-à-Vis the Covid-19 Pandemic' (2020) 11 Journal of International Humanitarian Legal Studies 218; Antonio Coco and Talita de Souza Dias, '"Cyber Due Diligence": A Patchwork of Protective Obligations in International Law' (2021) 32 European Journal of International Law 771.

The concept has been developed over the years through several landmark cases before international courts and tribunals, such as *Island of Palmas*,[35] *Trail Smelter*,[36] *Corfu Channel*,[37] and *Pulp Mills*.[38] It has gained particular prominence in the environmental realm, where prevention and precaution now inform every state action, and impact assessments are a pre-condition to any industrial activity.[39] International human rights law (IHRL) and IHL also enshrine rules requiring states to behave diligently to avoid certain harms. Given the potential dangers that AI technology poses to persons and objects protected under international law, it is crucial that states behave diligently and adopt appropriate safeguards. [40]

This Chapter argues that states must exercise due care and caution in the design, development, and deployment of AI technology not only as a matter of ethics and common sense, but also of international law. Section 2 starts by exploring the notion of due diligence, best understood as a standard of conduct that is found across a variety of 'protective' obligations in international law. In so doing, that Section unpacks some of those rules as they apply to AI, focussing on the so-called 'Corfu Channel' and 'no-harm' principles, and positive duties under IHL and IHRL. Section 3 then proposes several measures of due diligence that states may be required to adopt when developing and deploying AI technologies in order to comply with the said protective obligations. These notably include impact assessments, testing and training, continued oversight, human involvement, and regulatory measures.

2. Protective Duties and Due Diligence Standards Applicable to AI Technologies

There is often confusion as to the nature of 'due diligence' in international law. In *Island of Palmas*, the arbitral tribunal found that, as a corollary of territorial sovereignty, states have an 'obligation to protect within their territory the rights of other States, in particular their right to integrity and inviolability in peace and in war'.[41] The International Court of Justice (ICJ) referred to the same obligation in *Corfu Channel*, clarifying that it belonged to every state as a 'general and well-recognized principle' of international law.[42] Later on, in *Pulp Mills*, the Court reaffirmed the 'principle of prevention' and 'customary rule' which 'has its origins in the due diligence that is required of a state in its territory'.[43] In the literature, 'due diligence' has been described as a general principle of law, one or more self-standing state obligation(s) of conduct, or a standard of behaviour applying in different areas of international law.[44] This trend continues to this day in different contexts. For instance, in relation to states' use of information and communications

[35] *Island of Palmas Case (or Miangas) (United States v Netherlands)*, Award, 4 April 1928, II RIAA 829 (1928), ICGJ 392 (PCA 1928), 839.
[36] *Trail Smelter Case* (*USA v Canada*) (1941) 3 RIAA 1911, 1963.
[37] *Corfu Channel Case* (*United Kingdom v Albania*) (Judgment) [9 April 1949] ICJ Rep 4, 22.
[38] *Pulp Mills on the River Uruguay, Case Concerning (Argentina v Uruguay)* (Judgment) [20 April 2010] ICJ Rep 14, paras 101, 187, 197, 204, 223.
[39] Study Group on Due Diligence in International Law, 'Second Report' (Johannesburg 2016) (International Law Association (ILA), London 2016) 3-5 <https://www.ila-hq.org/index.php/study-groups> accessed 10 June 2021.
[40] So also Deeks, Lubell and Murray (n 9) 9; McGregor, Murray and Ng (n 11) 342.
[41] *Island of Palmas* (n 35), 839.
[42] *Corfu Channel* (n 37) 22.
[43] *Pulp Mills* (n 38), para 101.
[44] Neil McDonald, 'The Role of Due Diligence in International Law' (2019) 68 ICLQ 1041, 1043–1044, fn 13; Timo Koivurova, 'Due Diligence' , *Max Planck Encyclopedia of Public International Law* (2010) paras 1-2.

technologies, Finland,[45] France,[46] and the Netherlands[47] have referred to due diligence as both a principle and an obligation. Although these different articulations may be correct in substance, they have led to some confusion over the nature of due diligence and its applicability to different phenomena.

In seeking to find common ground, we have proposed to shift the focus of debate from label to substance.[48] After all, what matters is not the exact nature of due diligence but the extent to which states must prevent, halt, and redress different types of harm or injury. In addressing this substantive question, we found that several rules of international law require states to behave diligently. In this sense, 'due diligence' features as a flexible standard of conduct in a variety of 'protective' obligations.[49] Most of them are obligations of conduct, whereby state behaviour is measured against the requisite standard. Examples include the duty to prevent genocide under Article I of the Genocide Convention,[50] and the obligation to prevent marine pollution under Article 194(2) of the UN Convention on the Law of the Sea.[51] Yet other rules require a specific result as a type of diligent behaviour, such as the duty to carry out an environmental impact assessment.[52]

In sum, a 'patchwork' of international obligations spread across various regimes requires states to behave diligently, in different ways, with a view to preventing, stopping, or redressing a range of harms. Below, we delve deeper into some of these duties as they apply to AI technologies.

a. General Protective Obligations: the Corfu Channel and no-harm principles

At least two protective obligations of general application in international law require states to exercise due diligence in the development and deployment of AI technologies. The first is the abovementioned Corfu Channel principle, famously articulated by the ICJ as 'every state's obligation not to allow knowingly its territory to be used for acts contrary to the rights of other states'.[53] Grounded in customary international law, the principle requires states to take all necessary steps to prevent and stop acts contrary to the rights of other states which they know or should know emanate from their territory. These need not constitute internationally wrongful acts attributable to another state and may well be perpetrated by non-state actors.[54]

---

[45] 'International law and cyberspace: Finland's national positions' (15 October 2020) 4 <https://um.fi/documents/35732/0/Cyber+and+international+law%3B+Finland%27s+views.pdf/41404cbb-d300-a3b9-92e4-a7d675d5d585?t=1602758856859> accessed 14 June 2021 (emphasis added).

[46] 'France's response to the pre-draft report from the OEWG Chair' (*UN.org,* 2020) 3 <https://www.un.org/disarmament/open-ended-working-group/> accessed 14 June 2021.

[47] The Netherlands, 'Letter of 5 July 2019 from the Minister of Foreign Affairs to the President of the House of Representatives on the international legal order in cyberspace — Appendix: International law in cyberspace' (5 July 2019) 4-5 <https://www.government.nl/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace> accessed 14 June 2021.

[48] Coco and de Souza Dias, '"Cyber Due Diligence": A Patchwork of Protective Obligations in International Law' (n 33).

[49] ILA 2nd Report (n 39) 2; Krieger and Peters, 'Due Diligence and Structural Change in the International Legal Order', in H. Krieger, A. Peters and L. Kreuzer (eds.), *Due Diligence and Structural Change in the International Legal Order* (OUP 2020) 351.

[50] Convention on the Prevention and Punishment of the Crime of Genocide (1948) 78 UNTS 277; *Case Concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia-Herzegovina v. Yugoslavia)*, (Judgment) [26 February 2007] ICJ Rep 43, paras 430-431.

[51] United Nations Convention on the Law of the Sea (1982) 1833 UNTS 397.

[52] *Pulp Mills* (n 38), para 204; *Certain Activities Carried out by Nicaragua in the Border Area (Costa Rica v Nicaragua) and Construction of a Road in Costa Rica along the San Juan River (Nicaragua v Costa Rica)* (Judgment) [16 December 2015] ICJ Rep 665, para 104; Astrid Epiney, 'Environmental Impact Assessment', *MPEPIL* (January 2009), paras 1-4.

[53] *Corfu Channel* (n 37) 22.

[54] ibid; *Affaire des biens britanniques au Maroc espagnol (Spain v UK)* (1925) 2 RIAA 615, 643-644.

A variety of AI-enabled acts may be contrary to the rights of other states. For example, foreign states and actors have been accused of exploiting social media platforms' AI-enabled recommendation algorithms to spread disinformation about elections[55] and COVID-19 medical treatments,[56] as well as to spur hatred and division.[57] These information operations have undermined victim states' right to non-intervention, as well as their populations' rights to self-determination, health, security, freedom of expression, non-discrimination, and free participation in elections. Similarly, Cambridge Analytica's data mining AI software was employed in several states to extract confidential information about individuals located abroad, thereby violating their right to privacy.[58]

Importantly, the Corfu Channel principle only applies insofar as the 'host' state *knowingly* allows its territory to be used for such acts. This means that a state can only be held responsible for breaching this obligation if it had actual or constructive knowledge (i.e., it either knew or should have known) that an act contrary to the rights of another state emanated from its territory.

Given the unpredictability of AI technologies, the key question here is whether and to what extent a state knew or should have reasonably known that a particular system could fail or otherwise injure the rights of other states. Because it is well-documented that AI may fail if deployed in situations for which it was *not* trained, a state may be reasonably deemed to have constructive knowledge of any harm ensuing from the use of the technology in instances falling outside its trained dataset. In a hypothetical example, if state A allows corporations based in its territory to use AI facial recognition software on ethnic groups in state B which the algorithm was not trained to recognise, then state A should have known that such use of AI technology could undermine the rights of state B and its nationals.

Importantly, the Corfu Channel principle does not require states to do the impossible to successfully prevent the harm. Rather, states must behave reasonably in the circumstances according to their own capacity[59] and acquire the minimum governmental apparatus enabling them

---

[55] Samantha Bradshaw, Hannah Bailey and Philip N. Howard, 'Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation' (*Oxford Internet Institute*, 13 January 2021) < https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/> accessed 14 June 2021.

[56] Fabio Tagliabue, Luca Galassi and Pierpaolo Mariani, 'The "Pandemic" of Disinformation in COVID-19' (2020) 2 SN Compr Clin Med 1287, 1287-1289, <https://pubmed.ncbi.nlm.nih.gov/32838179/>; Sahil Loomba et al, 'Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA' (2021) 5 Nature Human Behaviour 337 <https://www.nature.com/articles/s41562-021-01056-1>; Melinda Mills, 'COVID-19 vaccine deployment: Behaviour, ethics, misinformation and policy strategies' (*The British Academy and The Royal Society*, 21 October 2020) <https://royalsociety.org/-/media/policy/projects/set-c/set-c-vaccine-deployment.pdf> all accessed 14 June 2021.

[57] Ryan Goodman, Mari Dugas and Nicholas Tonckens, 'Incitement Timeline: Year of Trump's Actions Leading to the Attack on the Capitol' (*Just Security*, 11 January 2021) <https://www.justsecurity.org/74138/incitement-timeline-year-of-trumps-actions-leading-to-the-attack-on-the-capitol/>; Amina Ahmed, 'A Tsunami Of Hate': The Covid-19 Hate Speech Pandemic' (*Human Rights Pulse*, 20 June 2020) <https://www.humanrightspulse.com/mastercontentblog/a-tsunami-of-hate-the-covid-19-hate-speech-pandemic> both accessed 14 June 2021; OHCHR 'Report of the independent international fact-finding mission on Myanmar' (12 September 2018) UN Doc A/HRC/39/64, para 73.

[58] Issie Lapowsky, 'How Cambridge Analytica Sparked the Great Privacy Awakening' *Wired* (17 March 2019) <https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/>; Lorna McGregor, 'Cambridge Analytica is more than a data breach – it's a human rights problem' *The Conversation* (4 June 2018) <https://theconversation.com/cambridge-analytica-is-more-than-a-data-breach-its-a-human-rights-problem-96601> both accessed 02 February 2023.

[59] *Alabama Claims Arbitration (USA v UK)* (1872) 29 RIAA 125, 129; ILA 2nd Report (n 39) 20, 47.

to do so.[60] Furthermore, the Corfu Channel principle is only breached when an act contrary to another state's rights materialises.[61]

The second protective obligation of general application is the no-harm principle, which is also customary in nature. As the International Law Commission (ILC) recognised in its Draft articles on Prevention of Transboundary Harm from Hazardous Activities, the no-harm or 'good neighbourliness' principle requires states to prevent and redress significant transboundary harm originating in their territory against persons, property or the environment, to the extent feasible in the circumstances.[62] Despite obvious similarities and significant overlaps, the Corfu Channel and no-harm principles have different requirements and scopes of application.

First, the no-harm principle contains a gravity threshold to the extent that it is triggered by 'significant' harm – i.e. something more than 'detectable', but not necessarily 'serious' or 'substantial.'[63] According to the ILC, 'the harm must lead to a real detrimental effect on matters such as, for example, human health, industry, property, environment or agriculture in other states'.[64] Reflecting a precautionary approach to harm prevention, this gravity threshold covers activities carrying a 'low probability of causing disastrous harm', as well as those having 'a high probability of causing significant harm'.[65] In short, '[t]he higher the degree of inadmissible harm, the greater would be the duty of care required to prevent it.'[66] Second, although the harm in question must cross a territorial boundary, it need not be caused by an internationally wrongful act or an act contrary to the rights of other states. In fact, the no-harm principle covers significant transboundary harm even if caused by lawful activities.[67] Third, a breach of this principle only arises after the harm materialises *and* the origin state fails to compensate the victim for the damage caused from its territory or jurisdiction.[68]

While some have questioned the applicability of the no-harm principle beyond the environmental context, it suffices to note that its origins date back to disputes involving the use of weapons, the treatment of aliens,[69] and hostile propaganda across borders,[70] well before the natural environment became a global concern. As noted by the ILC's Special Rapporteur on the topic, 'there was never an intention to propose a reduction in the scope of the topic to questions of an ecological nature,

[60] Riccardo Pisillo-Mazzeschi, 'The Due Diligence Rule and the Nature of the International Responsibility of States' [1992] German Yearbook of International Law 9, 26–27.
[61] ILC, 'Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA)' (12 December 2000) UN Doc A/Res/56/83 art 14(3); *Bosnian Genocide* (n 50), para 431; Karine Bannelier-Christakis, 'Cyber Diligence: A Low-Intensity Due Diligence Principle for Low-Intensity Cyber Operations?' (2014) 14 Baltic yearbook of international law 23, 37. *Contra* Constantine Antonopoulos, 'State responsibility in cyberspace' in Nicholas Tsagourias and Russell Buchan (eds), *Research handbook on international law and cyberspace* (Edward Elgar 2015) 55, 69.
[62] See ILC, 'Draft Articles on Prevention' (n 32), Art. 11(3) and Commentary, para 5, and Commentary to Art. 17, para 2. See also Russell Buchan, 'Cyberspace, Non-State Actors and the Obligation to Prevent Transboundary Harm' (2016) 21 Journal of Conflict & Security Law 429, 441; Oren Gross, 'Cyber Responsibility to Protect: Legal Obligations of States Directly Affected by Cyber-Incidents' (2015) 48 Cornell International Law Journal 481, 503.
[63] ILC, Draft Articles on Prevention (n 32), Commentary to Article 2, para 4 (emphasis in the original).
[64] ibid.
[65] ibid, para 3.
[66] ibid Commentary to Art. 3, para 18.
[67] ibid Commentary to Art. 1, para 6; 152, Commentary to Art. 2, para 5. See also Koivurova (n 44), para 11.
[68] Rebecca Crootof, 'International Cybertorts: Expanding State Accountability in Cyberspace' (2017) 103 Cornell Law Review 565, 603; Beatrice A Walton, 'Duties Owed: Low-Intensity Cyber Attacks and Liability for Transboundary Torts in International Law' (2016) 126 Yale Law Journal 1460, 1487–1488.
[69] See *Trail Smelter* (n 36), at 1963.
[70] See, e.g., Michael G. Kearney, *The Prohibition of Propaganda for War in International Law* (OUP 2007) 16; Arthur Larson, 'The Present Status of Propaganda in International Law' (1966) 31 Law and Contemporary Problems 439, 450; John B. Whitton, 'Hostile International Propaganda and International Law' (1971) 398 The Annals of the American Academy of Political and Social Science 14, 23-25.

or to any other subcategory of activities involving the physical uses of territory'.[71] And though, for practical purposes, the ILC Draft Articles only concern 'physical uses of territory giving rise to adverse physical transboundary effects',[72] state practice and *opinio juris* suggest that the no-harm principle applies to at least certain types of non-physical harm. Such forms of harm include lost revenues resulting from territorial delimitation,[73] 'anxiety' arising from potential nuclear damage,[74] and population relocation costs.[75]

Specific treaty-based articulations of the no-harm principle — relevant to AI technology — include: Article 38 of the Constitution of the International Communications Union,[76] requiring states parties 'to prevent the operation of electrical apparatus and installations of all kinds from disrupting the operation of telecommunication installations within the jurisdiction of other Member States'; and Article 1 of the 1936 International Convention concerning the Use of Broadcasting in the Cause of Peace,[77] requiring states parties to ensure that transmissions from their territory do not constitute incitement to war or are likely to harm 'good international understanding by incorrect statements'.

Thus, states are liable for any use of AI in their territory which causes significant harm to persons, property, or the environment in another state — e.g., when a state or non-state entity uses AI to intercept or exfiltrate private data. Likewise, a state might violate the no-harm principle if it allows its territory to be used for AI-enabled disinformation campaigns directed at another state's population. The same principle might be breached if the testing of AI-powered robots in the territory of one state causes significant harm to persons, property, or the environment in another state.

While the Corfu Channel principle requires actual or constructive knowledge of the relevant acts, the no-harm principle may be triggered by the mere existence of a *risk* of harm. Indeed, the ILC suggests that *objective* 'developments in scientific knowledge' that 'reveal a risk' of the requisite harm threshold suffice to trigger a state's duty to exercise due diligence in preventing it.[78] And this is because due diligence, in this context, 'requires a state to keep abreast of technological changes and scientific developments.'[79]

Such a precautionary approach is particularly important in the context of AI technologies, given their inherent unpredictability in performing new and specific tasks. In the case of dangerous AI applications, such as lethal AWS or AI-controlled power systems, particularly in nuclear facilities, even the slightest chance of error or malfunction — which science currently shows is possible for all machine-learning applications — requires states to take the utmost care in developing and deploying the technology.

---

[71] ILC, 'Fourth report on international liability for injurious consequences arising out of acts not prohibited by international law by Robert Q. Quentin-Baxter, Special Rapporteur' (27 June 1983) UN Doc A/CN.4/373 and Corr.1&.2, para 17.

[72] ILC, 'Draft Articles on Prevention' (n 32), Commentary to Article 1, para 16.

[73] UNGA 'Survey of state practice relevant to international liability for injurious consequences arising out of acts not prohibited by international law, prepared by the Secretariat' (16 October 1984) UN Doc A/CN.4/384, para 165, citing Separate Opinion of Judge Jessup, *North Sea Continental Shelf (Germany v Denmark and the Netherlands)* (Judgement) [1969] ICJ Rep 3.

[74] UNGA 'Survey of liability regimes relevant to the topic of international liability for injurious consequences arising out of acts not prohibited by international law (international liability in case of loss from transboundary harm arising out of hazardous activities), prepared by the Secretariat' (24 June 2004) UN Doc A/CN.4/543, para 520.

[75] UNGA 'Liability regimes relevant to the topic "International liability for injurious consequences arising out of acts not prohibited by international law": survey prepared by the Secretariat' (23 June 1995) UN Doc A/CN.4/471, para 259.

[76] Constitution of the International Communications Union (1992) 1825 UNTS 33.

[77] International Convention concerning the Use of Broadcasting in the Cause of Peace (1936) 186 UNTS 301.

[78] ILC, 'Draft Articles on Prevention' (n 32), Commentary to Art. 1, para 15.

[79] ILC, 'Draft Articles on Prevention' (n 32), Commentary to Art. 3, para 11.

b. International Humanitarian Law

There is no question that IHL applies to AI technologies used in armed conflict. States parties to the CCW have explicitly affirmed that IHL 'continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems.'[80] Likewise, when AI systems are powered by or operate through information and communication technologies (ICTs), IHL also applies to the extent that it is relevant.[81] In the following section, we unpack some of IHL's key protective obligations, analyzing the standards of due diligence they require from states and other actors when developing, purchasing, selling and deploying AI technologies.

i. The duty to ensure respect for IHL

The widest and perhaps most important positive IHL obligation applicable to AI technologies is the duty to ensure respect for IHL, including rules such as the prohibition of indiscriminate and disproportionate attacks. This protective obligation of customary nature, enshrined in Article 1 common to the 1949 Geneva Conventions as well as Article 1(1) of Additional Protocol I to the Conventions, is applicable to both international and non-international armed conflicts.[82]

Its importance is threefold. First, it applies not only during armed conflict but also in peacetime.[83] Second, and relatedly, it binds not only parties to a particular armed conflict but requires *all* states to do 'everything in their power to ensure that the humanitarian principles underlying the Conventions are applied universally', given the *erga omnes* nature of IHL.[84] Third, it requires states to refrain from committing or encouraging violations of IHL,[85] and to take positive steps to ensure that *other entities* comply with IHL.[86] This means that states must prevent violations of IHL committed not only by their own state agents but also by foreign agents, private entities and individuals over whom they exercise authority[87] or reasonable influence, including abroad,[88] to the extent feasible in the circumstances.[89] Lack of resources does not exempt states from compliance, since they remain bound to acquire all reasonable means to implement the obligation.[90]

Given its wide scope of application, the duty to ensure respect for IHL applies to all states developing, purchasing, selling, transferring, or deploying AI technologies which may be implemented in an armed conflict. Importantly, such states must not only ensure that AI systems themselves behave in accordance with IHL, but also that humans remain responsible for — and

---

[80] 'Guiding Principles' (n 33), lit. (a).
[81] UNGA 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security' (22 July 2015) UN Doc A/70/174, para 28(d).
[82] *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)* (Judgment) [27 June 1986] ICJ Rep 14, para 220; ICRC, *Commentary on the First Geneva Convention* (2016) art 1, paras 125-126 <https://ihl-databases.icrc.org/ihl/full/GCi-commentary> accessed 14 June 2021.
[83] ibid, paras 127-128 and 185.
[84] ICRC, *Geneva Convention Relative to the Protection of Civilian Persons in Time of War: Commentary* (1958) 16; *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory* (Advisory Opinion) [9 July 2004] ICJ Rep 136, paras 158-159.
[85] ICRC, *2016 Commentary* (n 82), paras 154, 158-163.
[86] ibid, paras 121, 153-154 and 164-173. On this obligation generally, Knut Dörmann and Jose Serralvo, 'Common Article 1 to the Geneva Conventions and the Obligation to Prevent International Humanitarian Law Violations' (2014) 96 International Review of the Red Cross 707.
[87] ICRC, *2016 Commentary* (n 82), para 150.
[88] ibid, paras 150, 153-154.
[89] ibid, paras 165-166 and *mutatis mutandis*, *Bosnian Genocide* (n 50), para 430. See also Marco Longobardo, 'The Relevance of the Concept of Due Diligence for International Humanitarian Law' (2020) 37 Wisconsin International Law Journal 44, 60–62.
[90] ICRC, *2016 Commentary* (n 82), para 187.

properly trained with a view to — ensuring respect for IHL throughout the systems' life cycle. These principles have been explicitly recognised by states parties to the CCW in relation to lethal AWS.[91]

In practice, this might require states to interpret and apply rules of IHL in a way that leaves no accountability gap, i.e. holding states and individuals responsible for violations of IHL caused by negligence[92] or in accordance with the principle of strict liability, including in the case of machine error or malfunction.[93] Accountability is particularly important to prevent automation bias, i.e. human over-reliance on the machine, and the moral buffer problem, i.e. the risk of humans transferring responsibility to the AI system. Such an interpretation would be in line with the type and degree of knowledge triggering states' duty to ensure respect for IHL in their use of AI technologies, that is, the objective foreseeability of a violation rather subjective knowledge thereof.[94] But it is worth noting that a violation of this positive duty only arises if the actual harm materialises.[95]

ii. Protection of civilians and precautions in the conduct of hostilities

Another key set of protective obligations under IHL, all part of customary international law, comprises a) the duty to protect civilians against the dangers arising from military operations and the ensuing obligations to take precautionary measures to b) spare civilians in the conduct of military operations and c) protect them against the effects of attacks.

The first is codified in Article 51(1) of Additional Protocol (AP) I to the Geneva Conventions,[96] providing that '[t]he civilian population and individual civilians shall enjoy general protection against dangers arising from military operations.'[97] This catch-all provision seeks to avoid gaps in the protection of civilians.[98] As such, it comprises not only an obligation to refrain from indiscriminate and disproportionate attacks against civilians, but also a duty to proactively take precautionary measures to protect them from military operations.[99] This proactive limb finds expression in the aforementioned duties to take precautions in the conduct of military operations and against the effects of attacks.[100]

The duty to take precautions in military operations, enshrined in Article 57 of AP I, requires constant care to spare the civilian population and civilian objects during any movements, manoeuvres or other activities carried out by the armed forces with a view to combat.[101] Furthermore, during armed attacks, several specific precautions must be taken. Of particular relevance for all AI military technologies are the duties a) to do everything feasible to verify that

---

91 'Guiding Principles' (n 33), lit. (b), (c) and (d).
92 See Tsvetelina van Benthem, DPhil Thesis (forthcoming).
93 Lawrence Hill-Cawthorne, 'Appealing the High Court's Judgment in the Public Law Challenge against UK Arms Export Licenses to Saudi Arabia' (*EJIL: Talk!*, 29 November 2018) comments <https://www.ejiltalk.org/appealing-the-high-courts-judgment-in-the-public-law-challenge-against-uk-arms-export-licenses-to-saudi-arabia/> accessed 14 June 2021.
94 ICRC, *2016 Commentary* (n 82), paras 150, 164.
95 ICRC, *2016 Commentary* (n 82), para 166 establishes a parallelism between common Article 1 and Article 1 of the 1948 Genocide Convention. *Bosnian Genocide* (n 50), para 431 established that a breach of the duty to prevent occurs only if genocide is committed, in line with ARSIWA art 14(3) (n 61).
96 Yves Sandoz, Christoph Swinarski and Bruno Zimmermann, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (ICRC 1987), para 1923.
97 AP I art 51. Generally Eric Talbot Jensen, 'Precautions against the Effects of Attacks in Urban Areas' (2016) 98 International Review of the Red Cross 147; Jean-François Quéguiner, 'Precautions under the Law Governing the Conduct of Hostilities' (2006) 88 International Review of the Red Cross 793.
98 ICRC, *1987 Commentary* (n 96), para 1991.
99 See Art. 51(2)-(8).
100 ICRC, *1987 Commentary* (n 96), para 2189.
101 ibid, para 2191.

only military objectives are engaged and targeted; b) to take all feasible precautions in the choice of means and methods of warfare to avoid or minimize incidental loss of civilian life, injury to civilians and damage to civilian objects; and c) to cancel or suspend an attack as soon as it becomes apparent that the objective is not a military one. Like the duty to ensure respect for IHL, these various precautionary requirements seem to necessitate meaningful human involvement and accountability throughout.

For the ICRC, the identification of objects, especially when distant or remote, should be carried out with great care. This means that those who decide or plan an attack must call for additional information and seriously evaluate its accuracy in case of doubt,[102] even if the identification of targets depends on technical means of detection, such as AI image recognition systems.[103] In fact, during the drafting of this provision, some delegates remarked that particularly dangerous weapons should contain 'safety devices to render them harmless if they fell out of the *control* of the user'.[104] Similarly, the ICRC Commentary highlights the importance of visual means in the identification of targets and the need for greater caution when the attacker lacks direct view.[105] It will be difficult to comply with this duty if humans lack control over AI detection, reconnaissance or targeting systems, given the technology's susceptibility to bias and error, as well as the lack of sufficient time to override the machine.[106] For these and other reasons, the ICRC has proposed a ban on anti-personnel AWS, as well as on autonomous weapons that are insufficiently understandable, predictable, or explainable.[107] Likewise, states must exercise extra caution when developing and deploying AI technology in all other military circumstances, such as automated systems used to gather intelligence in preparation of attacks.[108]

For its part, the duty to take precautions against the effects of attacks is spelled out in Article 58 of AP I. This positive obligation requires parties to protect their own territory and population against attacks conducted by their adversaries, i.e. passive or defensive precautions.[109] Though limited to feasible precautions, it includes not only the removal of civilians from the vicinity of military operations and vice-versa, but also any other necessary precautions to protect civilians and civilian objects from the dangers of military operations. In the AI context, this duty requires states and other parties to an armed conflict to protect their own nationals and territory from the dangers potentially caused by *any* AI application used during military operations, including, in particular, AWS. This obligation could be discharged by informing the population about the dangers of such AI systems, and adequately training military and civil defence operators to protect civilians and civilian objects from their unpredictable effects.[110]

iii. Development of new weapons, means and methods of warfare

Another positive IHL obligation having a particular bearing on AI military applications is the duty to carry out a legal review of new weapons, means or methods of warfare in their study, development, acquisition, or adoption, in accordance with IHL and other obligations under

[102] ibid., para 2195.
[103] ibid, 2199.
[104] ibid, 2201.
[105] ibid, 2221.
[106] ICRC, 'Position' (n 22) 7-9, 11.
[107] ibid 9.
[108] ibid.
[109] ICRC, *1987 Commentary* (n 96), paras 2239-2241.
[110] See Group of Governmental Experts of the High Contracting Parties related to emerging technologies in the area of lethal autonomous weapons systems, 'Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Addendum: Chair's summary of the discussion of the 2019 Group of Governmental Experts on emerging technologies in the area of lethal autonomous weapons systems' (8 November 2019) UN Doc CCW/GGE.1/2019/3/Add.1, para 20.

international law. Its applicability to lethal AWS has been explicitly recognised by states parties to the CCW,[111] and at least one state has recognised that it applies to cyber capabilities in some instances.[112]

This obligation, articulated in Article 36 AP I and viewed by many as part of customary international law,[113] aims to *prevent* the use of weapons, means and methods that would *always* violate IHL and other applicable rules of international law, as well as to *restrict* the use of those that would do so under certain circumstances. Much like other duties requiring diligent behaviour, the duty to carry out a legal review of weapons, means and methods of warfare is an important safeguard in the face of rapid developments in military technologies. It is not strictly limited to new means or methods, but also extends to existing ones that are a) modified in ways that alter their functions, b) studied, developed, acquired, or adopted for the first time by a certain state, or c) covered by a new legal instrument that the relevant state has joined.[114]

While states are not expected to foresee all possible (mis)uses of a new weapon, method or means of warfare – some of which may well be contrary to international law – they must assess their legality in *normal or expected* use(s).[115] The identification of foreseeably unlawful uses is particularly important in the context of dual-use and versatile technologies such as AI-enabled software and hardware.[116] However, in the case of technologies employing machine-learning algorithms, the dynamic, self-learning nature of the system may significantly complicate accurate foresight of its behaviour.[117] Moreover, as noted earlier, AWS and other AI military applications are prone to errors when deployed in unstable environments.[118] Thus, states which design, develop, acquire, or deploy AWS and AI-enabled means and methods of warfare must consider all the conceivable ways in which their *intended* use may violate IHL or other applicable rules of international law, as well as with ways to minimize such risk. In this light, it may be argued that AWS and AI-supported means and methods of warfare lacking in human control are likely to fail a diligent legal review process.[119]

Notably, in its commentary on Article 36, the ICRC warns against the 'uncontrolled rate of technological development',[120] and calls upon states not to adopt new weapons 'for the sole reason that they exist or because of a fear that others will develop them'.[121] Yet the opposite seems to be happening with current developments in AI military technologies, including among non-state actors.[122] For instance, the recent report of the US National Security Commission on Artificial

---

111 'Guiding Principles' (n 33), lit. (e). See also Chair's Summary (n 110), para 11.

112 Australian Government, 'Australia's submission on international law to be annexed to the report of the 2021 Group of Governmental Experts on Cyber' (28 May 2021) 4 <https://www.internationalcybertech.gov.au/sites/default/files/2021-06/Australia%20Annex%20-%20Final%2C%20as%20submitted%20to%20GGE%20Secretariat.pdf> accessed 06 February 2023 .

113 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' (January 2006) 4. See also Eric Talbot Jensen, 'The (Erroneous) Requirement for Human Judgment (and Error) in the Law of Armed Conflict' (2020) 96 International Law Studies Series 26; US Naval War College. [i], 38; Natalia Jevglevskaja, 'Weapons Review Obligation under Customary International Law' (2018) 94 International Legal Studies 186; Michael N. Schmitt and Jeffrey S. Thurnher, '"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict' (2013) 4 Harvard National Security 231, 271.

114 ICRC, *1987 Commentary* (n 96), para 1475; ICRC, 'Guide' (n 113) 9-10.

115 ICRC, *1987 Commentary* (n 96), paras 1469, 1480; ICRC, 'Guide' (n 113) 10.

116 ICRC, 'Guide' (n 113), 9-10.

117 Rebecca Crootof, 'Autonomous Weapon Systems and the Limits of Analogy' (2018) 9 Harvard National Security Journal 51, 65.

118 Access Now (n 1) 29.

119 ICRC, *1987 Commentary* (n 96), 1476: 'if man does not master technology, but allows it to master him, he will be destroyed by technology.'

120 ICRC, *1987 Commentary* (n 96), 1477.

121 ibid.

122 Chair's Summary (n 110), para 24; Access Now (n 1) 29.

Intelligence (NSCAI) invites the US government to accelerate its 'adoption of AI-enabled sensors and systems for command and control, weapons, and logistics' to 'win' and 'ensure its military-technical superiority' against 'technically sophisticated adversaries', especially China and Russia.[123] Worryingly, the same report encourages the 'relentless experimentation' of AI military technology as well as the adoption of 'a system that rewards agility and risk'.[124]

Article 36 AP I does not specify how exactly the legal review of weapons, means and methods of warfare must be carried out. Rather, states are free to do so by legal, administrative, regulatory, or any other appropriate means.[125] Nonetheless, regardless of the exact nature of the process followed, any meaningful legal review requires input from impartial experts in international law, technology and other relevant areas.[126] Impartiality is particularly important for AI, as its development is heavily concentrated in the hands of a few big tech companies which have invested large sums of money in research.[127] Casting doubts in this respect, 12 out of the 15 members of the US NSCAI are current or former senior executives of those same tech companies, including Google, Microsoft, and Amazon.[128] It thus comes as no surprise that the Commission has agreed by consensus to recommend that at least eight billion US dollars should be allocated annually towards AI research and development for military purposes.[129]

c. International Human Rights Law

The actual and potential impact of AI technologies on human rights is widespread and well-documented — endangering virtually every single human right guaranteed under international law.[130] Opaque AI algorithms used to predict recidivism and sentence individuals based on collective data can undermine individuals' rights to freedom of movement, the presumption of innocence, and other fair trial rights.[131] Similarly, AI systems used by public and private entities to filter job applications, admit students, calculate citizens' credit scores, and offer social benefits or health treatments are often fed by historical data and proxy indicators, such as individuals' purchasing habits, address, or social connections.[132] As such, they might violate individuals' rights to work, education, health, and an adequate standard of living.

Machine learning algorithms behind social media feeds tend to prioritise false or violent content and side-line dissenting and minority voices.[133] This may run counter to users' rights to receive and impart information, to freely form opinions, to vote, and, in extreme cases of online abuse, the

---

123 US National Security Commission (n 8) 61.
124 ibid 77.
125 ICRC, 'Guide' (n 113) 20.
126 ibid 21-22.
127 Amy Webb, *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity* (Public Affairs 2019).
128 US National Security Commission on Artificial Intelligence, 'Commissioners' <https://www.nscai.gov/commissioners/> accessed 14 June 2021.
129 US National Security Commission (n 8), 68.
130 UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (29 August 2018) UN Doc A/73/348.
131 Access Now (n 1) 15, 19, 24; McGregor, Murray and Ng (n 11) 310–311, 319, 326.
132 Access Now (n 1) 12, 14, 16; O'Neil (n 12) 141–160.
133 Special Rapporteur on freedom of opinion and expression (n 130), paras 9-18; Access Now (n 1) 16, 22-25; Yaël Eisenstat, 'Dear Facebook, this is how you're breaking democracy' (*TED*, August 2020) <https://www.ted.com/talks/yael_eisenstat_dear_facebook_this_is_how_you_re_breaking_democracy> accessed 14 June 2021; Carole Cadwalladr, 'If you're not terrified about Facebook, you haven't been paying attention', *The Guardian* (26 July 2020) <https://www.theguardian.com/commentisfree/2020/jul/26/with-facebook-we-are-already-through-the-looking-glass> accessed 14 June 2021; Cathy O'Neil, 'TikTok's Algorithm Can't Be Trusted' (*Bloomberg.com*, 21 September 2020) <https://www.bloomberg.com/opinion/articles/2020-09-21/tiktok-s-algorithm-can-t-be-trusted> accessed 10 June 2021; O'Neil (n 12) 180–185.

prohibition of incitement to discrimination, hostility, or violence.[134] AWS and other AI military applications that incorrectly identify and engage targets, or select individuals for security detention,[135] may directly infringe their rights to life, liberty, health, and freedom from inhumane treatment.[136] In addition, all these technologies may well be based on biased or unrepresentative data collected through state or private surveillance, which means that they might also violate the rights to non-discrimination and privacy.[137]

Admittedly, most human rights under international law are not absolute, but can be subject to necessary and proportionate limitations which are provided for by law and pursue a legitimate aim as specified in the relevant legal instruments.[138] However, some features of AI — especially machine learning algorithms — make it harder for these requirements to be satisfied. Although many AI applications may seek a legitimate aim and are provided for by law, the restrictions they impose on human rights are often unnecessary or disproportionate.[139] This is because many AI systems make decisions that affect individuals based on data and statistics that either a) consist of *proxy* indicators, i.e. features that are not causally or directly connected to the algorithm's intended output (e.g. using someone's Internet browsing history to calculate their credit score; identifying images of animals by correlation of pixels to the background);[140] or b) belong to *other* individuals or groups (e.g. facial recognition systems that are trained using photos of other individuals).[141] To safeguard against such arbitrary decisions, as well as the use of adversarial examples – slight changes in the dataset or the use of random inputs to fool the machine's output –[142] meaningful human control may be essential.[143] Moreover, the difficulty in explaining how AI's complex 'black box' algorithms reach their decisions hinders the effective verification of the lawfulness of limitations, as well as an individual's right to an effective remedy.[144]

For states, which have human rights obligations under both conventional and customary international law, this has two key implications. First, when developing, selling, purchasing, or deploying AI technologies for either military or civilian purposes, states have *negative* obligations to refrain from violating the human rights of individuals within their jurisdiction.[145] This means that they must refrain from using those technologies in ways that *arbitrarily* limit the human rights of individuals in their territory and arguably abroad, to the extent that they have functional control over the enjoyment of those rights.[146]

Second, and most importantly for the purposes of this contribution, states also have *positive* duties to protect the human rights of individuals within their jurisdiction against any use of AI by third parties, including both private and public entities (such as foreign states), in ways that may

---

134 Special Rapporteur on freedom of opinion and expression (n 130), paras 22-32.
135 Ashley Deeks, 'Detaining by Algorithm' (*Humanitarian Law & Policy Blog*, 25 March 2019) <https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/> accessed 14 June 2021.
136 UN Human Rights Committee 'General Comment No. 36 (2018) on Article 6 of the International Covenant on Civil and Political Rights, on the Right to Life' (30 October 2018) UN Doc CCPR/C/GC/36, para 65.
137 Special Rapporteur on freedom of opinion and expression (n 130), paras 33-38; Ashley Deeks, 'Introduction to the Symposium: How Will Artificial Intelligence Affect International Law?' (2020) 114 AJIL Unbound 138, 138.
138 Special Rapporteur on freedom of opinion and expression (n 130), para 28.
139 McGregor, Murray and Ng (n 11) 337–338; Access Now (n 1) 19.
140 Special Rapporteur on freedom of opinion and expression (n 130), para 7; Mitchell (n 1) 120-123; O'Neil (n 12) 12, 14–31, 145.
141 Special Rapporteur on freedom of opinion and expression (n 130), para 38; Mitchell (n 1) 124; Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 Proceedings of Machine Learning Research 1, 2–3.
142 Mitchell (n 1) 128-135.
143 HRC 'General Comment 36' (n 136), para 65.
144 Special Rapporteur on freedom of opinion and expression (n 130), paras 39-41.
145 ibid, para 19.
146 HRC 'General Comment 36' (n 136), para 63; Yuval Shany, 'Taking Universality Seriously: A Functional Approach to Extraterritoriality in International Human Rights Law' (2013) 7 Law & Ethics of Human Rights 47.

foreseeably violate human rights.[147] Compliance with such positive duties is measured by a standard of due diligence.[148] When preventing such harms is not viable, states must at least mitigate and redress their consequences insofar as possible, including by investigating and holding to account those responsible.[149] Again, ensuring meaningful human control and accountability for violations may be an important safeguard against AI's frailty. This is because shifting responsibility to humans not only forces machine operators to take extra care, but also prevents frequent machine errors that a human would be less likely to make.

A state's duty to protect human rights concerns not only potential violations it knows or should have known of, but also all those objectively foreseeable given the circumstances.[150] This knowledge threshold has important consequences in the context of AI technology. First, it means that not only actual harms but also risks of harm may trigger positive human rights obligations. As most AI applications have the potential to violate human rights, states must keep a close eye on those risks throughout the technology's entire life cycle. Second, AI's unpredictability cannot be raised as an excuse for inaction: states' current knowledge of AI's unpredictable behaviour already puts them on notice of the likely risks and harms caused by the technology, triggering both negative and positive human rights obligations.[151]

Yet this does not require states to prevent, mitigate and redress *all* human rights violations caused by AI technologies. Like other protective obligations, positive duties to protect human rights only require states to do what they are reasonably capable of in the circumstances — bearing in mind that, at the very least, they must put in place the minimal state apparatus necessary to fulfil such duties.[152] Thus, if all a state can do to prevent a human rights violation by a tech company located in its territory is to regulate the use of AI and hold to account those responsible for any violations, then it must do so. Of note, positive duties take a specific form with respect to certain human rights. For instance, states must actively monitor and seek information about possible violations of the right to life, as well as effectively investigate and punish such violations.[153]

3. Measures of Diligent State Behaviour

Having laid out the 'patchwork' of protective obligations applicable to AI technologies under international law, we now turn to their implementation. This section proposes measures that can constitute 'diligent behaviour' aimed at preventing, stopping and/or redressing AI-enabled harm, thus discharging one or more of the duties examined above. The list is by no means exhaustive,

---

147 Special Rapporteur on freedom of opinion and expression (n 130), para 20; HRC 'General comment No. 31 [80], The nature of the general legal obligation imposed on States Parties to the Covenant' (26 May 2004) UN Doc CCPR/C/21/Rev.1/Add.13, para 8; 'General Comment No. 3: The Nature of States Parties' Obligations (Art. 2, Para 1, of the Covenant)' (14 December 1990) UN Doc E/1991/23, para 1. See also International Covenant on Civil and Political Rights (1966) 999 UNTS 171 (ICCPR) art 2(1)-(2); International Covenant on Economic, Social and Cultural Rights (1966) 993 UNTS 3 (ICESCR) art 2(1); American Convention on Human Rights (1978) OAS Treaty Series No 36 1144 UNTS 123 (ACHR) art 1(1); European Convention for the Protection of Human Rights and Fundamental Freedoms (1953) ETS 5 (ECHR) art 1.

148 HRC, 'General Comment 31' (n 147), para 8; Samantha Besson, 'Due Diligence and Extraterritorial Human Rights Obligations – Mind the Gap!' (2020) 9 ESIL Reflections 2 <https://esil-sedi.eu/esil-reflection-due-diligence-and-extraterritorial-human-rights-obligations-mind-the-gap/> accessed 6 June 2020.

149 ICCPR art 2(3); HRC, 'General Comment 31' (n 147), paras 8, 15 and 18.

150 HRC, 'General Comment 36' (n 136), para 7; ECtHR, *Keller v. Russia* (2013) (Judgement) Appl. no. 26824/04, para 82; ECtHR, *Osman v. United Kingdom* (1998) (Judgment) Appl. no. 23452/94, para 116; *O'Keeffe v. Ireland* (2014) (Judgement) Appl. no. 35810/09, paras 16, 162; *Kurt v. Turkey* (1998) (Judgement) Appl. no. 15/1997/799/1002, para 69; Special Rapporteur on freedom of opinion and expression (n 130), paras 13, 14(f), 16.

151 McGregor, Murray and Ng (n 11) 341–342.

152 Besson (n 142) 5–7.

153 HRC, 'General Comment 36' (n 136), para 67; ECtHR, *McCann and Others v. United Kingdom* (1995) (Judgement) Appl. no. 19009/04, para 161; *Güzelyurtlu and Others v. Turkey* (2019) (Judgement) Appl. no. 36925/07, para 189.

but should serve as a guide for states in the course of developing, selling, acquiring, and employing AI systems. Crucially, all the proposed measures — e.g., those concerning monitoring — must be applied in a manner which is consistent with other international obligations.

a. Impact Assessments and Other Technical Measures

To discharge its protective obligations, a diligent state should ensure that any AI technology it designs and develops — or otherwise plans to purchase, sell, or employ — undergoes thorough and continuous impact assessments, to strive for safer technology and safer use.[154] The establishment of a procedure to verify the potential harm that a machine could cause — and intervene to avoid it — is not so different from the establishment of procedural safeguards against errors of judgment by another fallible entity, i.e. humans.[155] The assessment should cover the entire life-cycle of the technology, from the early stages of conception to deployment,[156] giving developers and public authorities sufficient time and opportunity to address violations of international law, whether foreseen or not.

Impact assessments have multiple aims. Above all, states should satisfy themselves that the machine is sufficiently predictable before it is deployed. This includes checking the instructions it follows, the origin, quantity and quality of data with which it has been trained, the accuracy with which it executes assigned functions, its behaviour in unpredicted circumstances, and the possibility of errors or malfunctions.[157] The ability to understand and explain how/why the algorithm reached a certain decision and/or acted upon it is thus a crucial element of a successful impact assessment,[158] even if this can only be done by reverse-engineering the relevant outputs or by providing counterfactual explanations.[159] If this is not possible, states should avoid using machine learning systems and perhaps revert to symbolic algorithms, which can be more easily understood and verified by computer programmers.[160]

The legal review of weapons as per Article 36 AP I is a special form of impact assessment, to the extent that it seeks to verify how AI weapons, means and methods of warfare will behave in normal or intended settings.[161] The review, ideally entrusted to a standing multi-disciplinary mechanism,[162] should take place as early as possible and prior to the development or purchase of a weapon.[163] In

---

[154] Special Rapporteur on freedom of opinion and expression (n 130), para 53.

[155] Indeed, AI impacts assessments could resemble human rights impact assessments (HRIA): Jessica Fjeld et al, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' [2020] Berkman Klein Center for Internet & Society 30 <https://dash.harvard.edu/handle/1/42160420> accessed 14 June 2021; Malcolm Langford, 'Taming the Digital Leviathan: Automated Decision-Making and International Human Rights' (2020) 114 AJIL Unbound 141, 145.

[156] McGregor, Murray and Ng (n 11) e.g. 323, 330.

[157] Deeks, Lubell and Murray (n 9) 24; McGregor, Murray and Ng (n 11) 334.

[158] Deeks, Lubell and Murray (n 9) 20–21, 24; McGregor, Murray and Ng (n 11) 320.

[159] Mitchell (n 1) 128; Access Now (n 1) 36-37. See generally Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 Harvard Journal of Law & Technology 841.

[160] Access Now (n 1) 26. See also Cynthia Rudin and Joanna Radin, 'Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition' (2019) 1 Harvard Data Science Review 10.

[161] With reference to predictive algorithms to be employed in relation to detention decisions in armed conflict, Lorna McGregor, 'The Need for Clear Governance Frameworks on Predictive Algorithms in Military Settings' (*Humanitarian Law & Policy Blog*, 28 March 2019) <https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings/> accessed 10 June 2021.

[162] Even though few states are known to have such mechanisms in place: Netta Goussec, 'Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting' (*Humanitarian Law & Policy Blog*, 18 April 2019) <https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/> accessed 10 June 2021.

[163] ICRC, 'Guide' (n 113) 23–24; Goussec (n 156).

the case of existing weapons, this review should take place whenever the weapon's functions are modified.[164] AWS powered by machine-learning should undergo a periodic legal review, since the technology autonomously adapts to stimuli based on the data it is fed and/or collects.[165] Additional review processes may be required under treaties that do not fall within the scope of this Chapter, such as the Arms Trade Treaty.[166]

Lewis has suggested a non-exhaustive list of parameters that a diligent state should verify before ever deploying AWS – a reasoning that can be extrapolated to all AI technology.[167] Through continuous cycles of testing and training, an effective impact assessment should verify, *inter alia*: whether humans would maintain legal and actual agency; whether the AI decision-making would be explainable; whether the AI risks engaging in a 'normative inversion', i.e. reversing the presumptions established by the relevant law (e.g. that of protected civilian status); whether there are sufficient limits to its use in unforeseen circumstances; how relevant facts and information are reflected and represented in the computational components; whether any bias is being transferred to the algorithm; whether ongoing monitoring of the AI's functioning would be possible.[168] Others have suggested an assessment of: the circumstances and timing of AI deployment;[169] the environment with which the technology is supposed to interact; the possible interactions with other AI technologies; and the persons or entities that would have access to the resulting data.[170] An ethical analysis may also be an important complement to a legal review.[171]

Prominently, the impact assessment should identify any potential harms caused by the machine,[172] including those foreseeably triggered by third parties (private or public) developing, selling or using the technology.[173] These can take the form of transparency reports issued by users and/or manufacturers of the technology,[174] as well as internal or external audits.[175] To ensure the accuracy of the results, any review or assessment should be undertaken by an impartial body of experts, and involve consultations with relevant stakeholders, such as academia and civil society, including underrepresented groups.[176] To prevent biases that are so common in the use of AI, experts should be as diverse as possible, both in terms of their area of expertise as well as their gender, race, cultural and socio-economic background.[177]

The information thus gathered should lead to the adoption of technical measures to avoid or mitigate the harms identified.[178] For instance, developers could: devise a deactivation threshold triggered by the lapse of a certain amount of time or by the degradation of a critical safety feature;[179] build in 'fail-safe' mechanisms to allow human operators to safely take over or override the

---

164 ICRC, 'Guide' (n 113) 24; Goussec (n 156).
165 Goussec (n 156); Special Rapporteur on freedom of opinion and expression (n 130), paras 8, 40.
166 See e.g. Art. 7 Arms Trade Treaty (2013) UN Doc A/Res/67/234 B, which requires an assessment of potential violations of international law before an arms export can be authorised.
167 Dustin Lewis, 'Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider' (*Humanitarian Law & Policy Blog*, 21 March 2019) <https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider/> accessed 10 June 2021.
168 ibid.
169 McGregor, Murray and Ng (n 11) 334; Murray (n 2) 159.
170 Murray (n 2) 159.
171 Special Rapporteur on freedom of opinion and expression (n 130), para 48.
172 McGregor, Murray and Ng (n 11) 334.
173 ibid.
174 Access Now (n 1) 33, 45.
175 Special Rapporteur on freedom of opinion and expression (n 130), para 55, 62, 69.
176 ibid, para 54.
177 Access Now (n 1) 34; ICRC, 'Guide' (n 113) 17, 21–22.
178 McGregor, Murray and Ng (n 11) 334.
179 Lewis (n 161); ICRC, 'Position' (n 22) 10.

system;[180] train the machine to prioritise specific data that is deemed to be particularly reliable;[181] set minimum levels of accuracy before any decision is made and acted upon;[182] or attempt to embed international human rights standards as faithfully as possible in the algorithm,[183] including by avoiding or redressing bias.[184] However, the precision required when coding an algorithm means that the machine will often be unable to accurately interpret or apply ambiguous legal provisions.[185] Even though dialogue between developers and lawyers can help bridge this gap,[186] value judgments are best left to humans.[187] The results of the impact assessment may also enable states to decide whether the technology should be limited to informing, rather than making, decisions.[188] The algorithm could also be tailored and trained considering the state's own policy objectives.[189]

While impact assessments can go a long way in preventing violations of international law, some degree of unpredictability is bound to remain, given the impossibility of simulating all possible settings in which the technology will operate.[190] Thus, investment in the people using the technology is pivotal: humans must be trained to use and rely on AI, as well as to resist automation bias.[191]

b. Human Oversight

There is often a tendency to anthropomorphise AI, accompanied by an expectation that machine outputs automatically align with legal standards made by and for humans. However, this focus on the machine alone, an 'artifact-centric conception of technology',[192] overlooks that it is humans — not machines — who remain accountable under international law. Thus, states must actively ensure that AI algorithms comply with international law.[193] This arguably necessitates an appropriate level of human oversight.[194] Although the exact level of oversight remains the subject of controversy among states, technologists, and scholars, consensus has emerged on the need for meaningful human control to counter AI's unpredictability,[195] bias, and unexplainability.[196] In this sense, providing for some level of human control over AI may be a sign of diligent state behaviour.

As reaffirmed at the Asilomar Conference on Beneficial AI, '[h]umans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.'[197] Likewise, the European Commission's High-Level Expert Group on AI has called upon member states to

---

[180] ICRC, *1987 Commentary* (n 96), para 2201; ICRC, 'Position' (n 22) 10.
[181] Deeks, Lubell and Murray (n 9) 15.
[182] ibid.
[183] Special Rapporteur on freedom of opinion and expression (n 130), para 48; McGregor, Murray and Ng (n 11) 342.
[184] Special Rapporteur on freedom of opinion and expression (n 130), paras 52, 67.
[185] Suresh Venkatasubramanian, 'Structural Disconnects between Algorithmic Decision-Making and the Law' (*Humanitarian Law & Policy Blog*, 25 April 2019) <https://blogs.icrc.org/law-and-policy/2019/04/25/structural-disconnects-algorithmic-decision-making-law/> accessed 10 June 2021.
[186] Deeks, Lubell and Murray (n 9) 15.
[187] One of the assessments listed by Lewis (n 161).
[188] Deeks, Lubell and Murray (n 9) 24.
[189] ibid 14.
[190] Goussec (n 156).
[191] Deeks, Lubell and Murray (n 9) 18.
[192] Langford (n 149) 145.
[193] Special Rapporteur on freedom of opinion and expression (n 130), para 19; Mathias Risse, 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda' (2018) HKS Faculty Research Working Paper Series RWP18-015, 8.
[194] See Chair's Summary (n 110), paras 10, 18-19.
[195] Rebecca Crootof, 'A Meaningful Floor for Meaningful Human Control Autonomous Legal Reasoning: Legal and Ethical Issues in the Technologies in Conflict' (2016) 30 Temple International & Comparative Law Journal 53, 53.
[196] ICRC, 'Guide' (n 113), at 3
[197] 'Asilomar AI Principles' (*Future of Life Institute,* 11 August 2017) no. 16 <https://futureoflife.org/ai-principles/> accessed 14 June 2021.

guarantee human agency and oversight through 'human-in-the-loop', 'human-on-the-loop', or 'human-in-command' approaches in their employment of AI technology.[198] Some states' national positions have gone in a similar direction. The French Ministry of Defence, for instance, chose the maintenance of sufficient human control as one of the guiding principles in the use of AI.[199] Similarly, the US Department of Defence, in adopting its Principles of AI Ethics, indicated that the use of AI must be 'responsible' and 'governable', including through appropriate forms of human control.[200]

However, the idea of ensuring some level of human control should not be taken to an extreme. At times, it may even be unnecessary, since not all uses of AI technology put humans, property, or the environment at risk. Secondly, keeping a human in the loop may undermine a clear advantage of employing AI technology, i.e. the immediacy and automation of certain decisions.[201] The key question then becomes which decisions can be delegated to AI, and which ones must remain under human control.[202] In answering this question, relevant benchmarks include the expected accuracy of the AI decision-making, as well as the seriousness and likelihood of the potential harm.[203]

In the same vein, human oversight is not a panacea against all AI-induced harms. Humans 'in the loop', 'on the loop' or 'in command' may well fall prey to automation bias and fail to question the validity of the machine's decision or to recognize an error or malfunction.[204] Algorithmic bias resulting from the lack of appropriate training data may also go unnoticed.[205] Machine decisions may even be used as 'moral buffer' to follow through with otherwise uncomfortable plans of action. For these and other reasons, Canada has recommended a case-by-case approach to decide on the appropriate degrees of machine autonomy and human involvement in various circumstances.[206] The real issue with human involvement is not so much whether a human is there, but whether that human is effective at reducing the risk of harm.[207] Again, this requires properly trained individuals with 'strong autonomy' and expert knowledge.[208] Only then can human oversight be a measure of due diligence. Otherwise, it will remain 'nothing more than a charade',[209] 'to make us feel better'[210] about the risks of new technology.

Granted, states have not agreed on the necessary level of human control: views range from human authorization for each decision, simple oversight coupled with the power to stop the machine, or

---

198 'Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence' (*European Commission,* 8 April 2019) 15–16 <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> accessed 14 June 2021.

199 ICRC, 'Autonomy' (n 25) 9.

200 Todd Lopez, 'DOD Adopts 5 Principles of Artificial Intelligence Ethics' (*Defense.gov,* 25 February 2020) <https://www.defense.gov/Explore/News/Article/Article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/> accessed 14 June 2021.

201 Deeks, Lubell and Murray (n 9) 9.

202 Thomas Burri, 'International Law and Artificial Intelligence' (2017) 60 German Yearbook of International Law 91, 103.

203 As suggested e.g. by Microsoft, cited in ICRC, 'Autonomy' (n 25) 9.

204 McGregor, Murray and Ng (n 11) 338.

205 ibid 317.

206 'Responsible Artificial Intelligence in the Government of Canada' (10 April 2018) 18-19 <https://docs.google.com/document/d/1Sn-qBZUXEUG4dVk909eSg5qvfbpNlRhzIefWPtBwbxY/edit> accessed 14 June 2021.

207 As put e.g. by McGregor (n 155).

208 Dimitri van den Meerssche, "'The Time Has Come for International Regulation on Artificial Intelligence" – An Interview with Andrew Murray' (*Opinio Juris*, 25 November 2020) <http://opiniojuris.org/2020/11/25/the-time-has-come-for-international-regulation-on-artificial-intelligence-an-interview-with-andrew-murray/> accessed 10 June 2021.

209 Burri (n 196) 99–100.

210 van den Meerssche (n 202).

an even lower standard of careful programming.[211] Yet the constructive ambiguity over what constitutes 'meaningful human control'[212] is not an excuse for inaction: diligent state behaviour is measured by genuine efforts to avoid foreseeable harm in the development and employment of AI technology.

c. Regulation

The enactment of an adequate national legal framework is not only necessary to protect human rights from the impact of AI, but also to satisfy the requirement of legality when limiting those rights through the use of the technology. Thus, states must put in place clear and comprehensive laws or regulations before authorising the development, sale, acquisition, or deployment of AI technologies which may foreseeably affect individual human rights within their jurisdiction.[213] Regulation is also an important step in ensuring compliance with positive duties under IHL[214] as well as the Corfu Channel and no-harm principles.

The role of domestic legislation or regulation in this regard is, first and foremost, to limit or condition the use of AI to the ways and circumstances in which it is lawful to do so under international law. With respect to AWS, the ICRC recently proposed that certain technologies and uses should be prohibited because of the intolerable risk they pose.[215] In all other cases, precise regulations should establish 'limits on the types of target, …on the duration, geographical scope and scale of use, … on situations of use, … [and] requirements for human–machine interaction, notably to ensure effective human supervision, and timely intervention and deactivation.'[216] Thus, states should, and in some cases must, adopt domestic legislation requiring meaningful human control over AI systems.

Moreover, due diligence may require states to ensure effective accountability for any harm or risk caused by AI technologies.[217] This includes carrying out appropriate criminal or civil investigations and prosecutions.[218] States must also ensure that affected individuals are able to access justice and obtain an effective remedy for any harm they suffer, including by notifying individuals of the use of AI and disclosing all the necessary information.[219] To ensure that personal data used for training AI algorithms is not collected without individual consent and remains subject to its owner's control throughout the technology's life cycle, states should put in place robust data protection laws.[220]

While most AI algorithms are proprietary, their legislative or regulatory approval for use should also be subject to independent auditing by a trusted body, ideally the same public organ that grants the relevant patent or licence.[221] The same goes for their training data, which should, as much as possible, be taken from freely available, open-source platforms and comply with data protection

[211] Crootof, 'A Meaningful Floor for Meaningful Human Control Autonomous Legal Reasoning' (n 189) 54. Other definitions, attempted by various institutions, are reported at 56-57.
[212] ibid 62.
[213] Special Rapporteur on freedom of opinion and expression (n 130), paras 42-46.
[214] Calling for clear governance frameworks are necessary especially for any employment of AI in military settings, McGregor (n 155).
[215] ie 'AWS that are designed or used in a manner such that their effects cannot be sufficiently understood, predicted and explained' and any anti-personnel use of AWS. See ICRC, 'Position' (n 22), respectively at 7 and 9.
[216] ibid 2, sub 3.
[217] Access Now (n 1) 33, 35.
[218] HRC, 'General Comment 31' (n 147), paras 8, 15, 18.
[219] ICCPR art 2(3); Special Rapporteur on freedom of opinion and expression (n 130), paras 49-50, 59-60; HRC 'General Comment 31' (n 147), paras 15-20; Access Now (n 1) 33-35.
[220] Access Now (n 1) 30-32.
[221] Special Rapporteur on freedom of opinion and expression (n 130), paras 55-56; Access Now (n 1) 33-34, 36.

laws or standards.[222] To inform such approval decisions, states should lay out the appropriate technical standards, including by drawing on the relevant industry and academic expertise.[223]

Finally, states should consider putting in place a moratorium on the development, acquisition, sale, and deployment of *fully* automated weapons or military systems — especially 'AWS that are designed or used in a manner such that their effects cannot be sufficiently understood, predicted and explained'.[224] This ought to be so at least until further research and development are carried out to ensure that these systems work with the necessary levels of predictability and transparency.

d. Other measures

At least three more sets of due diligence measures could discharge the protective obligations analyzed above in the context of AI technology: capacity-building, institutional measures, and international cooperation.

A state's efforts to *build capacity* could go in three directions. First, they could be aimed at further research on AI. Whilst calls for further research often aim to develop and expand the use of AI technologies, particularly by enhancing their explainability and making strides towards fully autonomous AI, states should also be supporting research into the harms and risks posed by current and future AI technologies. In fact, all efforts to innovate in the field should be accompanied by internal and external research into their impact on individuals and society at large.

Secondly, AI operators, whether civilian or military, should be adequately trained on their machines' inner workings. Operators should be familiar with training datasets and the inherent limitations of AI algorithms, such as their lack of contextual and qualitative knowledge, their susceptibility to bias and their complexity. Most importantly, operators must be trained to make a meaningful and informed judgment about AI outputs before they can be used.

Thirdly, civil society must be made aware of those same limitations. To this end, states could organise awareness-raising campaigns on how AI works and other educational initiatives targeting various societal groups, especially those most affected, such as socially disadvantaged groups and war-torn populations. This should enable individuals, companies, and states to make better informed choices when using AI technologies, and to be better prepared against potential harms.[225]

Most of the measures examined so far require *institutional arrangements* to be put in place. For instance, states should establish standing or ad hoc bodies to independently audit and review AI technologies during their entire life cycle. They should also support non-governmental institutions, especially academic and civil society initiatives, in carrying out similar studies and reviews. To that end, public-private partnerships between governments, tech companies, and civil society organisations are essential to foster mutually beneficial research and ensure that all the necessary information is duly exchanged between partners. Likewise, states should set up independent committees or panels of experts to review the legality of civilian and military AI applications that they seek to study, develop, acquire, sell, or deploy. Lastly, to enable individuals and groups to effectively challenge the use of AI and obtain an effective remedy for any harm(s) they have suffered, states should explore accessible and scalable avenues to justice.[226] For instance, they could require or encourage companies to establish internal complaint mechanisms, such as ombudspersons, or offer alternative dispute settlement or prevention mechanisms, such as

[222] Access Now (n 1) 35-36.
[223] ICRC, 'Position' (n 22) 4; Special Rapporteur on freedom of opinion and expression (n 130), paras 48, 65.
[224] ICRC, 'Position' (n 22) 7; HRC 'General Comment 36' (n 136), para 65.
[225] Special Rapporteur on freedom of opinion and expression (n 130), paras 49-50, 59.
[226] ibid, paras 29, 33, 41, 60, 70.

mediation and conciliation, and harness the power of ICTs to enable individuals to access domestic courts at a low cost.

Finally, given the extremely unequal development of AI technologies among developed and developing countries around the world, diligent states are expected to *cooperate* with one another to prevent, mitigate and redress the harms and risks that those technologies might pose. On the one hand, states should agree on specific measures to address the AI threat landscape in multilateral settings.[227] Thus, they should continue to engage in the ongoing discussions before the CCW's Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems and aim to achieve consensus on the necessary levels of human control and accountability. Moreover, states should cooperate to prevent an AI arms race and redirect efforts to study and develop lawful civilian applications of the technology that complement rather than replace human judgement and intelligence. Examples include AI for medical diagnosis, energy efficiency, and climate change models. [228] Whilst some have called for the establishment of an ad hoc UN body for AI,[229] states could use existing UN bodies, such as the Human Rights Council, as a forum to discuss the necessary measures to address the impact of all AI technologies on individuals and societies, such as clarifying the status of international law on the matter. On the other hand, states should exchange information on the outcomes of their legal reviews of AI technologies and research into their impact, to prevent further harm to other states and individuals.

4. Conclusion

Without doubt, AI technology has numerous beneficial applications, from ubiquitous search engines and climate change models to medical diagnosis and treatments, such as cancer screening and prosthetics. AI-informed statistics also help us to make better-informed, more objective, and justifiable decisions which cannot be based on common sense alone. Nonetheless, caution is needed when we do not or cannot know how AI systems will behave in complex human, social or natural environments, or when the risks of resorting to such technology are higher than their anticipated benefits. Those who are careless in developing or employing AI technologies might be opening a pandora's (black) box which we know can be destructive — even if we do not always know exactly how.

As shown in this Chapter, a variety of international rules require states to exercise due diligence when designing and using AI systems in real-world situations. Though not prescriptive of their exact means of compliance, these various protective duties may be discharged by a range of measures available to states at different levels of development. They include AI impact or risk assessments, meaningful human oversight in the use of the technology, basic domestic regulation, efforts to build or enhance our capacity to understand how AI operates, institutional arrangements, and international cooperation.

Whether AI will leave us better or worse, states must 'handle it with care' and strive — to the extent feasible in the circumstances — to prevent harm (and the risk thereof) to other states and individuals, even if scientific knowledge and understanding lag. The greater the harm or risk, the higher the degree of diligence required. While this does not mean that progress in AI should be stalled, it requires states to choose carefully which of its applications entails acceptable and manageable risks.

---

[227] The US DoD, e.g., made international cooperation one of the pillars of its AI strategy, and in 2019 the EU — through its Finnish Presidency — stressed the topicality of international cooperation, including with NATO: Hill (n 9) 148. Examples of multilateral cooperation initiatives, and other measures (including training) led by NATO follow at 149.

[228] Access Now (n 1) 14.

[229] van den Meerssche (n 202).