Combat Code Compliance: International Humanitarian Law and The Development Stages of AI for Targeting

Yasmin Afina

A thesis submitted for the degree of PhD in Law

Essex Law School

University of Essex

November 2024

This page is intentionally left blank.

Summary

The integration of artificial intelligence (AI) into the military domain presents both significant opportunities and profound challenges. International humanitarian law (IHL) offers a critical framework to navigating the opportunities while addressing the risks associated with the deployment of these technologies for targeting. While much of the attention has been dedicated to the legal implications of using these technologies, this thesis seeks to complement these endeavours by offering insights on their pre-deployment stages.

To this end, this thesis is divided into four chapters. The first chapter provides readers with an overview of what 'artificial intelligence' concretely means and demystifies its applications in the military domain. By exploring the building blocks of AI and some of its key aspects and dimensions, the reader will be equipped with a strong foundational basis for the subsequent legal discussions. The second chapter looks at data as a key intervention point to foster upstream compliance, in the light of its critical role in shaping the performance and outputs of AI technologies. The third chapter delves into select dimensions pertaining to AI's design and development, namely the permissibility of developing such technologies for anti-personnel targeting, the role and responsibilities of 'developers', and the suitability of Article 36 legal reviews. The fourth and final chapter examines the question of uncertainty by formulating IHL's approach and the subsequent implications of AI's inherently probabilistic nature.

Ultimately, it is hoped that this work will not only reduce the downstream risks from the already-ongoing use of these technologies, but also help States in implementing their obligation to respect and ensure respect for IHL through compliance by design. As these technologies are set to play an increasingly prominent role in shaping the battlefield, it is now more critical than ever to dedicate efforts in ensuring IHL remains central to their development.

To my dear, beloved Ikung, who taught me the value of persistence, resilience, kindness, and to dare chart paths where no one else would.

Acknowledgments

Completing this PhD would not have been possible without the tremendous support provided by my entourage and those that I have met along this journey. While words will never do justice to their role, I hope these acknowledgments will somewhat capture a glimpse of my profound gratitude.

First and foremost, I am most grateful for the unwavering support, invaluable guidance and encouragement from my supervisors, Prof. Noam Lubell and Dr Daragh Murray. You have both been incredibly generous with your time and the wealth of knowledge shared over these past years, and I will always be grateful for the baggage you have equipped me with. Knowledge is only as good as one's ability to challenge, constructively, their own assumptions, and I feel most privileged to have been able to learn this from the best. Your thoughtful and nuanced feedback and input have consistently helped me push my boundaries, and I am grateful for the tremendous and consistent support while this PhD journey has also been infused with quite a few personal endeavours and challenges. Thank you for representing the very qualities of the researcher I strive to become – thorough and rigorous, collaborative and insatiably curious, inquisitive of the other's perspective and open-minded, as well as kind and thoughtful. The healthy and intellectually stimulating environment you create have left an indelible mark on both my professional development and my personal growth, and I hope this thesis only marks the beginning of our continued work in this space.

Second, I wanted to extend my most sincere gratitude to the University of Essex, which has played a tremendous role in my academic, professional and personal development since I first set foot in Wivenhoe Park in September 2013 as an LLB Fresher's. The environment provided by the University was uniquely conducive to nurturing the bold, the rebel, the daring. From providing students with an incredible wealth of knowledge to fostering a culture of inclusivity and mutual learning and understanding, the campus was also where I have made life-long

friends whose support have been key in helping me reach the finish line. I am most grateful for the opportunity to return in 2019 as a PhD student, and for entrusting me with the university's humanities doctoral scholarship. As a young woman of colour hailing from the Global South, this support has been particularly meaningful as I strive to make a positive contribution in what remains a largely Western, male-dominated field.

This PhD journey would not have been made possible without those who I have met from various walks of life and who were a driving force through their encouragement, support and guidance. John, Kerstin, Beyza, Patricia, Marjorie and Giacomo, thank you for your mentorship and friendship. Your unwavering dedication to the field of technology governance and international peace and security is admirable, and learning from your experiences and wisdom truly is a luxury. To friends and colleagues – this endeavour would not be as fulfilling and rewarding without everything I have learned from you and without all the shared moments. Wilfred, Alisha, Daniel, Lenka, Joanne, Claire, Jenny, Calum, Julia, Emma, Hallie, Amrit, Rowan, Nikki, Michael, Alexi, Magda, Netta, Kazuo, Bobby, Jingjie, Jamal, Jessica EA, Sarah, Federico, Edward and Jessica LA, and all I could not possibly name but whose absence from this list does not in any way reflect my lack of gratitude – thank you.

Special gratitude goes to my closest friends Yasmeen, Hanik and Chris. Hanik and Chris, I think it is no coincidence that we met from the very first week of university back in 2013, and I am grateful to have you by my side in my personal and professional growth. From our Model UN years at university to the heartwarming meals we shared and what should remain outside of an academic thesis, it truly is a privilege to be part of a decade of 'MCL', memes and friendship. Yasmeen, Duphe, I found a most profound, unbreakable bond in our friendship and sisterhood, and you of all people know that a mere few sentences will never do justice to how grateful I am for you. Thank you for all the love, for all the tears, for all the laughters and for all the soulful food, and it is an honour to share all my major milestones in life with you. Your

consistent support and unwavering care are the very reason why I have made this far, and I thank you, and Matt, for always making sure that my stomach, and heart, are full.

To my family – thank you. While I cannot possibly name all, I am most grateful for the support and love shown and shared over these twenty-nine years of existence. To Mama Atty and Yangti, Tante Ea, Om Peter, Bude Tian, Pakde Bobby, Bude Lily, Savanah, Om Pudja, Tante Nino, Tante Bonny, Bude Ila, Pakde Aung, Andung Odi, Kyran, Torin, Mbak Ica, Masqy, Nana, and all my 'keluarga besar' whose absence from this list does not in any way equate to a lack of gratitude – thank you. To Ikung, who has transitioned into the next realm along the way – I am profoundly grateful for your support through the values, invaluable lessons and tender memories we share. I may not be the engineer you had hoped me to be, but I wish to dedicate this quest for knowledge and impact to your legacy. To my in-laws, thank you for the regular check-in's and for always showing that you care. And to my 'adopted' family in Geneva, merci infiniment. All the soulful food, positivity and laughter shared with my beloved tante and om, and of course the unique friendship shared with my Indoluge-iar clique, will always be dear to me.

To my closest family – my parents, Maman Cessy and Papa Hedi, and my brother, Rayhan, your unconditional love and consistent support are most precious to me. Maman and Papa, thank you for always standing at the forefront through the ups and downs. Thank you for the infinite source of comfort and inspiration, both at home in the outside world. I will always strive to follow your footsteps and to learn from your dedication to our communities, our family, our faith and those most in need. Ray, while our relationship is (heavily) marked by what you would expect between a brother and a sister, part of it is that I will always be grateful to have you in my life. You are the very embodiment of persistence, and I will always cherish our heart-to-heart conversations. The home that you three provide give me much comfort, and I hope my work will always reflect all my love and dedication to you.

Last but not least, I wanted to dedicate this final paragraph to my dear husband, Dr Mohammed Azzouzi, and my most beloved daughter Soraya. Bou, you have consistently reminded me, even in the most challenging times, what I should be striving for in this PhD, and it is clear to me that I would not have been able to even start this thesis without your consistent encouragement, support and love. Through your patience and kindness, you have been my rock and you never failed to provide me with a moral compass in this challenging, at times existential dread-inducing field. And of course, never without a mental sparring or two, with humour and a touch of science fiction. Words will never do justice to my profound gratitude, and I feel most privileged to have been able to share this journey with you. May this PhD be an ode to our shared values, ambitions and dreams as we sail through life together. Soraya, you are the very reason why I will always strive for the best. As bleak and as turbulent today's world is, it remains our generation's imperative to ensure you inherit a peaceful, prosperous future. You came in the middle of this journey and you have blessed me, every single day, with so much love, hope and patience, and your thirst for knowledge – always with a quirky touch of cheekiness – will always remind me of what is important. May this journey and everything I have learned along the way help me be a better Mother to you. Terima kasih, sayangku.

Table of Contents

Abbreviations

AGI	Artificial general intelligence
AI	Artificial intelligence
AP	Associated Press News
API	1977 Additional Protocol I to the 1949 Geneva Conventions
ATR	Automated target recognition
AUKUS	Australia, UK, and US
BC	Before Christ
CCM	2008 Convention on Cluster Munitions
CCTV	Closed-circuit television
CCW	1980 Convention on Certain Conventional Weapons
CDAO	Chief Digital and Artificial Intelligence Office
DARPA	Defense Advanced Research Projects Agency
DIANA	Defence Innovation Accelerator for the North Atlantic
DoD	Department of Defense
EMP	Electromagnetic pulse
EU	European Union
GCHQ	Government Communication Headquarters
GDPR	General Data Protection Regulation
GGE	Group of Governmental Experts
GPS	Global positioning system
GSM	Global system for mobile communications
HRW	Human Rights Watch
HUMINT	Human intelligence
ICJ	International Court of Justice

ICRC	International Committee of the Red Cross
ICTY	International Criminal Tribunal for the former Yougoslavia
IDF	Israel Defense Forces
IHL	International humanitarian law
ILC	International Law Commission
INTERPOL	International Criminal Police Organization
IRA	Irish Republican Army
ISR	Intelligence, surveillance and reconnaissance
ISTAR	Intelligence, surveillance, target acquisition and reconnaissance
JAIC	Joint Artificial Intelligence Center
LAWS	Lethal autonomous weapons systems
MI5	Security Service
MSF	Médecins Sans Frontières
NATO	North Atlantic Treaty Organization
NSA	National Security Agency
NSAG	Non-State armed group
OECD	Organization for Economic Co-operation and Development
OTP	Office of the Prosecutor
REAIM	Responsible AI in the military domain
SIGINT	Signals intelligence
SIPRI	Stockholm International Peace Research Institute
TEVV	Testing, evaluation, verification and validation
UK	United Kingdom
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization

UNHCR	United Nations High Commissioner for Refugees
UNICRI	United Nations Interregional Crime and Justice Research Institute
UNIDIR	United Nations Institute for Disarmament Research
URSA	Urban Reconnaissance through Supervised Autonomy
US	United States
VUCA	Volatility, uncertainty, complexity and ambiguity
WMD	Weapons of mass destruction

TABLE OF CASES

TABLE OF INTERNATIONAL TREATIES AND CONVENTIONS

TABLE OF UN DOCUMENTS

TABLE OF INTERNATIONAL POLICY DOCUMENTS

TABLE OF REGIONAL TREATIES

TABLE OF NATIONAL POLICY DOCUMENTS

TABLE OF NATIONAL REPORTS

Introduction

1 Prologue

"*Hence to fight and conquer in all your battles is not supreme excellence; supreme excellence consists in breaking the enemy's resistance without fighting*." – Sun Tzu, The Art of War

Just like any technological revolution in the military domain, artificial intelligence (AI)'s effects will be far-reaching and diverse. In some respects, it holds promises of tactical and strategic opportunities leading to a frictionless victory; in others, it is the source of dread, worries and fears. Both are, by no means, mutually exclusive. The opportunities these technologies could offer in warfare are non-negligible. Acknowledged across the world, these range from enhancing existing intelligence, surveillance and reconnaissance (ISR) capabilities to increasing the efficiency of military logistics at scale and at speed.[1] The appeal of these opportunities does not preclude apprehension with regards to the risks stemming from these technologies' development, deployment and use. From concerns over an 'accountability gap' to agitation over the prospects of an 'inhumane' war, a host of worries are shared across communities and across geographies, albeit at varying degrees.[2]

Military AI is not the product of fiction anymore. In fact, various forms of autonomy have arguably 'been used in military systems for over seventy years', for instance in the form of homing munitions during World War II.[3] As at November 2024, reported use-cases of military AI in Gaza and Ukraine have made their way to the mainstream press. States are both

[1] Sarah Grand-Clément, 'Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain' (2023) UNIDIR 9 <https://unidir.org/publication/artificial-intelligence-beyond-weapons-application-and-impact-of-ai-in-the-military-domain/> accessed 11 November 2024

[2] Yasmin Afina, 'The Global Kaleidoscope of Military AI Governance: Decoding the 2024 Regional Consultations on Responsible AI in the Military Domain' (2024) UNIDIR 14 <https://unidir.org/publication/the-global-kaleidoscope-of-military-ai-governance/> accessed 11 November 2024

[3] Michael Horowitz and Paul Scharre, 'An Introduction to Autonomy in Weapon Systems' (2015) CNAS 8 <https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems> accessed 11 November 2024

developing such capabilities in-house, and/or procuring either from other States, or even technology companies, depending on needs and the availability of resources.[4] Growing realization that the train has indeed departed the station, and increased understanding of the issue over the past years, have led to today's rich and nuanced debate as to how these technologies should be developed, deployed and used. To some, this calls for a prohibition on specific applications such as autonomous weapons systems: this is the case of the Campaign to Stop Killer Robots, a movement of over a hundred advocacy organizations. To others, this calls for a nuanced approach based on the premise that these technologies should be developed, deployed and used responsibly: this is the case, among others, of States who have endorsed the 2023 Responsible AI in the Military Domain (REAIM) Call to Action, including China and the United States.[5]

While the international and multistakeholder community remains relatively divided on both the opportunities and risks presented from military AI, there is general agreement that international humanitarian law (IHL) applies in armed conflict including in the context of these technologies' development, deployment and use. Yet, despite this shared understanding, much uncertainty and disagreement remain as to what IHL provisions and considerations specifically apply in specific contexts and how. While academic literature is growing in this space, States' position and approach at the higher levels and in the public domain remain general and, at times, inconsistent.[6] While a handful of States have been developing their national position on

[4] Sam Biddle, 'U.S. Military Makes First Confirmed OpenAI Purchase for War-Fighting Forces' (*The Intercept*, 25 October 2024) <https://theintercept.com/2024/10/25/africom-microsoft-openai-military/> accessed 11 November 2024

[5] 'REAIM 2023 Call to Action' (*Government of the Netherlands*, 16 February 2023) <https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action> accessed 11 November 2024

[6] Netta Goussac and Magdalena Pacholska 'The Interpretation and Application of International Humanitarian Law in Relation to Lethal Autonomous Weapon Systems: Background paper on the views of States, scholars and other experts' (2025) UNIDIR 10 <https://unidir.org/publication/the-interpretation-and-application-of-international-humanitarian-law-in-relation-to-lethal-autonomous-weapon-systems/> accessed 6 June 2025

the interpretation of international law on this issue for a number of years, the majority of the international community remains yet to build their internal capacity on this front.

The present thesis seeks to make a meaningful contribution to this emerging understanding. Specifically, this thesis will focus on what are the relevant IHL considerations for the pre-deployment stages of AI for military targeting. By focusing on the pre-deployment stages, the present research aims to contribute to a much-needed body of knowledge for the responsible development, deployment and use of AI in the military domain. Not only would such understanding help inform States for their deliberations at the multilateral level; it ultimately aims to foster compliance with international law from the outset instead of being an afterthought. States procuring these capabilities, whether in-house or externally, will have some reference as to what are some of the criteria that must be met in order to ensure their inherent compliance and thus decrease downstream risks of IHL violations. Entities involved in the development of these technologies will know what IHL obligations their clients (i.e., States) are bound to and thus, will be incentivized by the idea of developing technologies more likely to pass compliance testing and evaluation. Furthermore, the focus on military targeting is deliberate due to the high stakes of such applications and their timeliness in today's conflicts. Achieving such understanding will, first, require taking stock on the background and wider context in which this discussion sits. A careful consideration of its scope will ensue, before then proceeding with this thesis' substantive chapters.

2 Background

2.1. Tracing Back Today's Artificial Intelligence Fever

Progress in the field of AI has kicked off a general frenzy from the promises held by the development and use of these technologies. This AI fever particularly skyrocketed over the past couple of years following the rapid and widespread adoption of OpenAI's ChatGPT. The latter led to a mainstreaming of generative models in the public with over 250 million reported

active users on a weekly basis. This has subsequently led to a dramatic uptick in investment in the companies developing these technologies, valued in the range of billions of dollars.[7] As their name suggests, these models generate novel data, including images (e.g., Dall-E, Stable Diffusion), text (e.g., ChatGPT, Claude), and audio (e.g., AudioCraft). These models are developed and used with a host of potential, from assisting education through enhanced human learning to aiding artistic creative processes, or even material discovery and optimisation to advance scientific research.[8] Despite the overwhelming focus on generative AI, artificial intelligence stretches far beyond these applications. Before ChatGPT was even launched in 2022, these technologies were already omnipresent in our daily lives, catering for a wide range of applications. From supporting search engines to air traffic management, AI also featured prominently (and, at times, controversially) in informing and even driving public health policies during the COVID-19 pandemic.[9]

AI's users expand beyond public bodies. At the individual level, for instance, smartphones are capable of notifying their user every morning, at a certain time, that it is time to go to work, and what is the best route to what the programme infers to be the user's workplace. This function is made possible from patterns the programme identifies and defines, based on data collected from the user's daily usage of their device: They may follow a certain routine, where they drive from point A to point B on a daily basis at a certain hour, following some preferred

[7] OpenAI has recently been reported to have raised USD $6.6 billion in funding, see: Dan Milmo, 'OpenAI raises $6.6bn in funding, is valued at $157bn' (*The Guardian*, 2 October 2024) <https://www.theguardian.com/technology/2024/oct/02/openai-raises-66bn-in-funding-is-valued-at-157bn> accessed 11 November 2024

[8] See, respectively, Lixiang Yan and others, 'Promises and challenges of generative artificial intelligence for human learning' (2024) 8 Nature Human Behaviour, <https://www.nature.com/articles/s41562-024-02004-5> accessed 11 November 2024; Juniper Lovato, Julia Witte Zimmerman, 'Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership and Fairness in AI Generative Art' (2024) 7(1) Proceedings of the AAAI/ACM Conference on AI, Ethics and Society <https://ojs.aaai.org/index.php/AIES/article/view/31691> accessed 11 November 2024; and Dhruv Menon and Raghavan Ranganathan, 'A Generative Approach to Materials Discovery, Design, and Optimization' (2022) ACS Omega <https://pubs.acs.org/doi/10.1021/acsomega.2c03264> accessed 27 February 2023

[9] Chenrui Lv and others 'Innovative applications of artificial intelligence during the COVID-19 pandemic' (2024) 3(1) Infectious Medicine <https://www.sciencedirect.com/science/article/pii/S2772431X24000091> accessed 11 November 2024

routes and with defined variables such as traffic and certain times of the year. In this sense, the navigation programme is designed to not simply help the user navigate, but to navigate better. Another example relates to 'smart' fridges: These appliances can be connected to the user's smartphone and would then notify them when a certain item usually present in the fridge (e.g. milk) is missing. This is done through image recognition programmes embedded into the fridge's system, coupled with a programme that identifies common patterns in the user's fridge composition, thus allowing it to identify particular objects that may be missing and would usually present, subsequently prompting the user to buy more of this missing item to avoid inconvenience.

It is expected that these technologies' role in society will only grow in prevalence. Beyond informing shopping lists and movie recommendations, AI is likely going to be increasingly informing the development, implementation and review of national, regional, and international policies.[10] This prospect is met both with optimism and scepticism.[11] While a number of commentators and analysts are calling this frenzy yet another 'tech bubble' and thus ephemeral, it is nevertheless generally accepted that AI is having, and will continue to have, a tremendous and even disruptive impact with profound and possibly long-lasting societal implications.[12]

At the macro level, AI is poised to have an impact on how our society is organized and functions (i.e., power relationships in the social contract); and the prevalence of these

[10] At the international level, for example, a growing movement is advocating for the use of artificial intelligence to help implement the UN's Sustainable Development Goals. See: Anna Visvizi, 'Artificial Intelligence (AI) and Sustainable Development Goals (SDGs): Exploring the Impact of AI on Politics and Society' (2022) 14(3) Sustainability <https://www.mdpi.com/2071-1050/14/3/1730> accessed 11 November 2024; and Jaron Porciello and others, 'Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning' (2020) 2 Nature Machine Intelligence <https://www.nature.com/articles/s42256-020-00235-5> accessed 11 November 2024

[11] Over the years, the concept of 'AI for good' has emerged as a driver for responsible innovation. There is, however, a growing number of commentaries pointing out the limitations and, at times, dangers of such narrative, see for example: Mirca Madianou, 'Nonhuman humanitarianism: when 'AI for good' can be harmful' (2021) 6 Information, Communication & Society <https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.1909100> accessed 11 November 2024

[12] Luciano Floridi, 'Why the AI Hype is Another Tech Bubble' (2024) 37 Philosophy & Technology <https://link.springer.com/article/10.1007/s13347-024-00817-w> accessed 11 November 2024

technologies will re-calibrate, or even disrupt our relationship with technology not only as individuals but also as a society.[13] Consequently, there has been increased scrutiny over the implications of AI as not only will we grow increasingly dependent on these technologies; their implications are far-reaching and effective solutions are necessary to ensure that these technologies are being developed responsibly and in compliance with existing laws. Governments, international organizations and civil society groups have all initiated discussions, deliberations and processes in various forms to formulate the necessary policy responses, and possible governance pathways framing the development, deployment and use of these technologies. In October 2023, State representatives of the Group of Seven (G7) have developed and released the Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems and the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.[14] Shortly, after, in November 2023, the United Kingdom (UK) hosted the UK AI Safety Summit, convening a number of countries, industry and civil society representatives to deliberate on safety measures for the development and monitoring of these technologies. An AI Safety Institute was subsequently established, kicking off the mushrooming of similar institutes across continents.[15] The Organisation for Economic Co-operation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) are also both known to have been convening deliberations on AI governance for a few years now, while the United Nations (UN) Secretary-General has appointed, also in late October 2023, his High-Level Advisory Body on Artificial Intelligence.

---

[13] Arthur Holland Michel, 'Recalibrating assumptions on AI' (2023) Chatham House 5 <https://www.chathamhouse.org/2023/04/recalibrating-assumptions-ai/about-author> accessed 13 December 2023

[14] 'G7 Leaders' Statement on the Hiroshima AI Process' (*Ministry of Foreign Affairs of Japan*, 30 October 2023) <https://www.mofa.go.jp/ecm/ec/page5e_000076.html> accessed 20 December 2023

[15] Renan Araujo, Kristina Fort, Oliver Guest, 'Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges' (2024) ArXiv <https://arxiv.org/abs/2410.09219> accessed 11 November 2024

At the micro level, the number of domain-specific efforts and initiatives for the governance of these technologies has also increased. From the education sector to healthcare, governance frameworks, policies and proposals have proliferated.[16] The areas of security and defence are far from being spared. For example, there is a growing body of research on the security implications of generative AI, both through intentionally malicious uses and unintentional or incidental risks.[17] While they fall outside the scope of the present thesis, it is also worth mentioning the number of initiatives undertaken for the governance of AI in law enforcement. While the number of academic literature on this front is growing, international bodies such as the United Nations Interregional Crime and Justice Research Institute (UNICRI) and the International Criminal Police Organization (INTERPOL) have collaborated to develop the Toolkit for Responsible AI Innovation in Law Enforcement.[18] In defence, and as this thesis will discuss a few times, several States-led governance discussions and parallel processes are underway and ever-growing since this PhD's start in 2019.

This relatively nascent governance fever is, however, not representative of the technological state of the art. Way before the current hype around AI, the idea, and appeal, of developing autonomy and other AI-enabled solutions by the military is evidenced from as early as the 1950s, when the United States' (US) Air Force funded a program for translation.[19]

---

[16] See for example Aashish Ghimire and John Edwards, 'From Guidelines to Governance: A Study of AI Policies in Education' (International Conference on Artificial Intelligence in Education, Recife, Brazil, 8-12 July 2024) <https://link.springer.com/book/10.1007/978-3-031-64312-5> accessed 11 November 2024; Shivansh Khanna, Shraddha Srivastava, 'AI Governance in Healthcare: Explainability Standards, Safety Protocols, and Human-AI Interactions Dynamics in Contemporary Medical AI Systems' (2021) 1(1) Empirical Quests for Management Essences <https://www.researchberg.com/index.php/eqme/article/view/166> accessed 11 November 2024

[17] Ardi Janjeva and others, 'The Rapid Rise of Generative AI: Assessing risks to safety and security' (2023) Centre for Emerging Technology and Security, The Alan Turing Institute, Research Report <https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai> accessed 20 December 2023

[18] On UNICRI and INTERPOL's Responsible AI Toolkit, see: 'The Toolkit for Responsible Artificial Intelligence Innovation in Law Enforcement' (*UNICRI*) <https://unicri.it/topics/Toolkit-Responsible-AI-for-Law-Enforcement-INTERPOL-UNICRI> accessed 11 November 2024; for an example of academic literature see: Mareille Kaufmann, 'Chapter 22: AI in policing and law enforcement' in Regine Paul, Emma Carmel and Jennifer Cobbe (eds), *Handbook on Public Policy and Artificial Intelligence* (Edgar 2024) <https://www.elgaronline.com/edcollchap-oa/book/9781803922171/book-part-9781803922171-31.xml> accessed 11 November 2024

[19] Margaret Boden, *Artificial Intelligence and Natural Man* (Basic, 1977) 360, as cited in Manuel DeLanda, *War in the Age of Intelligent Machines* (Zone Books, 1991) 214

Coincidentally, in the civilian space, the Dartmouth Summer Research Project on AI was held around the same time in the summer of 1956 – an event widely considered as the birth of artificial intelligence as a discipline, and when it is generally accepted that the term 'artificial intelligence' was first coined.[20] This event did not take place in a vacuum with, in fact, embryonic work generally recognized as done in the early 1940s eventually leading towards the advent of AI.[21] Akin to what many commentators qualify as 'disruptive', 'new', and/or 'emerging' technologies – from the nuclear, biology and chemical, to outer space, quantum or even the internet – technological progress in the civilian sphere is indeed often very much intrinsically intertwined with that of the military realm. On a more conceptual level, the idea of 'automating the process of thinking' has been present in the minds of philosophers and scientists from as early as the 17th century, with the idea of intelligent machines arguably going back centuries, or even millennia.[22]

2.2. Artificial Intelligence in Defence: Navigating Undefined Innovation

Despite these technologies' prevalence and the massive investment poured into their research and development, there is no consensus as to what 'artificial intelligence' concretely means. Governments and the research community alike are divided on its definition, its very notion, and its implications. States, or even regional and international organizations, are proposing their own definitions, while much disagreement remains on the concerns and implications posed by these technologies.[23]

---

[20] Stuart Russell, Peter Norvig (eds) *Artificial Intelligence: A Modern Approach* (3rd edn, Pearson 2016) 1

[21] Ibid, 16

[22] Martijn Kuipers, Ramjee Prasad, 'Journey of Artificial Intelligence' (2022) 123 Wireless Personal Communications 3277 <https://link.springer.com/article/10.1007/s11277-021-09288-0> accessed 5 January 2024; see also Genevieve Liveley and Sam Thomas, 'Homer's Intelligent Machines: AI in Antiquity' in Stephen Cave and Kanta Dihal (eds), *Imagining AI: How the World Sees Intelligent Machines* (Oxford University Press 2023) 25 <https://global.oup.com/academic/product/imagining-ai-9780192865366?cc=ch&lang=en&#> accessed 11 November 2024

[23] States' national policies and strategy documents on AI has dramatically increased over the years, where most of them propose a definition of what corresponds to 'artificial intelligence', see: Yasmin Afina 'Draft Guidelines for the Development of a National Strategy on AI in Security and Defence: A Policy Brief' (2024) UNIDIR <https://unidir.org/publication/draft-guidelines-for-the-development-of-a-national-strategy-on-ai-in-security-

As Chapter I will discuss in further detail, there are broadly two conceptual ways of defining or at least, describing AI:

- Artificial general intelligence (AGI): Alternatively referred to, at times, as 'true' AI or 'general' AI, these programmes are meant to exhibit general intelligence and cognitive capabilities similar to, or even surpassing, that of humans. In other words, AGI would correspond to the 'holy grail' of artificial intelligence, which a handful of technology companies are driven by. While AGI has not been achieved to this day (and whether it is indeed truly reachable or not remains up to much debate), it is generally expected to be equally, if not more intelligent than humans for the completion of all kinds of tasks by optimizing their learning and adapting their experiences and skills effectively.
- Narrow AI: Narrow AI is programmed to surpass a human's capabilities through higher efficiency, speed and performance. However, unlike AGI, this superiority is strictly limited to the specific and narrow tasks that it was designed for and trained to complete. Most of the AI technologies deployed and used today are narrow AI, whether it is to power search engines, navigation assistance, or speech recognition.

It is however important to note that, as of 2024, the difference between narrow AI and AGI is not as black-or-white as it may have been a few years ago, especially with the development of increasingly sophisticated systems. A couple of years ago, Google DeepMind released Gato,

and-defence/> accessed 11 November 2024. At the supra-national level, see for example European Union (EU) Regulation 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (EU AI Act) [2024] OJ L2024/1689. It is however worth noting that some international organizations have expressed reluctance in providing a definition of artificial intelligence; for example the United Nations Educational, Scientific and Cultural Organization (UNESCO) argued, in one of their flagship AI governance frameworks, that such definition would need to be changed over time 'in accordance with technological developments'. See: 'Recommendation on the Ethics of Artificial Intelligence' (2021) UNESCO <https://unesdoc.unesco.org/ark:/48223/pf0000381137> accessed 11 November 2024. On disagreements regarding these technologies' policy implications, concerns and necessary responses, see: Charlotte Stix and Matthijs M. Maas 'Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy' (2021) 1 AI and Ethics 267 <https://link.springer.com/article/10.1007/s43681-020-00037-w> accessed 11 November 2024

one same agent that can essentially 'play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens.'[24] Today, ChatGPT can generate outputs on a text-to-image basis: the system can generate the picture of a calico cat dressed as a US Supreme Court justice when prompted to. Not only does this require the system to understand the text in the prompt; it is also capable of generating how it conceptualizes specifically a cat with a tri-colour coat wearing the attire of a judge serving in the US Supreme Court. The same system can be used to solve equations, analyse multiple documents at once, suggest travel itineraries, as well as write and debug code on JavaScript. ChatGPT has been trained in such way that, based on the prompt, it is capable of producing multiple kinds of outputs. This marks a dramatic shift from pre-conceptions of what constitutes a narrow AI, which had much more limited capabilities even just a few years ago.[25]

Similarly, in defence, significant progress has been made despite a lack of consensus of what 'artificial intelligence' concretely means. As of early 2024, the United States' Department of Defense (DoD)'s research agency, the Defense Advanced Research Projects Agency (DARPA), is reported to be using AI, machine learning and autonomy in about 70% of its programmes 'in some form or another'.[26] Back in 2018 already, the same Agency had announced an investment of over USD $2 billion as part of its now archived 'AI Next'

[24] Scott Reed and others, 'A Generalist Agent' (*Google DeepMind*, 12 May 2022) <https://deepmind.google/discover/blog/a-generalist-agent/> accessed 11 November 2024

[25] This shift in the approach to what would constitute 'narrow' artificial intelligence was particularly striking in 2022 when Gato was first released by Google DeepMind, the reach of which however remained pretty limited to those actively following, closely, the artificial intelligence space; and then through ChatGPT when it was released in November 2022. While more on the difference between narrow AI and AGI will be discussed in Chapter 1, for an example of how narrow AI was previously seen, see: Andreas Kaplan, Michael Haenlein, 'Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence' (2019) 62(1) Business Horizons, <https://www.sciencedirect.com/science/article/pii/S0007681318301393> accessed 11 November 2024

[26] David Vergun 'DARPA Aims to Develop AI, Autonomy Applications Warfighters Can Trust' (*DOD News*, 27 March 2024) <https://www.defense.gov/News/News-Stories/Article/Article/3722849/darpa-aims-to-develop-ai-autonomy-applications-warfighters-can-trust/> accessed 12 November 2024

campaign.[27] The latter included the development and sophistication of new capabilities, including the advancement of AI technologies for 'multi-modality automatic target recognition'.[28] In other words, and in principle, the system could identify targets based on varying kinds of data combined altogether, e.g., satellite imagery combined with intercepted Global System for Mobile Communications (GSM) metadata and radar.[29]

Beyond States, military alliances are also keen on leveraging the potential AI technologies hold to support their operations. For example, the North Atlantic Treaty Organization (NATO)'s 2021 AI strategy, which has recently been updated in the summer of 2024, outlines both the international security challenges posed by these technologies and their opportunities.[30] Even before its strategy's adoption, NATO has been reported to collaborate with the private sector to develop AI capabilities, while member States have been advancing the Alliance's own capabilities.[31] The Netherlands does so by, for example, heading work on big data and AI for decision making within the NATO Science and Technology board. [32] Beyond purely military applications, NATO has also launched its Defence Innovation Accelerator for the North

[27] 'AI Next Campaign' (*DARPA*) <https://www.darpa.mil/work-with-us/ai-next-campaign> accessed 9 January 2020

[28] Ibid

[29] More on automatic target recognition will be discussed later in Chapter 2, Section 2. For now, it is worth noting that the sophistication of automated target recognition is not new and has in fact been the subject of much research for years. See for example: David Blacknell, Luc Vignaud, 'ATR of Ground Targets: Fundamentals and Key Challenges' (2013) EN-SET-172-2013-01, NATO Science & Technology Organization, <https://www.sto.nato.int/publications/STO%20Educational%20Notes/STO-EN-SET-172-2013/EN-SET-172-2013-01.pdf> accessed 12 November 2024

[30] Zoe Stanley-Lockman, Edward Hunter Christie 'An Artificial Intelligence Strategy for NATO' (*NATO Review*, 25 October 2021) <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html> accessed 2 February 2024; see also 'Summary of NATO's revised Artificial Intelligence (AI) Strategy' (*NATO*, 10 July 2024) <https://www.nato.int/cps/en/natohq/official_texts_227237.htm> accessed 12 November 2024

[31] 'Dataiku Selected As Platform For NATO's Allied Command Transformation Focusing On AI Projects' (*Dataiku*, 14 February 2020) <https://pages.dataiku.com/nato> accessed 20 May 2020; more recently, it has been reported that NATO has invested in a number of European tech companies as part of its one billion euro innovation fund dedicated to AI, among others. See: Martin Coulter 'NATO targets AI, robots and space tech in $1.1 billion fund' (*Reuters*, 18 June 2024) <https://www.reuters.com/technology/artificial-intelligence/nato-targets-ai-robots-space-tech-11-billion-fund-2024-06-18/> accessed 12 November 2024

[32] J.A.P. Antoine Smallegange and others, 'Big Data and Artificial Intelligence for Decision Making: Dutch Position Paper' (2018) STO-MP-IST-160, NATO Science & Technology Organization <https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-IST-160/MP-IST-160-PP-1.pdf> accessed 20 May 2020

Atlantic (DIANA) programme, with the aim of leveraging dual-use solutions (i.e., those with both civilian and military utility) including in the AI space.[33]

It is also important to note that this innovation, despite being undefined, does not happen in a vacuum. A number of States have already, or are at least starting to adopt governance frameworks surrounding the development, deployment and use of these technologies. The United States is generally seen as one of the pioneers in this space. Back in 2018 already, the DoD adopted a national AI strategy document: 'Harnessing AI to Advance our Security and Prosperity', which articulates the Department's approach and methodology for accelerating the adoption of AI-enabled capabilities.[34] This document is complemented by a number of others, including the Responsible AI Strategy and Implementation Pathway adopted in 2022, the DoD Directive 3000.09 which very specifically covers autonomy in weapon systems, as well as the more general Responsible AI Toolkit presented in 2023.[35]

On the other side of the pond, the United Kingdom (UK)'s Ministry of Defence (MoD) 'Defence Technology Framework' includes AI as one of the primary 'technology families' with potential impact in defence, and subsequently published in 2022 its Defence AI Strategy.[36]

---

[33] 'About DIANA' (*NATO*) <https://www.diana.nato.int/about-diana.html> accessed 2 February 2024

[34] U.S. Department of Defense (DoD), *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* (2019) <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF> accessed 10 March 2020

[35] Respectively, U.S. Department of Defense, DoD Responsible AI Working Council, *Responsible Artificial Intelligence Strategy and Implementation Pathway* (2022), <https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF> accessed 17 January 2024; US DoD, *Autonomy in Weapon Systems, Directive 3000.09* (2023) <https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF> accessed 12 November 2024; and US DoD, CDAO Public Affairs, *Responsible AI Toolkit* (2023), <https://www.ai.mil/blog_9_19_23_RAI_Toolkit.html> accessed 17 January 2024

[36] Interestingly, the Framework differentiates between 'AI, Machine Learning and Data Science (i.e. software)' from 'Autonomous systems and Robotics' as two separate technology families. See: U.K. Ministry of Defence, Defence Science and Technology, *Defence Technology Framework* (2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/830139/20190829-DTF_FINAL.pdf> accessed 10 March 2020. For the Defence AI Strategy, see: U.K. Ministry of Defence, *Defence Artificial Intelligence Strategy* (2022), <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf> accessed 17 January 2024

France's national AI strategy document identifies security and defence as one of its primary areas of focus, and stipulates that in the upcoming years, the use of AI will constitute a necessity in this area, citing for instance applications to enhance data processing.[37] The Ministry of Armed Forces has subsequently published, in 2019, a report with specific guidelines for action to be adopted for AI in support of defence.[38] In 2024, France's national strategy on AI in defence has been updated, with respect for international law featuring as one of its three key pillars for the responsible development of these technologies.[39]

2.3. Discussions on the international stage

Beyond national interest, developments and progress in the realm of AI quickly caught the attention of the international community to develop, or at least initiate processes towards the development of governance frameworks surrounding AI applications in security and defence. In fact, the push towards international governance processes and deliberations in this space came, interestingly, much earlier than at the national level for most States. Although the International Committee of the Red Cross (ICRC) has been raising concerns related to autonomous weapons systems from as early as in 2011,[40] this subject was first formally brought into discussions within the UN in 2013 by the then-Special Rapporteur on extrajudicial, summary or arbitrary executions.[41] Christof Heyns referred to 'lethal autonomous robotics' as

---

[37] Cédric Villani, 'Donner un Sens à l'Intelligence Artificielle : Pour une Stratégie Nationale et Européenne' (2018) Mission Parlementaire <https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf> accessed 15 November 2019, 219-221

[38] French Ministry of Armed Forces, AI Task Force, *Artificial Intelligence in Support of Defence* (2019) <https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf> accessed 17 January 2024

[39] Ambassador Camille Petit, Permanent Representative of France to the Conference on Disarmament, Remarks at the UNIDIR Side-Event on 'Guidelines for the development of national strategies on AI in security and defence' (New York City, 25 October 2024)

[40] International Committee of the Red Cross 'International Humanitarian Law and the challenges of contemporary armed conflicts' (2011) Report for the 31st International Conference of the Red Cross and Red Crescent <https://www.icrc.org/en/doc/assets/files/red-cross-crescent-movement/31st-international-conference/31-int-conference-ihl-challenges-report-11-5-1-2-en.pdf> accessed 24 May 2020

[41] United Nations General Assembly 'Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns' Human Rights Council 23rd Session, A/HRC/23/47 (2013) <https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf> accessed 10 March 2020

‘weapons systems that, once activated, can select and engage targets without further human intervention. They raise far-reaching concerns about the protection of life during war and peace. This includes the question of the extent to which they can be programmed to comply with the requirements of international humanitarian law and the standards protecting life under international human rights law.’[42]

Christof Heyns concluded his report by recommending that ‘States establish national moratoria on aspects of [lethal autonomous robotics (LARs)], and calls for the establishment of a high level panel on LARs to articulate a policy for the international community on the issue.’ In his report, Heyns elaborated, among others, on how statements from Governments with the ability to produce lethal autonomous robotics indicated, to the Special Rapporteur, that their use during armed conflict or elsewhere was not ‘currently’ envisioned; as well as how some of those he consulted argued these technologies could ‘never meet the requirements of international humanitarian law or international human rights law’. In addition, Heyns covered a number of elements including, but not limited to, the drivers of, and impediments to the development of these technologies; the potential use of these technologies vis-à-vis the decision to go to war or otherwise use of force; the use of these robotics during armed conflict; legal responsibility for their use; and their implications on states who do not hold such technologies.

Some states, however, manifested reluctance in pursuing discussions within the framework of the Human Rights Council (HRC), particularly as the subject matter pertains to security and defence – a field which States are usually reluctant to discuss in human rights fora. In 2013, the discussion was then subsequently ‘moved’ under the 1980 Convention on Certain Conventional Weapons (CCW)’s auspices, when the CCW Meeting of States Parties decided

[42] Ibid 1

that the Chairperson would convene, in 2014, an informal Meeting of Experts to discuss the questions related to emerging technologies in the area of lethal autonomous weapons systems (LAWS).[43]

Albeit the limited scope of the CCW discussions (i.e., 'lethal autonomous weapons systems'), the environment in which the discussions have been held was conducive to further research in the wider realm of military AI and their (potential) implications by states, international entities, as well as non-governmental organizations. For instance, the United Nations Institute for Disarmament Research (UNIDIR) has been extensively researching and informing international deliberations and States on the weaponization of increasingly autonomous technologies. The ICRC has been organising expert meetings on a regular basis and published meeting reports, which have informed states and the research community.[44] Furthermore, the CCW's unique rules of procedure allows for civil society participation in meetings, including the sessions held by the CCW's group of governmental experts (GGE) on this issue. The mandate currently held by the GGE at present, however, is limited to discussions on 'emerging technologies in the area of lethal autonomous weapons systems', without negotiating powers within this particular framework, to the regret of many.

Over the years, while the GGE discussions have grown in maturity, the issue of AI applications in security and defence has proliferated beyond the CCW's auspices. In October 2022, the HRC adopted Resolution 51/22, which calls for the Human Rights Council Advisory Committee to prepare a study examining the human rights implications of new and emerging technologies in

[43] Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects 'Final report' (2013) CCW/MSP/2013/10, 4 para 32 <https://documents.un.org/doc/undoc/gen/g13/646/33/pdf/g1364633.pdf> accessed 10 March 2020

[44] For instance, the ICRC held an expert meeting on 'Autonomous weapon systems technical, military, legal and humanitarian aspects' on 26-28 March 2014. The meeting report may be consulted at: 'Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects' (2014) Expert Meeting Report, ICRC <https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014> accessed 25 March 2020

the military domain – inevitably covering AI technologies.[45] More recently, the First Committee of the General Assembly, responsible for international security and disarmament affairs, approved in October 2023 a draft resolution calling for the UN Secretary-General to conduct a comprehensive study of LAWS.[46] Most recently in early November 2024, two draft First Committee Resolutions were adopted by the UN General Assembly: one on AI in the military domain and its implications for international peace and security, and the other on LAWS.[47] In addition, a number of States have attempted to initiate parallel deliberations and processes held outside of the UN. In February 2023, the Netherlands and the Republic of Korea have co-hosted the inaugural REAIM Summit in The Hague, the second iteration of which was held in Seoul in September 2024. In April 2024, Austria hosted a conference on lethal autonomous weapons systems, particularly aimed at aligning the international community to the LAWS resolution passed shortly beforehand that Austria itself shepherded the development, negotiation and eventual adoption of.

Yet, despite the mushrooming of many processes in this space, coupled with a growing desire from civil society and the majority of non-aligned movement States to move to formal negotiations for a treaty (with most of them advocating for a prohibition treaty), there is no such negotiations currently underway, whether it is for lethal autonomous weapons systems or more generally on the use of AI technologies by the military. The CCW GGE on LAWS remains the sole forum in which States have an actual, deliberative mandate that has been consolidated over the years, more recently in November 2023 following the CCW's Meeting

[45] UNGA 'Human rights implications of new and emerging technologies in the military domain' (2022) Human Rights Council 51st Session, A/HRC/RES/51/22 <https://undocs.org/A/HRC/RES/51/22> accessed 18 January 2024

[46] UNGA 'Lethal autonomous weapons systems' (2023) First Committee 78th Session, Resolution A/C.1/78/L.56 <https://undocs.org/A/C.1/78/L.56> accessed 18 January 2024

[47] Respectively, UNGA 'Artificial intelligence in the military domain and its implications for international peace and security' (2024) First Committee 79th Session, Resolution A/C.1/79/L.43 <https://undocs.org/en/A/C.1/79/L.43> accessed 12 November 2024; and UNGA 'Lethal autonomous weapons systems' (2024) First Committee 79th Session, Resolution A/C.1/79/L.77 <https://documents.un.org/doc/undoc/ltd/n24/305/45/pdf/n2430545.pdf> accessed 21 November 2024

of the High Contracting Parties.[48] The latter renewed and strengthened the GGE's mandate, which many see as the sole avenue where fulfilling the UN Secretary-General's objective of an international treaty on LAWS by 2026 would be realistic.[49]

The international community is, in this sense, divided as to what is needed, for the way ahead, to govern the development, testing, deployment and use of AI applications in security and defence. These differences are often politically motivated, with mainly States likely to develop, deploy and sell (if not already) such AI capabilities opposing to, or at least, sceptical about the adoption of such treaty. On the other hand, proponents of an international treaty, in particular those in favour of setting prohibitions on the development, deployment and use of select AI applications, are mainly those generally concerned about losing on the AI arms race and getting caught in the middle.[50] Regardless of where one stands in this debate, one thing that seems to draw consensus in the international community is the applicability of international humanitarian law in the development, deployment and use of military AI technologies. In 2019, the CCW GGE on LAWS adopted 11 guiding principles, the first of which specifically asserts that 'international humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems'.[51] Considering the application of the rule of consensus in the CCW's procedures (i.e., that all states parties must agree on the adoption of the text), this attests to the general sense that IHL plays indeed a key

[48] Meeting of the High Contracting Parties to the CCW 'Final Report' (2023) CCW/MSP/2023/7, 4 para 20 <https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Meeting_of_High_Contracting_Parties_(2023)/CCW_MSP_2023_7_Advance_version.pdf> accessed 19 January 2024

[49] United Nations 'A New Agenda for Peace' (2023) Our Common Agenda, Policy Brief 9, 27 <https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf> accessed 19 January 2024

[50] The representative of a State in an armed conflict against another (as of December 2023) voiced their concerns on the adversary's possession and use of advanced technological capabilities, including the use of AI and its impact on civilian populations and objects. The author cannot divulge the identity of those States given the use of the Chatham House rule at the convening in question.

[51] Meeting of the High Contracting Parties to the CCW 'Final Report' (2019) CCW/MSP/2019/9 Annex III, 10 <https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/343/64/PDF/G1934364.pdf?OpenElement> accessed 19 January 2024

role in framing and, eventually, restricting not only the use of military AI, but also to that of their development and all pre-deployment stages.

3. Scope: The Importance of AI Technologies' Pre-Deployment Stages

All research undertaken in the context of this thesis will revolve around the pre-deployment stages of AI technologies developed for military targeting. While the upcoming Chapter 1 will delve deeper into the breadth and depth of what AI and military targeting constitute respectively, the present section will attempt to justify the author's motivations to frame the present study around both aspects.[52] Beyond the recognition, by States and many other actors, that IHL is indeed central in framing the development of these technologies, the author is also of the conviction that an 'upstream' approach to their governance, in particular to ensure their legal compliance 'by design', is one that ought to be promoted and further explored. This section will also strive to unpack more in detail in what sense, and how, IHL can be applied to the pre-deployment stages of AI technologies – and the relevance of military targeting as a key, timely and contentious issue.

3.1. AI in military targeting: Applications beyond weapons

This thesis examines AI technologies for military targeting and as such, its scope will expand beyond weapons systems. In fact, most of the attention dedicated to AI in the military domain, whether it is in multilateral fora or in academic scholarship, has been dedicated to lethal autonomous weapons systems. While the landscape has now changed and especially over the past 2-3 years as the international and multistakeholder community's awareness and understanding on this issue are growing, the very focus on LAWS was particularly striking when this thesis was first started back in 2019. The CCW discussions on LAWS is the most

[52] For a more elaborate discussion on what 'artificial intelligence' concretely means, see Chapter 1, Section 2. For a more elaborate discussion on what military targeting consists of, see Chapter 1, Section 3.1, as well as Section 3.2 for both the potential uses and issues arising from the use of AI for military targeting.

evident example of such focus, an over-ten years-old endeavour States and civil society alike are still grappling with. While there is not even a universally accepted definition of what LAWS constitute, such applications remain very much limited and would exclude other AI applications that are as much relevant and bring as many legal considerations in the context of military targeting. This very limited scope of the LAWS framing is for example reflected in China's position: The latter has expressed its view, within the CCW GGE, that LAWS should be prohibited. However, their definition of LAWS is one that would be exclusively limited to 'fully autonomous lethal weapon systems', thus excluding technologies that are not 'fully autonomous' and that, realistically, is very unlikely to be developed and deployed in the first place by armed forces.[53] This narrow definition, other than being a deliberate, political choice, is also one that leaves little room for granular deliberation as to how applicable laws, including IHL, should be applied for systems that fall under this high threshold. Surveillance and intelligence collection systems, for example, may not necessarily fall within the scope of LAWS, yet would still have a number of legal implications especially as their outputs may inform potentially lethal decision-making.[54]

An intelligence collection programme can be deployed to help identify targetable fighters in armed conflict, thus playing a critical role in military targeting without necessarily being part of a weapon system. One thing to note here is that, unless stated otherwise, the present thesis will often refer to 'targetable fighters' as encompassing both combatants in an international armed conflict, and civilians who have lost their protected status in non-international armed conflict, i.e., civilians directly participating in hostilities. There are debates as to what would

[53] Group of Governmental Experts 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 'Position Paper: Submitted by China' (2018) CCW/GGE.1/2018/WP.7, 1 para 3 <https://undocs.org/CCW/GGE.1/2018/WP.7> accessed 5 February 2024

[54] Grand Clément (n 1); see also Yasmin Afina 'Intelligence is dead: long live Artificial Intelligence' (*Chatham House*, 14 July 2022) <https://www.chathamhouse.org/2022/07/intelligence-dead-long-live-artificial-intelligence> accessed 5 February 2024

constitute the latter category, i.e., what would constitute a civilian directly participating in hostilities. Divergences exist, for example, between the ICRC's position on continuous combat function on the one hand, and a number of States' approach with regards to membership on the other hand. This PhD does not seek to find common ground on this issue, which would fall outside of its scope: it goes with the assumption that there will be a category of targetable individuals in non-international armed conflict, while leaving their status' defining criteria up to the discretion of the relevant organizations.

In this sense, it is important, for the author, that academic scholarship in the field of AI in security and defence expands beyond the restricted scope of lethal autonomous weapons systems. Not only will a widened scope of research help move past definitional concerns; it will also add much-needed nuance and, hopefully, feed into policy deliberations in this space. In fact, there is a growing understanding that the development and use of AI for military applications are not necessarily inherently unlawful, and that there is a need to move beyond the weapons systems discourse in the light of the far-reaching integration of AI across the security and defence ecosystem.[55]

3.2. The case for an anticipatory approach to legal compliance

This thesis will specifically focus on the pre-deployment stages of AI technologies designed to assist with military targeting operations. It is hoped that by focusing the analysis on the earlier stages of these technologies' lifecycle, this work will provide a contribution to efforts at promoting what may be called an anticipatory approach to legal compliance. As states and a number of industries seek, or even race towards the development of AI-enabled solutions for security and defence applications, many civil society and advocacy groups, along with certain States, are sounding the alarm on the downstream risks stemming from these technologies'

[55] Frank Slijper 'Slippery Slope: The arms industry and increasingly autonomous weapons' (2019) PAX for Peace 4 <https://www.paxforpeace.nl/publications/all-publications/slippery-slope> accessed 11 November 2019; see also Grand-Clément (n 1)

deployment and use in the battlefield. From fears of States losing on the potential power dynamics that may unfold from these technologies' development and deployment, to (perhaps more importantly) shared concerns of IHL violation and potential harm on civilians, the international community is actively exploring a number of options for the governance of these technologies' development, deployment and use in security and defence.

One relatively under-explored approach is to look at their pre-deployment stages and address, or at least, reduce risks of non-compliance from the outset. This approach seeks to anticipate some key, downstream IHL concerns that stem from AI-enabled technologies for military targeting, and subsequently find ways to address them from their embryonic stages (i.e., their conceptualisation, design, development, testing, evaluation, verification and validation or, in other, colloquial words, 'from the lab'). This thesis will thus present some of the author's findings what these key IHL issues are, and how the law can specifically be leveraged to inform and eventually shape, or even restrict, these technologies' conception. The objective is that, on the assumption that certain AI technologies are bound to be researched, and developed to assist with military targeting, they will be deployed with confidence that they are reliable for the purpose of compliance with applicable laws, and risks of IHL violation have been minimised.

This approach draws inspiration, among others, from the many anticipatory efforts undertaken for civilian applications of AI technologies. There has been a particular push for this approach through AI safety considerations, which the UK is particularly known to have championed when it held the 2023 AI Safety Summit at Bletchley Park. Many efforts are indeed underway to ensure the safety of AI technologies from their design, development and testing stages, particularly those considered as 'foundation models' or 'frontier models'. The former has been defined, by the UK Competition & Markets Authority, as a 'machine learning model which is

trained on vast amounts of data and can be adapted to a wide range of tasks and operations.' [56] In the same vein, the latter has been defined as 'highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models.'[57] The deployment and use of both are set to have far-reaching and profound implications; thus a number of solutions have been put forward to ensure the safety of these models 'by design'. Partnership on AI, a renowned coalition of industries, civil society and academic institutions has, for example, developed a framework to guide the responsible development and deployment of a range of AI models.[58] Major industry players have taken the initiative to research and present their perceived risks on public safety, as well as recommendations to address those risks – including through pre-deployment risk assessments and continuous auditing throughout the technology's lifecycle.[59]

In the security and defence sector, a similar trend has emerged through the promotion of responsible behaviour in the development, deployment and use of AI technologies. As mentioned in the previous section, the Netherlands and the Republic of Korea have launched the REAIM initiative, aimed to be a regular occurrence in promoting, as its name suggests, responsible AI in the military domain.[60] This focus on the promotion of responsible behaviour has gained particular traction to not only frame AI discussions, but also in other security fields such as cyber and outer space affairs.[61]

---

[56] 'AI Foundation Models: Initial Report' (2023) Competition & Markets Authority 8 para 2.2 <https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf> accessed 23 January 2024

[57] Government Office for Science (UK) 'Future Risks of Frontier AI: Which capabilities and risks could emerge at the cutting edge of AI in the future?' (2023) Technology & Science Insights and Foresight 31, 4 para 2 <https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf> accessed 23 January 2024

[58] 'PAI's Guidance for Safe Foundation Model Deployment: A Framework for Collective Action' (Partnership on AI) <https://partnershiponai.org/modeldeployment/> accessed 23 January 2024

[59] Markus Anderljung and others, 'Frontier AI regulation: Managing emerging risks to public safety' (2023) OpenAI <https://openai.com/research/frontier-ai-regulation> accessed 23 January 2024

[60] Section 1.3, Chapter 1

[61] Izumi Nakamitsu, Under-Secretary-General and High Representative for Disarmament Affairs, United Nations Office for Disarmament Affairs at the REAIM session on 'AI, Or Not AI? – Debunking Commonly-Held Assumptions on Military AI' (The Hague, 15 February 2023)

Similarly, the notion of reliability has emerged – at least, in informal settings – as a framework to promote legal compliance from the technology's pre-deployment stages.[62] There are a number of ways of defining reliability; this diversity spans across disciplines. In engineering, there are two distinct models of system reliability. First, it can either be binary: in other words, either the machine is working, or has failed. Second, the model may be based on the understanding that reliability can be at multi-state: in other words, 'both the system and its components are allowed to experience more than two possible states.'[63] For instance, the machine could be completely working, or it could be partially working, or partially failed, or even completely failed. Reliability can also mean different things across fields. From an operational perspective, reliability could correspond to that of ensuring mission success.[64] From a legal perspective, reliability could correspond to minimum compliance with existing laws, which may at times raise tensions with operational considerations. Ultimately, one of the main motivations of the present thesis is to bridge the discussions between the technical and the legal realms, and ultimately provide concrete recommendations to foster compliance with IHL in the age of military AI. As such, reliability serves as a useful concept to connect the technical characteristics of AI on the one hand, with legal compliance on the other hand.

However, constraining the thesis' scope to the concept of reliability would have raised too many definitional issues that could overshadow, or even undermine the core, legal issues that this thesis treats. In the absence of a universally agreed definition on what reliability constitutes – and even more so from a legal perspective – there would have been risks that the discussions would have deviated too much and too often into a reflection on what reliability would mean in specific contexts, at the expense of the concrete, practical analysis and recommendations

[62] For example, a number of state representatives have promoted reliability as a useful approach at a workshop the author attended under the Chatham House rule in Stockholm, on 22-23 January 2024.
[63] Ming J. Zuo, Jinsheng Huang and Way Kuo 'Multi-state *k*-out-of-*n* Systems' in Hoang Pham (ed), *Handbook of Reliability Engineering* (Springer 2003) 3
[64] This was discussed, again, at the workshop the author was invited to, held under the Chatham House rule, on 22-23 January 2024 in Stockholm.

that the author seeks to offer throughout this thesis. In this sense, narrowing down this study's scope to the development stages of AI technologies for military targeting would still serve, very well, the overarching purpose of this thesis while putting aside all concerns with regards to the concept of reliability which, in itself, could perhaps serve as a focus area for subsequent research in this space.

3.3. The applicability and centrality of international humanitarian law in the development of AI technologies for military targeting

Beyond the author's personal motivations and interests, it is also important to examine in the first place the relevance of international humanitarian law in the development of AI technologies for military targeting. All of the following aspects will be discussed in further detail throughout this thesis; however, it would be important to at least provide a general overview, from this early stage, on the legal landscape and the nexus with the development stages of AI-enabled technologies for military targeting. In fact, not only will establishing the relevance of IHL consolidate the arguments of the present thesis, but this exercise will also assert the credibility and strength of IHL as a governance framework for these technologies' development in the eyes of the wider, non-legal community such as in policy and technical circles. While many of these legal considerations will be unpacked in further details throughout the chapters ahead, there are a number of ways through which IHL is relevant and, in fact, central to these technologies' early stages.

First, there are statutory provisions that directly pertain to the upstream compliance of military capabilities. First, Article 36 of the 1977 Additional Protocol I to the 1949 Geneva Conventions (API) specifically provides:

> '*In the study, development, acquisition or adoption of a new weapons, means or method of warfare, a High Contracting Party is under an obligation to determine whether its*

employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.'

The status of this provision is contested as many, including the United States, consider the provisions of Article 36 as not reflective of customary international law (in addition to Additional Protocol I not being universally adopted and ratified by states – including by the United States). Yet, there is evidence that the majority of states – including the US – would still undertake legal review processes prior to the deployment of new weapons, albeit sometimes under the framework of national policies as the US does, not in the spirit of Article 36 legal reviews.[65] Further discussion as to whether Article 36 legal reviews are, in fact, sufficient and appropriate for the purpose of AI development and deployment will be unpacked in further detail in Chapter 3.

Essentially, these legal reviews are two-pronged. First, states ought to assess, prior to their deployment and eventual use, whether the new weapons, means or method of warfare in question are unlawful by nature. These inherent prohibitions could, for one, stem from international treaties specifically banning the development and use of certain capabilities: This is for example the case of 'laser weapons specifically designed, as their sole combat function or as one of their combat functions, to cause permanent blindness to unenhanced vision'.[66] This is also the case for cluster munitions, the development, production, acquisition, stockpiling, retention and transfer being prohibited under all circumstances within the framework of the Convention on Cluster Munitions (CCM).[67] It is however important to note that these treaties

[65] The US has, in fact, an entire chapter dedicated to the legal review of weapons that employ cyber capabilities in its DoD Law of War Manual. See: Department of Defense (US), Office of General Counsel, *Law of War Manual* (2015, updated 2023) 1038-1039 <https://media.defense.gov/2023/Jul/31/2003271432/-1/-1/0/DOD-LAW-OF-WAR-MANUAL-JUNE-2015-UPDATED-JULY%202023.PDF> accessed 6 February 2024

[66] Additional Protocol 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 (Protocol IV, entitled Protocol on Blinding Laser Weapons) (adopted 13 October 1995, entered into force 30 July 1998) 1380 UNTS 370 (CCW Additional Protocol IV) art 1

[67] Convention on Cluster Munitions (adopted 30 May 2008, entered into force 1 August 2020) 2688 UNTS 39 (CCM) art 1

have limited ratification and as such, their inherent prohibition would only be applicable to State parties. Second, legal review processes ought to assess the circumstances under which the use of the weapon system, means or methods of warfare in question would violate the applicable laws: In other words, this second prong consists of identifying the red lines not to cross with regards to legal compliance. Such red lines can, for example, stem from specific contexts under which the use of certain weapons could be prohibited under international treaties the state is party to: This is for example the case of the CCW's third protocol, prohibiting the use of incendiary weapons where civilian population as such, individual civilians or civilian objects are the object of the attack – conversely, the use of such weapons could be permissible against military objectives far from concentrations of civilians.[68]

Beyond international treaties prohibiting and/or restricting the development, deployment and use of certain weapons, a number of additional considerations ought to be factored into the legality assessments. This includes, for example, the weapon's characteristics and the subsequent ability to use it in compliance with the rule of distinction: The International Court of Justice (ICJ) has, indeed, held in its Advisory Opinion on the *Legality of the Threat or Use of Nuclear Weapons*, that 'methods and means of warfare, which would preclude any distinction between civilians and military targets, or which would result in unnecessary suffering to combatants, are prohibited.'[69] Beyond the rule of distinction, the adjacent, customary rules pertaining to the conduct of hostilities ought to be considered too, including proportionality and precautions in attack. [70] This is further complemented by the provisions made under Article 35 API, specifically:

[68] 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 as amended on 21 December 2001 (Protocol III) (adopted 10 October 1980, entered into force 2 December 1983) 1342 UNTS 137 (CCW Protocol III) art 2

[69] *Legality of the Threat or Use of Nuclear Weapons* (ICJ Advisory Opinion) 1996 <https://www.icj-cij.org/files/case-related/95/095-19960708-ADV-01-00-EN.pdf> accessed 21 November 2024 [95]

[70] These three IHL principles and requirements are particularly recognized in the context of the CCW GGE on LAWS, see for example: Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons

'*1. In any armed conflict, the right of the Parties to the conflict to choose methods or means of warfare is not unlimited.*

2. It is prohibited to employ weapons, projectiles and material and methods of warfare of a nature to cause superfluous injury or unnecessary suffering.

3. It is prohibited to employ methods or means of warfare which are intended, or may be expected, to cause widespread, long-term and severe damage to the natural environment.'

As such, there are clear legal provisions that would subsequently require, for the purpose of their implementation and operationalisation, interventions from the early stages of the design, development, training and testing of AI-enabled capabilities meant to serve as means and methods of warfare. How they translate and their intricacies will be discussed in subsequent chapters, such as with regards to data practices and the rule of precautions, as well as the appropriateness of Article 36 legal reviews.[71]

Another, and perhaps more general and high-level, relevant legal consideration pertains to the positive obligation by all states to respect and ensure respect for international humanitarian law at all times. The four 1949 Geneva Conventions begin, indeed, with the following provision:

'*The High Contracting Parties undertake to respect and to ensure respect for the present Convention in all circumstances*.'[72]

Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 'Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System: Chairperson's Summary' (2021) CCW/GGE.1/2020/WP.7, 3 para 3 <https://documents.un.org/doc/undoc/gen/g21/090/11/pdf/g2109011.pdf> accessed 12 November 2024

[71] See, respectively, Chapter 2, Section 2.1.2. and Chapter 3, Section 4

[72] Geneva Convention for the amelioration of the condition of the wounded and sick in armed forces in the field (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31 (Geneva Convention I) art 1; Geneva Convention for the amelioration of the condition of the wounded, sick and shipwrecked members of the armed forces at sea (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 85 (Geneva Convention II) art 1; Geneva Convention relative to the treatment of prisoners of war (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 135 (Geneva Convention III) art 1; Geneva Convention relative to the treatment of prisoners of war (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 287 (Geneva Convention IV) art 1

This provision is generally seen as of customary rule – and even as an obligation *erga omnes* by many.[73] It further reinforces the case for early interventions, from the embryonic stages of AI-enabled technologies, if these actions are in fact *sine qua non* to their subsequent compliance, post-deployment, with the applicable laws. Again, how this obligation pertains specifically to the development of AI technologies for military targeting will be discussed later, particularly as part of reflections on the legal basis for an iterative legal review process.[74]

3.4. Exclusion from the pre-deployment stages

The present thesis focuses on the pre-deployment stages for the reasons stated in earlier sub-sections. This implies the subsequent exclusion of a number of considerations that, albeit some of them being very much present in most law and policy discourses surrounding AI in the military domain, fall outside of this thesis' scope. This exclusion notably includes discussions surrounding humanity as grounds for a prohibition on anti-personnel AI.

Humanity has substantially guided and framed the discussions surrounding military AI governance, particularly in the advocacy of specific prohibitions. In fact, the motive and drive behind most of advocacy groups and the position of certain countries fundamentally lie on the (lack of) humanity in discussions surrounding military AI, and the need to preserve human dignity.[75] Beyond advocacy movements, and regardless of where one stands in the debate on how to regulate military AI, the concept of humanity has had a deeply influencing role in shaping policy and academic discussions that transcend disciplines. For example, in human rights, there is argument for the right to a dignified life in the light of the development of

[73] Michael N. Schmitt, Sean Watts 'Common Article 1 and the Duty to "Ensure Respect"' (2020) 96 Int'l L. Stud. 695 <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=2936&context=ils> accessed 7 February 2024

[74] Chapter 3, Section 4.2.2.

[75] 'Losing Humanity: The Case against Killer Robots' (*Human Rights Watch*, 19 November 2012) <https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots> accessed 18 November 2022

autonomous weapons.[76] Recently, the Human Rights Committee asserted the obligation of 'States parties engaged in the deployment, use, sale or purchase of existing weapons and in the study, development, acquisition or adoption of weapons, and means or methods of warfare' to 'always consider their impact on the right to life.'[77] In ethics, a number of studies have examined the interplay between autonomous weapons systems and human dignity, and a growing number of research is dedicated to this issue.[78]

In international humanitarian law, arguments have also been made with regards to the role of humanity especially through the Martens Clause.[79] Re-affirmed by the ICJ, the Martens Clause specifically establishes humanity and the dictates of public conscience as the minimum threshold parties to a conflict must observe, even in the absence of written norms and treaties in some matters.[80] The Court has, in fact, found the Martens Clause to be an 'effective means of addressing the rapid evolution of military technology' by maintaining customary principles of humanity as the minimum yardstick applicable to these inventions.[81] Proponents of a ban on autonomous weapons systems have used this as a case for a prohibition on the development of

[76] Christof Heyns 'Autonomous weapons in armed conflict and the right to a dignified life: an African perspective' (2017) 33(1) South African Journal on Human Rights <https://journals.co.za/doi/abs/10.1080/02587203.2017.1303903> accessed 18 November 2022

[77] International Covenant on Civil and Political Rights, 'General comment No. 36, Article 6: right to life' (2019) Human Rights Committee, CCPR/C/GC/36, para 65 <https://documents.un.org/doc/undoc/gen/g19/261/15/pdf/g1926115.pdf> accessed 29 September 2024

[78] Amanda Sharkey 'Autonomous weapons systems, killer robots and human dignity' (2019) 21 Ethics and Information Technology, <https://link.springer.com/article/10.1007/s10676-018-9494-0> accessed 18 November 2022; see also Michael C. Horowitz 'The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons' (2016) 145(4) Daedalus <https://direct.mit.edu/daed/article/145/4/25/27111/The-Ethics-amp-Morality-of-Robotic-Warfare> accessed 18 November 2022. For a recent example of leading work in this space, see: Neil Renic and Elke Schwarz 'Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing' (2023) 37(3) Ethics & International Affairs <https://www.cambridge.org/core/journals/ethics-and-international-affairs/article/crimes-of-dispassion-autonomous-weapons-and-the-moral-challenge-of-systematic-killing/ECDC02C13BA23C432C9B11D84ACBB303> accessed 30 September 2024

[79] Theodore Mero 'The Martens Clause, Principles of Humanity, and Dictates of Public Conscience' (2000) 94(1) American Journal of International Law <https://www.cambridge.org/core/journals/american-journal-of-international-law/article/abs/martens-clause-principles-of-humanity-and-dictates-of-public-conscience/F55EECE5BED3DDB9D78162DA4509A03A> accessed 18 November 2022

[80] *Legality of the Threat or Use of Nuclear Weapons* (n 69) [78]

[81] Ibid

AI-enabled programmes designed to decide on whether or not to engage, in lethal force, with an individual target.[82]

While there are interesting and at times compelling legal and philosophical arguments as to the importance of maintaining human control, a more detailed discussion on this aspect would fall outside of this thesis' scope. In fact, much of this debate would relate to use, while the present thesis seeks to focus on the design and development stages and how they might be IHL-compliant. As to the role of humanity in the pre-deployment stages, building on shared narratives by States, this thesis has been written on the assumption that the development, testing and evaluation of these technologies will never be fully autonomous.[83] Moreover, given what seems to be an emerging consensus on retaining some form of human control during deployment, this thesis proceeds on the assumption that human commanders will play a role in the deployment and use of the systems. Accordingly, insofar as considerations of humanity are relevant in this thesis, it is in the parts that reflect on the challenges of the human-machine interaction (such as automation bias and black boxes), and how to ensure that the development of these systems proceeds in such manner as to allow for human judgment to be effective and minimise the risks that may arise.[84] As such, discussions on the underlying legal considerations surrounding concepts akin to humanity, as salient as they are in governance discussions, will be excluded from pre-deployment in the light of their centrality to use.

[82] One prominent example of such proponents is the Holy See, having argued for the need for 'human qualities' for decisions over life and death ever since the embryonic stages of multilateral discussions on autonomous weapons; and eventually used the 'dictates of morality and public conscience' and the Martens Clause to consolidate their position. See: Diego Mauri 'The Holy See's Position on Lethal Autonomous Weapons Systems: An Appraisal through the Lens of the Martens Clause' (2020) 11(1) Journal of International Humanitarian Legal Studies, <https://brill.com/view/journals/ihls/11/1/article-p116_116.xml?ebody=pdf-60564> accessed 18 November 2022

[83] More on the role of humans in the development of military AI technologies will be discussed in Chapter 3, Section 3 and Chapter 4, Section 1.2.2

[84] Automation bias is discussed in Chapter 4, Section 1.2 and the black box issue is discussed in detail in Chapter 4, Section 1.3

4. Methodology, Structure and Guiding Questions

4.1. Methodology

Exploring and addressing the international humanitarian law considerations for the development of AI technologies for military targeting required an interdisciplinary approach: While, ultimately, this thesis is of legal nature, the qualitative research that underpins its discussions and findings necessitated a combination of the following elements: multidisciplinary desk research, informal consultations, and external engagements.

First, this thesis relies heavily on multidisciplinary desk research. This thesis was motivated by the desire to 'connect the dots' and look at what constitutes a timely issue from a legal perspective – an area that, in the author's opinion, would merit more attention and granularity in the light of its relevance in today's conflicts and the implications the development, deployment and use of these technologies may pose. As such, while reflecting on possible applications of AI in the military domain, the author sought to identify key IHL issues (e.g., distinction) and relate those with existing legal scholarship and, if relevant, cases. In recognition of the growing number of legal reflections done in this space with regards to both the law as it is and the law as it should be, the author would navigate these dynamics by contextualizing these sources against the wider policy landscape, which will be treated shortly. The thesis overall does not seek to provide any statement of what the law should be in the author's opinion but rather, ground considerations for the development of military AI technologies in the law as it is.

In recognition of artificial intelligence's inherently technical nature and the pace at which it is being developed and growing in sophistication, it was also important that a knowledge base on its technological dimension would be established. This explains the decision to dedicate Chapter 1 to technological foundations, not only reflecting the author's journey in deepening her technical understanding to better engage with the legal implications and considerations that

may arise from their development, deployment and use for military targeting, but also in ensuring readers from all disciplines are able to explore the thesis with a level playing field on their minimum technical knowledge.

Furthermore, the research underpinning this thesis also relied, heavily, on policy research. Specifically, the desk research focused on two main types of sources: policy documents available publicly (e.g., statements and national positions), and what is often referred to as 'grey literature' encompassing both academic writings and reports from think-tanks, research institutes, civil society organizations and the work of industries in this space. This choice was not only in recognition of the multistakeholder nature of the issue, where States and regional organizations seek to develop norms and potential regulations for the governance of these technologies, industries and research laboratories particularly from the civilian space are leading on cutting-edge innovation, and academics and research institutes build expertise in this space. Infusing legal research with policy considerations would also ensure that these discussions and reflections would be and remain relevant, cognizant of the many on-going processes unfolding in this space and the variety of perspectives and approaches that exist. A comprehensive awareness of the many perspectives surrounding policy discussions will ensure the research undertaken in this thesis are useful and concrete, paving the way for its findings' direct application to inform policy choices and decisions, and most importantly concrete and feasible. This will ultimately help the author's quest in ascertaining the centrality of international law in the governance of AI in the military domain beyond theory but also in practice.

Second, over the five years and a half of research undertaken to write this thesis, the author had the privilege of conducting a series of informal consultations with different stakeholders to inform her research. Whether it is with military officers, legal advisors, academic experts, technologists, industry representatives or diplomats, the author sought to leverage her own

network built as part of her doctoral research and as part of her professional affiliations. These informal consultations would generally be of conversational nature to delve deeper into certain issues of relevance to this thesis. This approach has not only increased the author's understanding of the issue from various perspectives (e.g., operational versus legal versus policy considerations), ascertaining yet again the inherently multidisciplinary nature of this issue. It has also helped consolidate her literature review and desk research, oftentimes putting into context some of the policy documents and outputs that are available in the public domain, yet there may not be much on their equivalent of their *travaux préparatoires*. A better understanding of the wider context in which this research may fall into from various perspectives truly helped in ensuring the thesis' findings are somewhat relevant and resonate with the interests, concerns and questions stemming from all stakeholders in the space of military AI.

Third, and finally, plenty of work has been undertaken based on the author's external engagements. Whether it is participation in global summits, e.g., the 2023 and 2024 editions of the REAIM Summit, through her professional work both at Chatham House or the United Nations, in addition to partaking in roundtables, speaking engagements and briefings, these external engagements added further layers to the author's understanding of the 'bigger picture' and her reflections as to how and where her legal research could be useful and why. Active participation in these external engagements truly helped to consolidate some of the findings drawn from the desk research and informal consultations, in addition to challenging the author's own assumptions and knowledge, helping to further consolidate the critical dimension of this work.

Ultimately, it is hoped that these three elements helped the author in zooming out and reach beyond descriptive legal analyses by providing concrete approaches and measures towards integrating IHL considerations from the early stages of a technology's lifecycle.

4.2. Argument, Structure and Guiding Questions

Ultimately, this thesis seeks to identify the main IHL considerations for the development of AI for military targeting, with the assumption that these technologies will only grow in prominence in the battlefield. As debates intensify on the legality of their development, deployment and subsequent use, the framing of this thesis is indeed centered around the observation that armed forces around the world, big and small, are seeking to integrate these technologies for all sorts of applications around targeting, from intelligence, surveillance and reconnaissance to supporting decision-making and the eventual use of lethal force. Current conflicts and geopolitical tensions are such that, while scrutiny over the use of these technologies intensifies particularly from civil society organizations, paradoxically, a number of States are seeking to develop and deploy them faster and at scale. In parallel, the private sector is ramping up on its investment in research and development for the rapid development and sales of these technologies in the light of their financial profitability, both traditional defence industries and technological companies alike. Amidst this storm of hype, appetite and investment, and in the light of its high stakes, concerns are – understandably – growing around unchecked innovation at the cost of civilian lives as battlefields become testing grounds. This thesis seeks to demonstrate that technological innovation and legal compliance are not mutually exclusive. In fact, by factoring in IHL considerations from the outset and ensuring that they inform and shape design choices and processes surrounding their development, by promoting a 'compliance by design' approach, it is hoped that this thesis will constitute more than an addition to legal scholarship: It will ultimately pave the way towards responsible innovation that will foster compliance and reinforce IHL implementation.

To this end, this thesis is structured around four substantive chapters.

The first chapter will be dedicated to unpacking what 'artificial intelligence' concretely means and more specifically in the military domain. This will be an opportunity to explore the building

blocks of AI and some of its key dimensions, including what algorithms are, paradigms of machine learning, as well as deep learning. Many policy, legal and general governance discussions surrounding AI use these technical terms, at times employed as a basis for proposed solutions. A clear understanding of what these terms even mean in the first place will be critical. It is hoped that the first chapter will provide the reader with a solid foundational understanding of the technology, subsequently enabling in-depth discussions on the relevant legal implications in the following chapters.

The second chapter will look at data. Data plays a critical role in shaping the performance and outputs of AI-enabled technologies: The quality and quantity of data used for the training of these capabilities are indeed highly influential vis-à-vis its overall reliability and, ultimately, compliance with applicable laws. As such, Chapter 2 will seek to answer the following questions:

- What is data and what is its relevance to the development of AI for military targeting?
- How can IHL inform key processes in ensuring the quality of training and testing datasets, and how can these processes be leveraged to foster legal compliance?
- To what extent does the quantity of training data influence the design of IHL-compliant AI systems for military targeting?

To this end, this chapter will unpack some of the main dimensions pertaining to data and the key, relevant legal considerations for each of these aspects. These include, for instance, the veracity of training datasets, their diversity, as well as the use of proxy data. It is hoped that the relevant legal considerations will influence, frame and shape practices surrounding the collection, curation and use of datasets underpinning the training and overall performance of AI-enabled technologies for military targeting.

The third chapter will delve into critical dimensions, beyond training datasets, pertaining to the design and development of AI-enabled programmes for military targeting. Specific attention will be provided to the following guiding questions:

- Is the development of machine learning for anti-personnel targeting permissible under IHL? What are some of the technical issues and considerations that could affect the legality and permissibility of developing such capabilities?
- Who are the 'developers' and what is their role in the design, training, testing and evaluation of these technologies? How does State responsibility relate to developers and how can due diligence be used as a means for fostering compliance with IHL?
- Are Article 36 legal reviews fit for the purpose of evaluating machine learning-based systems? Why is there a need for an iterative review process?

The fourth and final chapter looks into the question of accuracies and uncertainty. AI is, by nature, a probabilistic technology with inherent inaccuracies and uncertainties. This chapter aims to explore what are some of the main issues stemming out of the lack of certainty surrounding these technologies, notably the opacity induced by their black box, or their opacity. To this end, the following guiding questions will frame this chapter's discussions:

- How does international humanitarian law approach the issue of uncertainty in war? Does IHL have any expectations for accuracy in military targeting and, conversely, is there room for error and if so, to what extent?
- What are the legal implications of AI technologies' inherently probabilistic nature?
- Does the black box nature of machine learning raise compliance issues with regards to the legal duty to investigate?

This thesis will conclude with an epilogue, summarising some of the main findings from what has been an intense, five-year research journey. The conclusion will also be an invitation, to

the reader, to reflect on the many adjacent legal issues surrounding the design, development, training, testing, deployment and use of AI-enabled technologies for military targeting, and where this thesis sits amidst the on-going frenzy in this space.

Chapter I – Military Applications of Artificial Intelligence

1 Introduction

This inaugural chapter seeks to unpack the definition and many dimensions of 'artificial intelligence', as well as to examine their (potential) applications in the military sphere with a particular focus on targeting processes. These foundational discussions will enable us to examine, in more detail, specific dimensions relating to the development of military AI in subsequent chapters (i.e., data; the design and development processes; and uncertainties). Our discussions will also bring to the fore some notable examples of military AI development and their eventual deployment in the battlefield. It is hoped that this chapter will add nuance to the general discourse surrounding 'killer robots' and debunk some of the commonly-held, misleading assumptions on what AI-enabled technologies for military targeting consist of.

Academic research in law and policy is sailing through waves of unprecedented excitement on AI and its implications. Yet, narratives are often shrouded with hype, thus requiring extra care in ensuring the technology is well understood. Overhyping AI can in fact bear a number of negative consequences, from compromising legal compliance to even public safety.[85] As such, the present chapter seeks to break away from this common trend and provide readers with a strong foundational basis on the technology which, in turn, will enable granular discussions in the following chapters.

Beyond providing basic definitions, this chapter will seek to provide the lay reader with a general overview of these technologies' characteristics and capabilities. When reports of such examples are available in the public domain, the author will provide the reader with concrete use-cases of these technologies to illustrate some of the points being made. This in turn will

[85] Kevin LaGrandeur 'The consequences of AI hype' (2023) 4 AI Ethics 653 <https://link.springer.com/article/10.1007/s43681-023-00352-y> accessed 13 March 2024; see also Yasmin Afina 'Rage Against the Algorithm: the Risks of Overestimating Military Artificial Intelligence' (*Chatham House*, 27 August 2020) <https://www.chathamhouse.org/2020/08/rage-against-algorithm-risks-overestimating-military-artificial-intelligence> accessed 13 March 2024

allow for evidence-based reflections, ultimately ensuring that law and policy analyses stemming from the present study and beyond are indeed relevant, robust, and survive the test of time as technological progress is proceeding apace.

The first section of this chapter will explore the meaning of the term 'artificial intelligence', and how it is to be understood and framed for the purpose of this thesis. It is important that this section sets clear boundaries on what is meant by 'AI', and sheds light on terminology commonly associated with AI-related studies, such as 'machine learning', 'deep learning', and 'neural network'. Clarity on the terminology and their scope is needed to ensure that our subsequent discussions will be made based on a shared understanding of what could constitute an 'artificial intelligence'. The second section will then look into military targeting practices and provide the reader with a taxonomy on where, and how, artificial intelligence can be integrated into military targeting, if not already. This second half of the chapter will also provide a general overview of the issues that may arise from the development and subsequent deployment of AI technologies in military targeting, paving the way for the detailed discussions in the following chapters.

2 Artificial Intelligence

2.1 What is AI? What makes a technology an 'AI'?

The term 'artificial intelligence' is everywhere: In the span of a few years, it has gone from what most would consider as science fiction to, now, a common household name. As discussed in the introduction, there is, however, no universally agreed definition: Paradoxically, everyone seems to have their own idea on what 'artificial intelligence' is, yet no one agrees on what such a mainstream term even means in the first place. As such, there have been numerous attempts at defining this construct. These attempts can be categorised, broadly, into two approaches:

1. A human intelligence-centric approach:

Many of the definitions presented on artificial intelligence uses human intelligence as the sole point of reference. Some of these definitions would approach artificial intelligence as being capable of 'duplicating human faculties', including creativity, self-learning and improvement, and the learning and use of language; in other words, putting artificial intelligence on par with human intelligence. [86] Some would even go as stating human intelligence as the minimum threshold, with the view that artificial intelligence surpasses the former. This is usually the approach taken to define artificial general intelligence, which will be discussed in further detail shortly.

2. A technology-centric approach:

These approaches to defining artificial intelligence would actively steer away from comparing these systems' functions and performance to human intelligence. This is for example the approach taken in the recently adopted European Union (EU) AI Act, which defines an AI system as a 'machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.'[87]

These definitions are often in response to concerns against the tendency to anthropomorphise AI and the adjacent issues that stem from such practice. These include, among others, risks of misplaced trust and over-confidence; the need to uphold responsibility and accountability; privacy concerns; and even what some would argue as AI's inherent association with the white racial frame (i.e., AI being cast predominantly as white).[88] At the UN, states discussing the

[86] Russell and Norvig (eds) (n 20) 18
[87] EU AI Act (n 23) article 3(1),
[88] Respectively Christoph Bartneck and others 'Psychological Aspects of AI' in Christoph Bartneck and others (eds) *An Introduction to Ethics in Robotics and AI* (Springer, 2021) <https://link.springer.com/chapter/10.1007/978-3-030-51110-4_7> accessed 17 March 2023; Christine Rosen 'Technosolutionism Isn't the Fix' (2020) 22(3) The Hedgehog Review <https://go.gale.com/ps/i.do?id=GALE%7CA645242019&sid=googleScholar&v=2.1&it=r&linkaccess=abs&iss

issue of LAWS within the framework of the CCW GGE even went as far as ascertaining, in one of its Guiding Principles adopted in 2019, that 'in crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons should not be anthropomorphized'.[89]

A similar approach is taken by the Council of Europe's recently adopted Framework Convention on Artificial Intelligence, which contains the following definition of an 'artificial intelligence system': 'a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments. Different artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment.'[90] This definition is primarily centred around the system's technical features and characteristics.

There is, again, no universal consensus as to what constitutes the best approach to defining AI. Beyond these two general categories, there is a multitude of layers in the definition of these technologies, depending on a number of factors including, but not limited to, the discipline in question, the needs and objectives (i.e., what is this definition for), as well as the entities and stakeholders that will be impacted by the definition in question. There is a rich and growing literature in this space, a deep examination of which would fall beyond the scope of the present thesis. However, for the purpose of subsequent discussions, it would be worth differentiating

---

n=15279677&p=AONE&sw=w&userGroupName=anon~fdd2b5a8> accessed 17 March 2023; Mark Ryan 'In AI We Trust: Ethics, Artificial Intelligence, and Reliability' (2020) 26 Science and Engineering Ethics <https://link.springer.com/article/10.1007/s11948-020-00228-y> accessed 17 March 2023; Bianca Kronemann and others, 'How AI encourages consumers to share their secrets? The role of anthropomorphism, personalisation, and privacy concerns and avenues for future research' (2023) 27(1) Spanish Journal of Marketing – ESIC <https://www.emerald.com/insight/content/doi/10.1108/SJME-10-2022-0213/full/html> accessed 17 March 2023; and Stephen Cave and Kanta Dihal 'Race and AI: the Diversity Dilemma' (2021) 34 Philosophy & Technology <https://link.springer.com/article/10.1007/s13347-021-00486-z> accessed 17 March 2023

[89] Meeting of the High Contracting Parties to the CCW 'Final Report' (n 51) 10 para (i). More on the UN discussions surrounding lethal autonomous weapons systems will be covered in detail later throughout the thesis.

[90] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Framework AI Convention) art 2

'artificial general intelligence' from 'applied AI' or 'narrow AI', a dichotomy that transcends the three categories covered earlier. This differentiation is essential as part of the objective to keep the present thesis grounded, evidence-based, and relevant to existing applications of these technologies in the military targeting space.

2.1.1 Artificial General Intelligence

As alluded in the previous section, one way of approaching artificial intelligence is to see those systems as surpassing that of humans' intelligence and cognitive capabilities. This 'superhuman' category of AI is often referred to as 'artificial general intelligence', although other colloquial terms are also used, such as 'true AI' or 'singularity'.[91] These terms stem from the idea that artificial general intelligence can, essentially, emulate the many facets and layer of complexity inherent to human reasoning within one single computer programme, and apply those to the domain of rational knowledge.[92] In other words, artificial general intelligence would not be constrained by its initial programming and the objectives set by the developers. Rather, artificial general intelligence will have the ability to reason and exercise judgment against highly complex settings, self-initiate new learning, develop and expand its capabilities and functionalities, and define its own priorities. Historically, this approach reflects the underlying desires and ambitions that led to the establishment of AI as a discipline. In 1950, Alan Turing developed the imitation game, or what is today colloquially known as the 'Turing test'. The test essentially aims to provide metrics to answer the question 'Can machines think?' and has since been developed and used to assess and evaluate the performance of AI systems.[93]

---

[91] See, respectively, Alan Bundy 'Preparing for the future of Artificial Intelligence' (2017) 32 AI & Society <https://dl.acm.org/doi/abs/10.1007/s00146-016-0685-0> accessed 21 November 2024; Michael L. Anderson 'Why is AI so scary?' (2005) 169(2) Artificial Intelligence <https://www.sciencedirect.com/science/article/pii/S0004370205001529> accessed 21 November 2024; and Vernor Vinge 'Signs of the singularity' (2008) 45(6) IEEE Spectrum <https://ieeexplore.ieee.org/document/4531467> accessed 21 November 2024

[92] Alain Cardon *Beyond Artificial Intelligence: From Human Consciousness to Artificial Consciousness* (ISTE Ltd. 2018) ix

[93] Ayse Pinar Saygin, Ilyas Cicekli, Varol Akman 'Turing Test: 50 Years Later' (2000) 10 Minds and Machines <https://link.springer.com/article/10.1023/A:1011288000451> accessed 20 March 2024

The original proposal laid for the two-month Dartmouth workshop in 1956 stated the following objective: 'an attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.'[94] Participants to said convening were of the belief that intelligence could be described so precisely that it could be codified into machines.[95] Such systems will then be capable of solving problems that humans only have been able to solve so far, and learn how to self-optimise by themselves against the set objectives, learn and decide on its own the training data it requires to develop novel functionalities and capabilities, all without further assistance by humans.

Following the Dartmouth workshop, the optimism that artificial general intelligence could be achieved was a driving force in the first decade of AI research. Yet, due to a number of factors, including hardware limitation and the unavailability of data back then, promises were far from being met. Theories and experiments in this space were deemed as 'weak methods', and the systems developed could not be scaled up to solve 'larger or difficult problem instances.'[96] In addition, there was the growing realisation that deciphering human intelligence was not as straightforward: the level of knowledge and research undertaken in the field of neuroscience were enough to come to that realisation, yet remained limited to enable any attempt at 'coding' it into algorithms. [97] As time went by, the focus shifted from developing artificial general intelligence to 'narrow' AI, designed to solve very specific tasks and which defined most of the research undertaken in the field over the past few decades.

[94] As cited in Russell and Norvig (eds) (n 20) 17
[95] Jørgen Veisdal 'The Birthplace of AI' (*Medium*, 12 September 2019) <https://medium.com/cantors-paradise/the-birthplace-of-ai-9ab7d4e5fb00> accessed 24 May 2020
[96] Russell and Norvig (eds) (n 20) 22
[97] Ben Goertzel 'Human-level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's *The Singularity Is Near*, and McDermott's critique of Kurzweil' (2007) 171(18) Artificial Intelligence <https://www.sciencedirect.com/science/article/pii/S0004370207001464> accessed 24 May 2020

In the past few years, as progress is proceeding apace in the AI field, there has been a re-emergence of interest in developing artificial general intelligence. Google DeepMind was notorious for putting 'solving intelligence' and achieving artificial general intelligence at the heart of its mission since its establishment as DeepMind in the early 2010s.[98] Their relatively early focus on artificial general intelligence was met with plenty of scepticism, deeming it as far-fetched, as a product of science fiction, with the overall assumption that there is a 'huge gulf' between the technology's state of the art and the idea of a generalist, multi-purpose system qualifying as artificial general intelligence.[99] The narrative has, however, dramatically changed upon OpenAI's public release of ChatGPT in November 2022: the latter has catalysed an unprecedented AI frenzy, which in turn brought artificial general intelligence back to the table as a concrete possibility and not a product of science fiction anymore. There is a growing number of researchers of the view that while current AI applications are not artificial general intelligence yet, they exhibit signs, or sparks, of such technologies.[100] The dominating discourse has shifted from 'whether' to 'when' artificial general intelligence can be achieved, with predictions for the near future mushrooming in the mainstream media every other week.[101] Beyond the press, it is clear that well-resourced industries, including Google and OpenAI, are

[98] Google DeepMind was established as DeepMind in 2010, with its initial mission being to simply 'solve intelligence'. However, since its fusion with the Brain Team from Google Research in April 2023, this has featured less prominently on their public-facing material. On the history of Google DeepMind and its early vision, see: David Rowan 'DeepMind: inside Google's super-brain' (*WIRED*, 22 June 2015) <https://www.wired.com/story/deepmind/> accessed 20 March 2024; on its current status, see: 'Build AI responsibly to benefit humanity' (*Google DeepMind*) <https://deepmind.google/about/> accessed 21 November 2024

[99] Alan Winfield 'Artificial intelligence will not turn into a Frankenstein's monster' (*The Guardian*, 10 August 2014) <https://www.theguardian.com/technology/2014/aug/10/artificial-intelligence-will-not-become-a-frankensteins-monster-ian-winfield> accessed 20 March 2024

[100] See for example Sébastien Bubeck and others 'Sparks of Artificial General Intelligence: Early experiments with GPT-4' (2023) ArXiv, <https://arxiv.org/abs/2303.12712> accessed 20 March 2024

[101] See for example, Eric J. Savitz 'Nvidia CEO Sees Artificial General Intelligence Within 5 Years – With a Big Caveat' (*Barron's*, 20 March 2024), <https://www.barrons.com/livecoverage/nvidia-gtc-ai-conference/card/nvidia-ceo-sees-artificial-general-intelligence-within-5-years-with-a-big-caveat-e2NjhpHnNlqI1qv3QvQw> accessed 20 March 2024, see also Ben Goertzel 'Artificial Superintelligence Could Arrive by 2027, Scientist Predicts' (*Futurism*, 7 March 2024) <https://futurism.com/artificial-superintelligence-agi-2027-goertzel> accessed 29 March 2024

investing heavily in the development of artificial general intelligence: OpenAI has notably published a note specifically dedicated to 'Planning for AGI and beyond'.[102]

Whether artificial general intelligence is truly achievable, or whether it remains a distant dream, falls beyond the scope of this thesis. While artificial general intelligence is not there yet, it is, however, still important at this stage to acknowledge the changing discourse on AGI and differentiate these discussions from 'narrow' AI. The present thesis will indeed focus exclusively on the latter, as most, if not all evidenced research and development in the context of military AI remains in this space at present.

2.1.2 Narrow AI

Artificial general intelligence is meant to be, as its name suggests, generalist. On the contrary, the performance of 'narrow AI', or alternatively 'applied AI', is limited to the intended purpose and the specific, narrow tasks the system has been designed, developed and tested for. As discussed earlier, the explosion in narrow AI is a response to the limitations faced by efforts at developing artificial general intelligence in the 20th century. [103] As such, the focus has been to develop machines that are essentially smarter and faster than humans, with however the caveat that these systems excel for the narrow tasks they were designed for only. [104]

Continued research and progress with this approach has paved the way for the development of advanced systems that have considerably helped users in completing tasks not only faster, but also better. The programmes supporting today's available tools, for example in language translation, navigation assistance, or even to play games such as Google DeepMind's AlphaGo, are all narrow AI.[105] Navigation assistance systems are able to recommend users with suggested

[102] Sam Altman 'Planning for AGI and beyond' (*OpenAI*, 24 February 2023) <https://openai.com/blog/planning-for-agi-and-beyond> acccessed 23 March 2023
[103] Russell and Norvig (eds) (n 20) 22
[104] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Company, 2018)
[105] How the programme will manage to complete this task is another question that will be covered later, in the section dedicated to the different AI learning techniques (Chapter 1, Section 2.2)

routes that not only leads them to their destination, but also taking into account a number of factors, including the user's preferences (e.g., avoid the motorway), traffic (e.g. live information on road accidents and the immediate suggestion of an alternative route), and information surrounding public transportation (e.g., on-going strikes, accidents, delays, etc.) However, these systems remain, again, narrow in their performance and cannot perform beyond their intended purpose. Systems designed and developed for navigation assistance would be obsolete for language translation.

As mentioned in the previous section, this thesis will exclusively focus on narrow AI solutions developed to support military targeting operations. This choice is motivated by the desire to keep the present discussions relevant to on-going research and development in this particular space, all falling into the category of narrow AI. While it would indeed be an interesting scholarly exercise to reflect on the long-term implications of general AI in the security and defence sectors, it is not what this thesis is trying to achieve.[106]

2.2 Machine learning

Today, vast amounts of data are generated daily: put together, this data can provide strategic advantage to those able to access and leverage the insights stemming out of it. In parallel, progress in computing power reach unprecedented levels, with the economic and societal importance of access to compute hardware and power often put on par with access to the internet and the subsea cables.[107] Put together, an avalanche of data and powerful compute pave the way for the development, testing, and subsequent deployment of artificial intelligence.

[106] Research organisms are specifically dedicated to this 'long-term' research focusing more on existential risks pertaining from artificial general intelligence – e.g., the University of Cambridge's Centre for the Study of Existential Risk, the University of Oxford's Future of Humanity Institute, the Centre for Long-Term Resilience, etc.

[107] Girish Sastry and others, 'Computing Power and the Governance of Artificial Intelligence' (2024), ArXiv 12 <https://arxiv.org/pdf/2402.08797.pdf> accessed 21 March 2024

These systems can subsequently complete tasks and provide users with the desired advantage sought from their development, testing and use.

Yet, in order to reach the desired endpoint, AI systems must be underpinned by a series of processes and calculations that eventually lead towards the output. This is where machine learning steps in. Machine learning goes beyond the 'code', instructing the system on what to do with what parameters and what data. More than simply executing tasks, machine learning corresponds to computational methods capable of learning through experience to improve its performance and produce better results.[108] This entails designing and developing the right features for the right models, so it achieves the right tasks to obtain the best of desired results.[109] One of these features is, thus, the ability for the system to learn continuously and adapt its methods when facing new circumstances.[110] How it achieves this optimisation and high levels of accuracy varies from one learning technique to another.

2.2.1 Algorithms

Algorithms lay at the foundation of machine learning and, more generally, computing. Broadly speaking, an algorithm may be defined as a 'computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output.'[111] In other words, an algorithm is akin to a recipe: it is a series of steps and processes that the system needs to follow in order to go from the input to achieve the desired output. What tends to be referred to as the 'code' in many discussions corresponds, essentially, to the algorithm. The sequence of code that constitute the algorithm will indeed 'tell' the system what the desired output is, and instructions as to how the system should achieve it.

[108] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, *Foundations of Machine Learning* (2nd edn, The MIT Press 2018) 1
[109] Peter Flach, *Machine Learning: The Art and Science of Algorithms* (1st edn, Cambridge University Press 2012) 13
[110] Russell and Norvig (eds) (n 20) 2.
[111] Thomas H. Cormen and others, *Introduction to algorithms* (3rd edn, The MIT Press 2009) 5

As alluded to earlier, algorithms in machine learning stretch beyond rigid instructions which the system would be bound to. The system is capable, on its own and in varying ways, of 'learning' on its own to produce better results. In other words, machine learning algorithms are supposed to have the ability to learn what to do of the data it is being fed with and overcome the limitations that come with hand-coded knowledge engineering.[112] This focus on the ability to learn essentially stems from the desire to address issues pertaining to 'code' knowledge by producing new ones. As such, there are generally three major categories of algorithms, namely: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning.

1. Supervised learning

Supervised learning corresponds to a learning method where a human supervisor will train the algorithm by coupling training datasets as point of reference to the desired output. Based on the association made by the human supervisor, the algorithm will use this knowledge to generate a function and process the input raw data based on this function.[113] To this end, the human supervisor will label each training data based on what they want the algorithm to come up with in its outputs; the algorithm will then identify, on its own, key characteristics to differentiate one label from another.

For example, if developers want an algorithm to have the ability to differentiate between pictures of cats and pictures of dog. In a supervised learning setting, the human supervisor will provide the algorithm with a training dataset of cats and dogs pictures, each labelled with either 'cat' or 'dog', depending on what animal each picture shows. To this end, the training dataset will have gone through a series of processes, where humans will have 'cleaned' it from

[112] Russell and Norvig (eds) (n 20) 28
[113] Dhairya Parikh 'Learning Paradigms in Machine Learning' (*Medium*, 7 July 2018) <https://medium.com/datadriveninvestor/learning-paradigms-in-machine-learning-146ebf8b5943> accessed 24 May 2020

irrelevant and inaccurate data (e.g., pictures where there is neither a cat nor a dog), aggregated, and labelled the data as either 'cat' or 'dog'. This process of ensuring the 'cleanliness' of data is often colloquially referred to as 'data hygiene' or 'data cleansing'.[114] As such, the algorithm will 'learn' to identify cats and dogs by establishing a series of correlations and identifying patterns as to what the unique characteristics are of a cat on the one hand, and of a dog on the other hand. What these patterns and correlations are is often subject to much debate, with the inability in many instances to scrutinise what these are: this corresponds to the black box issue, which will be discussed in further detail in a few sections and as part of chapter 4.[115] These patterns can be straightforward, or less so in many instances.[116] The system will then undergo a number of iterations to train itself and either form new patterns, of refine them, in order to improve and optimise its outputs. Once the system has gone through extensive rounds of training, and its performance levels have reached the desired accuracy rate, the algorithm can then be tested against a whole new set of pictures of cats and dogs. The latter will be unlabelled and will aim to assess the performance of the system when it is exposed to new data: this would be the last test before being validated as apt for deployment and use. If its performance remains unsatisfactory with this testing dataset, the system will usually go back to the training stage and conduct further iterations of training, while the developer tries to tweak the algorithm and its parameters to help optimise further its performance.

Supervised learning algorithms are generally useful when its intended purpose is to identify something known and that has been pre-defined, yet humans may either find it difficult to conduct such task, or it may be too time-consuming, or humans may simply not be as good at

[114] Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon 'A Review on Data Cleansing Methods for Big Data' (2019) 161 Procedia Computer Science, <https://www.sciencedirect.com/science/article/pii/S1877050919318885> accessed 22 March 2024

[115] Section 1.3, Chapter 4

[116] See for example Brandon Carter and others, 'Overinterpretation reveals image classification model pathologies' (Conference on Neural Information Processing Systems, NeurIPS 2021) <https://proceedings.neurips.cc/paper/2021/file/8217bb4e7fa0541e0f5e04fea764ab91-Paper.pdf> accessed 21 February 2023

doing it. Beyond sorting cats and dogs, there are a number of ways in which supervised learning methods can be integrated into our daily lives. This technique is indeed being studied, for example, to assist with waste-sorting by identifying objects as per recycling requirements.[117] In the medical field, supervised learning algorithms are being developed to help with cancer detection.[118]

With regards to military applications, supervised learning algorithms can for example be of use to detect targets from satellite imagery. Perhaps the most straightforward application would be target recognition and classification: supervised learning algorithms can be trained to recognise and classify targets of various types, based on data collected from satellite imagery, surveillance drones and radars.[119] Supervised learning can also be used to support mission planning and decision-making. Drawing on historical data, these systems will recommend the user with courses of action based on what it has been taught as success factors.[120] More discussions on military applications will be covered in the second half of this chapter.

2. Unsupervised learning

While supervised learning methods hold many promises, it also faces a number of limitations. One of them is the algorithm being constrained to the labels and parameters set in the training stage. Unsupervised learning thus corresponds to the method where algorithms are provided with training data, without however any human supervisor pre-defining the desired sets of

[117] Shoufeng Jin and others 'Garbage detection and classification using a new deep learning-based machine vision system as a tool for sustainable waste recycling' (2023), 162 Waste Management, <https://www.sciencedirect.com/science/article/abs/pii/S0956053X23001915> accessed 22 March 2024

[118] S Sivakumar and others, 'An empirical study of supervised learning methods for breast cancer diseases' (2018) 175 Optik, <https://www.sciencedirect.com/science/article/abs/pii/S0030402618312622> accessed 22 March 2024

[119] Kaeye Dästner and others 'Classification of Military Aircraft in Real-time Radar Systems based on Supervised Machine Learning with Labelled ADS-B Data' (IEEE Conference, Sensor Data Fusion: Trends, Solutions, Applications, Bonn, 2018) <https://ieeexplore.ieee.org/abstract/document/8547077> accessed 22 March 2024

[120] Kyle P. Hanratty 'Artificial (military) intelligence: enabling decision dominance through machine learning' (SPIE Defense + Commercial Sensing, Orlando, 2023) <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12538/125380J/Artificial-military-intelligence-enabling-decision-dominance-through-machine-learning/10.1117/12.2663413.short#_=_> accessed 22 March 2024

output, nor labelling. The algorithm will essentially observe the data it has been fed with, and identify on its own patterns, structures and clusters. In other words, unsupervised learning has more of an exploratory nature. It can particularly be useful for discovering novel groupings in the data (i.e., clustering) or at least, the likelihood that the data points will belong to a particular distribution (i.e., probabilistic clustering).[121] Unsupervised learning can also be helpful in discovering 'rules that describe large portions' of the data, or put simply, find patterns within patterns (i.e., association).[122]

Unsupervised learning has been used for many of the AI applications we encounter on a daily basis. For example, many recommender systems are notoriously trained in unsupervised learning settings. These systems, as their name suggests, aim to provide the users with recommendations that are likely to match their interests: recommender systems are used for search engines online, music and movie platforms, service provision, as well as in advertising.[123] The algorithm will analyse data collected from users of a specific platform (e.g., music streaming) and will identify patterns and clusters of users based on their preferences, habits and search queries. Based on these groupings, the algorithm can then provide users with personalised recommendations that, in reality, are based on the preferences of fellow users belonging to similar clusters. These groupings can be based on straightforward criteria, such as preferences for a specific music genre (e.g., users listening to Gesaffelstein may enjoy listening to Marcus L, both artists in electronic music). These groupings can also be less straightforward, such as for example frequency patterns in the songs the users generally prefer to listen to. Again, the system's calculations and subsequent clustering leading to its outputs

[121] 'What is unsupervised learning?' (IBM), <https://www.ibm.com/topics/unsupervised-learning> accessed 22 March 2024

[122] Dhairya Parikh 'Learning Paradigms in Machine Learning' (*Medium*, 7 July 2018) <https://medium.datadriveninvestor.com/learning-paradigms-in-machine-learning-146ebf8b5943> accessed 21 November 2024

[123] Ivens Portugal, Paulo Alencar, Donald Cowan 'The use of machine learning algorithms in recommender systems: A systematic review' (2018) 97 Expert Systems with Applications, <https://www.sciencedirect.com/science/article/abs/pii/S0957417417308333> accessed 22 March 2024

can be very opaque, i.e., a black box. The avid listener to Gesaffelstein may, surprisingly, enjoy the system's recommendation to listen to Ibrahim Maalouf, a jazz trumpeter. However, it would be unclear as to how it made the association between the two artists.

In a military targeting setting, unsupervised learning can be used for a number of applications. One of such examples is for the detection of anomalies. This can be in the context of cyber defence, specifically with regards to network intrusion detection systems. The unsupervised learning-based system can detect any form of disruption to its pre-existing clusters and detect suspicious activity at very high speed, even forms of attack that were previously unknown.[124] Another example of such use would be for intelligence collection: drawing from large, unstructured volumes of data (e.g., social media feeds, intercepted communications), an unsupervised algorithm can help analyse these in bulk. By identifying certain underlying themes and topics, the system can feed into risks assessments and the identification of emerging threats.[125] These are a preview of how artificial intelligence can be used for military applications, which again will be discussed in further detail later in the present chapter.

3. Reinforcement learning

While supervised and unsupervised learning can be particularly useful for classification (i.e., tasks involving grouping and clustering), another approach further adds to the kaleidoscope of machine learning techniques: reinforcement learning. Reinforcement learning essentially corresponds to an approach that seeks to mimic the learning approach of the human brain on

[124] Guo Pu and others 'A hybrid unsupervised clustering-based anomaly detection method' (2021) 26(2) Tsinghua Science and Technology, <https://ieeexplore.ieee.org/abstract/document/9147152> accessed 22 March 2024
[125] Jamie Bartlett, Louis Reynolds 'The state of the art 2015: A literature review of social media intelligence capabilities for counter-terrorism' (2015) Demos <https://demos.co.uk/wp-content/uploads/2015/09/State_of_the_Arts_2015-1.pdf> accessed 22 March 2024

the basis of rewards. Within that context, the programme will optimise its performance and determine the ideal behaviour in order to reach said reward.[126] As opposed to supervised and unsupervised learning techniques, reinforcement learning might not be provided with training data. Instead, it will train itself through the repetition of processes needed to obtain the reward. Through trial and error, the system will use its own data generated from these iterations to train itself and find the best and most optimal behaviour leading to the reward.

Perhaps one of the most prominent examples of reinforcement learning is Google DeepMind's AlphaGo. AlphaGo is a system the company developed to learn and master the game of Go, an ancient Chinese game known to be profoundly complex. Players can opt for a vast number of possible moves on the game's board, known to be greater than that in chess. While the rules of the game, in themselves, are simple, the game of Go is also known for its strategic depth, where players must be able to plan on the short- and long-term, in addition to crafting the right balance between offensive and defensive moves. Through games of self-play, AlphaGo eventually developed its own strategies and optimised the combination and sequences of moves based on one defined reward: win the game of Go. As such, AlphaGo was able to achieve a 99.8% winning rate against other Go programmes and defeated, in October 2015, the then human European Go champion by 5 games to 0: this was the first time ever a computer programme has managed to defeat a human professional player.[127]

In the military context, reinforcement learning has been explored for a number of applications. One example is the use of reinforcement learning to help with the navigation of autonomous systems, such as unmanned aerial vehicles (UAVs), ground robots, or unmanned maritime vehicles (UMVs). The appeal of developing these autonomous systems is the ability to operate in cluttered and complex environments with bad weather and limited access to communication:

[126] Parikh 'Learning Paradigms in Machine Learning' (n 122)
[127] David Silver and others 'Mastering the game of Go with deep neural networks and tree search' (2016) 529 Nature, <https://www.nature.com/articles/nature16961> accessed 22 March 2024

in this context, a human operator may be more of a liability than an asset. As such, through reinforcement learning, these systems can learn to navigate these difficult environments, avoid obstacles, and accomplish the said task (i.e., reward) such as surveillance, or even target identification and engagement if such systems were armed.[128] Another example of reinforcement learning use in the military is for training and planning. More specifically, the algorithm can be programmed into playing the role of an adversary to help members of the armed forces to train and plan, for example in tactical simulation technologies.[129] Similar to AlphaGo, the idea is to use reinforcement learning almost as a synthetic player or character that must be defeated; hence, this exercise can be used as a way of training military personnel and testing their tactics, strategies, and overall decision-making for combat.[130]

2.2.2 Deep learning

Another learning method which, however, does not fall under one of the three abovementioned categories, corresponds to deep learning. Conceptually similar to reinforcement learning, deep learning draws inspiration from the human brain, as it 'allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.'[131] Deep learning is designed to discover 'intricate structure in large data sets […] to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer.'[132] In other words, deep learning takes advantage of increases in the amount of data, and will 'develop intelligence' by learning through the available data and by processing them in layers. This

[128] Nan Yan, Subin Huang, Chao Kong 'Reinforcement Learning-Based Autonomous Navigation and Obstacle Avoidance for USVs under Partially Observable Conditions' (2021) 2021 Mathematical Problems in Engineering 2 <https://www.hindawi.com/journals/mpe/2021/5519033/> accessed 22 March 2024
[129] Arif Furkan Mendi and others, 'Applications of Reinforcement Learning and its Extension to Tactical Simulation Technologies' (2021) 22(1) International Journal of Simulation: Systems, Science & Technology, <https://ijssst.info/Vol-22/No-1/paper14.pdf> accessed 22 March 2024
[130] Volan Ustun and others 'Adaptive Synthetic Characters for Military Trainig' (2021) ArXiv, <https://arxiv.org/abs/2101.02185> accessed 22 March 2024
[131] Yann LeCun, Yoshua Bengio, Geoffrey Hiton 'Deep learning' (2015) 521 Nature <https://www.nature.com/articles/nature14539> accessed 24 May 2020
[132] Ibid

learning process is notably relevant for image categorization (including facial recognition), language translation, and speech recognition.[133]

Deep learning is driven and powered by neural network models, inspired by the electrochemical activity in the human brain with, in a nutshell, nodes or units 'connected by directed links' akin to networks of brain cells (neurons).[134] In other words, it uses the human brain as a source of inspiration, i.e., how brain cells are interconnected within the brain, and how information goes through this network of interconnected brain cells through electric signals. In the case of artificial neural networks, those brain cells are called 'units'.[135] These units can either receive information that the machine will learn about (input units); or they can signal how the machine will need to respond to the information it has learned via the input units (output units); or they can be the units in-between the input and output units (hidden units) the information goes through.[136] There can be layers of these hidden units and they constitute most of the neural network. These units are interconnected by 'weights', which can be adjusted either to positive or negative, depending on whether one unit excites the other. Learning occurs by changing the weights connecting the units as the machine receives feedback in order to make more accurate predictions.[137]

While deep learning and reinforcement learning have the similar goal of mimicking the human brain, both approaches are fundamentally different. While reinforcement learning will be limited to data generated from their own experiences for decision-making, deep learning will be expected to combine their experience along with historical data for tasks that stretch beyond

[133] On facial recognition, see for example: Neelabh Shanker Singh, S. Hariharan, Monika Gupta 'Facial Recognition Using Deep Learning' in Vanita Jain, Gopal Chaudhary, M. Cengiz Taplamacioglu, M. S. Agarwal (eds) *Advances in Data Sciences, Security and Applications* (Springer, Singapore 2020)
[134] Russell and Norvig (eds) (n 20) 727-728
[135] Charu C. Aggarwal, *Neural Networks and Deep Learning: A Textbook* (Springer 2018), VII
[136] David L. Poole, Alan K. Mackworth, *Artificial Intelligence: Foundations of Computational Agents* (1st edn, Cambridge University Press 2010)
[137] Charu C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*, 1-2

decision-making such as classification and pattern recognition.[138] Deep learning algorithms are not constrained by their own experiences and trial and errors, as they can draw from 'external' data, i.e., datasets drawn from external sources. In other words, these models are capable of processing vast amounts of data, and they can be used for a number of tasks including for regression, classification, and generative purposes at scale.

There are two notable limitations or at least, considerations related to deep learning. First, due to their sheer scale, with the large number of trainable parameters and massive amounts of training data, these programmes are particularly difficult or even impossible to understand and explain. This will have implications on the black box issue, which will be discussed in further detail in Chapter 4.[139] Second, the scale of their parameters and data means that their functioning requires a level of sophistication in their hardware and overall infrastructure surrounding these systems. From cutting-edge semiconductors to major data centres, the latter would for example require high levels of energy, thus explaining the high levels of investments of big technology companies such as Microsoft and Google in such areas.[140]

Deep learning is used extensively today and can be seen as the key towards so-called 'foundation models'.[141] Thanks to their ability to process massive volumes of data, deep learning algorithms has consequently enabled the tremendous leaps witnessed over the past couple of years in the field of AI. ChatGPT, Claude and Llama all operate on a deep learning architecture. In the legal space, for example, there is a mushrooming of products

[138] Lucas M. Cappel 'What is the difference between Deep Learning and Reinforcement Learning?' (*Big Data Made Simple*, 2019) <https://bigdata-madesimple.com/what-is-the-difference-between-deep-learning-and-reinforcement-learning/> accessed 24 May 2020

[139] Chapter 4, section 1.3

[140] Google has, for example, been reported having ordered a number of small nuclear reactors to power its datacenters, see: Alex Lawson 'Google to buy nuclear power for AI datacentres in 'world first' deal' (The Guardian, 15 October 2024) <https://www.theguardian.com/technology/2024/oct/15/google-buy-nuclear-power-ai-datacentres-kairos-power> accessed 6 June 2025

[141] Rishi Bommasani and others 'On the Opportunities and Risks of Foundation Models' (2022) ArXiv <https://arxiv.org/abs/2108.07258> accessed 14 November 2024

commercialised for document analysis.[142] Deep learning algorithms are trained on large datasets of existing legal documents, which subsequently enables the programme to recognise patterns and differences in the text. As such, the programme is then capable of comparing documents, and eventually classify key, recurring terms in new legal documents. These models can play a key, enabling role in legal research by analysing vast amounts of historical case law data and other legal documents in a much faster and more efficient way than humans.

In the military context, there are also vast efforts at developing and, eventually, integrating deep learning solutions not only to enhance decision-making, but also to increase efficiency and effectiveness. Perhaps the most prominent example – and one of the oldest ones – of deep learning applications in the military, today, corresponds to supporting automated target recognition (ATR). These programmes can, in fact, 'scan' very quickly through vast amounts of imagery collected from ISR systems, or even identify potential, land-based targets based on vibration signatures in data picked from laser radars.[143]

2.2.3 Modelling

Modelling is a practice commonly used in sciences to 'model' a reality based on a set of parameters, assumptions, data and pre-defined interactions.[144] These will allow modellers to obtain results (predictions) on a specific question that could potentially reflect reality – with the caveat that this reality is confined to the same features set in the model. Traditionally, modelling approaches rely on reproducing reality through the mathematical description of a

[142] Qian Han, Derek Snaiduf 'Comparison of Deep Learning Technologies in Legal Document Classification' (IEEE International Conference on Big Data, Orlando, 2021) <https://ieeexplore.ieee.org/document/9671486> accessed 7 April 2024

[143] See Steven K. Rogers and others 'Neural networks for automatic target recognition' (1995) 8(7-8) Neural Networks <https://www.sciencedirect.com/science/article/abs/pii/089360809500050X> accessed 7 April 2024; see also Zhe Geng and others 'Target Recognition in SAR Images by Deep Learning with Trainign Data Augmentation' (2023) 23(2) Sensors <https://pubmed.ncbi.nlm.nih.gov/36679740/> accessed 7 April 2024; and Serkan Aktas 'Applying Deep Learning Methods to Identify Targets in Synthetic Aperture Radar Images' (Thesis, Naval Postgraduate School 2020) <https://apps.dtic.mil/sti/trecms/pdf/AD1126746.pdf> accessed 7 April 2024

[144] Yasmin Afina, Calum Inverarity 'Predictions and Policymaking : Complex Modelling Beyond COVID-19' (Chatham House, Expert Comment, 2020) <https://www.chathamhouse.org/expert/comment/predictions-and-policymaking-complex-modelling-beyond-covid-19> accessed 24 May 2020

specific phenomenon (e.g., the fall of an object to illustrate Newton's law of gravity). These models are based on human observations, analysis of the data, and their development requires the establishment of a set of approximations and simplifications to establish correlations. With technological progress and the increasing availability and production of data, machine learning increases our capacity for analysis in scale and in complexity. As such, AI-enabled models are fundamentally data-driven: its functioning relies mostly on the data to infer the correlations sought after.[145]

For example, in the relatively early stages of the COVID-19 pandemic, Imperial College London modelled the potential impact of the adoption of different public health policies – thus modelling possible realities in the UK based on different policies the government could adopt. These models have reportedly been used by the UK government and driven their approach from a strategy of mitigation to one of suppression.[146]

Similarly, the military uses predictive modelling extensively today and for a range of purposes. Perhaps the most prominent example corresponds to simulation and training: modelling enables military personnel to undergo training and planning using simulators that replicate real-world scenarios. By predicting the behaviour of actors, events and orders of consequences based on historical data and within certain sets of parameters, these simulations can involve a range of situations, including combat scenarios, vehicle operations and flight training. This approach has been embraced by armed forces for many years, thus attesting to the influential role these AI applications hold in shaping the future of warfare. NATO has, for example, its own Modelling and Simulation Centre of Excellence, located in Italy which hosted, among

[145] Will Koehrsen 'Modeling: Teaching a Machine Learning Algorithm to Deliver Business Value' (*Medium*, 15 November 2018), <https://towardsdatascience.com/modeling-teaching-a-machine-learning-algorithm-to-deliver-business-value-ad0205ca4c86> accessed 24 May 2020

[146] Sabine L. van Esland, Ryan O'Hare 'COVID-19: Imperial researchers model likely impact of public health measures' (*Imperial College London*, 17 March 2020) <https://www.imperial.ac.uk/news/196234/covid19-imperial-researchers-model-likely-impact/> accessed 24 May 2020

others, a conference for military, academia and industry representatives and dedicated to modelling and simulation for autonomous systems in late 2023.[147] Beyond helping with the prediction of outcomes and consequences of action for both training and planning purposes, modelling can even go as far as aiming to predict the adversary's behaviour. In China, commercial systems are being used by the military to feed into a system designed to predict the behaviour and decision-making of human adversaries.[148]

2.3 AI characteristics

AI technologies present a number of characteristics that, without appropriate attention or consideration, will have a number of implications on the technologies' development, testing, deployment and eventual use. This will have subsequent implications on their compliance with the law – as we will discuss in further detail in the following chapters. At this stage, however, it would still be useful to provide an introductory overview on three specific characteristics: 1) autonomy; 2) bias; and 3) the black box.

2.3.1 Autonomous versus Automated

First, it would be useful to discuss the difference between 'autonomy' and 'automated', as the former is often used in relation to systems supported by AI. Although an extensive discussion on this topic would fall beyond the scope of this thesis, it would still be useful for us to cover briefly attempts at differentiating both terms in the light of their prominence in literature and in governance discussions. For example, on-going multilateral discussions with the GGE on LAWS are limited by a mandate focusing on 'emerging technologies in the area of 'lethal autonomous weapons systems'. In addition to the background information on the GGE

[147] 'Modelling & Simulation for Autonomous Systems Conference' (*NATO M&S COE*) <https://www.mscoe.org/event/mesas-2023/about-us/> accessed 7 April 2023

[148] Stephen Chen 'China's military lab AI connects to commercial large language models for the first time to learn more about humans' (*South China Morning Post*, 12 January 2024) <https://www.scmp.com/news/china/science/article/3248050/chinas-military-lab-ai-connects-commercial-large-language-models-first-time-learn-more-about-humans> accessed 7 April 2024

provided in this thesis' introduction, it is useful to note that the Group has reflected, in length, on the difference between 'automated' and 'autonomous'. This discussion comes with the caveat that there are still existing disagreements on whether this exercise is useful at all.[149]

One approach would be to define separate both automation and autonomy. 'Automated' systems follow a set of rigid, pre-defined rules (e.g. if/else). For example, a smoke detector functions based on a very simple rule: if it detects smoke, then it will turn on, else it will remain idle. The objective is simple, and the way it achieves this objective is also simple and predictable. The same rule applies for landmines: if they are subject to a certain pressure resulting from the weight of an external element, then they will activate and trigger an explosion, else they will remain unexploded.

On the other hand, 'autonomous' systems would correspond to machines where the actions and learning process of the technology are not strictly constrained by the pre-defined rules, as long as it meets its pre-set objectives. [150] Autonomous technologies are programmed to complete more complex tasks, and thus will need to function on more than a simple and rigid if/else basis. For example, an agent trained to take part in a role-playing game would require the programme to learn and adapt its actions against a number of factors – including the rules of the game, its adversary's choices and actions in the game, the level of difficulty, the complexity of the game, the interactions and subsequent consequences between agents taking part of the game, etc. The same would apply for AI-powered synthetic environments designed and developed to assist in the training and planning of armed forces.[151] The objectives of programmes underpinning such environments will be to train the user and augment their

[149] The differentiation between 'autonomous' and 'automated' were mostly prominent and contentious during the early stages of discussions on military AI in the international sphere. For instance, the ICRC has shifted over time from this approach of differentiating 'autonomous' from 'automated' to a focus on the role of human operators and human control.
[150] 'Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects' (2014) (n 44)
[151] Raghuveer Rao, Celso de Melo, Hamid Krim 'Synthetic Environments for Artificial Intelligence (AI) and Machine Learning (ML) in Multi-Domain Operations' (2021) ARL-TR-9198, DevCom: Army Research Laboratory <https://apps.dtic.mil/sti/trecms/pdf/AD1135395.pdf> accessed 21 March 2024

combat skills through battlefield-like simulations. This requires the programme to adapt its results to the user's skills, choices and decisions. The more skilled the user is, the more difficult the training environment needs to be, thus requiring the software to adapt the training environment accordingly.

Another approach would be to consider autonomy as part of the automation spectrum, 'in which independent decision-making can be tailored for a specific mission'. [152] In other words, systems with the greatest degree of autonomy will have humans taken out of the loop from decision-making, with AI systems effectively replacing them. Systems with limited degrees of autonomy will be able to take decisions without the intervention of human operators, however the latter will still play a role in the decision-making chain. The degree of autonomy conferred to systems will highly depend on a number of factors, including the utility, the magnitude of consequences, as well as legal considerations. For example, some argue that AI-enabled cyber defence solutions must be able to operate without human operators, due to the sheer speed of advanced cyber offensive capabilities.[153] On the flipside of the coin, integrating AI into nuclear decision-making may carry too high of a risk and as such, human operators must always remain in the loop of such decisions.[154]

This approach is ultimately based on the degree at which humans are involved (or not) in the decision-making chain, which is usually described in three prongs in discussions pertaining to autonomous weapons systems:

---

[152] Congressional Research Service (US) 'U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress' (2018) R45392, Version 3 <https://fas.org/sgp/crs/weapons/R45392.pdf> accessed 7 April 2020

[153] Blessing Guembe and others 'The Emerging Threat of Ai-driven Cyber Attacks: A Review' (2022) 36(1) Applied Artificial Intelligence, <https://www.tandfonline.com/doi/full/10.1080/08839514.2022.2037254> accessed 21 March 2024

[154] Alice Saltini 'To avoid nuclear instability, a moratorium on integrating AI into nuclear decision-making is urgently needed: The NPT PrepCom can serve as a springboard' (*ELN*, 28 July 2023) <https://www.europeanleadershipnetwork.org/commentary/to-avoid-nuclear-instability-a-moratorium-on-integrating-ai-into-nuclear-decision-making-is-urgently-needed-the-npt-prepcom-can-serve-as-a-springboard/> accessed 21 March 2024

1. ‘Human out of the loop’, i.e., the ability to act without any human interface’;
2. ‘Human in the loop’, whereby a system ‘requires a human to complete the chain of decision-making, and
3. ‘Human on the loop’, where the system can undertake certain actions, albeit with a degree of oversight exercised by human operators.[155]

This nomenclature, while widely used in governance discussions surrounding AI in security and defence, has also been the subject of much critique. There have been attempts, including by states, at coining alternative terms to describe the degree of human control over such systems such as ‘human-supervised autonomous systems’[156], ‘human-machine interaction’[157], ‘meaningful human control’[158], as well as ‘appropriate human judgment’[159]. For this thesis, however, we will go beyond this debate on terminology for various reasons. In addition to their often-politicised definitions, the nuances in terminology adds little to discussions surrounding legal compliance considerations for the development of AI-enabled technologies, which this thesis is fundamentally about.

[155] This three-pronged approach to humans in/on/out of the loop was notably made prominent in discussions pertaining to autonomous weapons by Human Rights Watch, as cited in: Michael N. Schmitt & Jeffrey Thurnher ‘“Out of the loop”: autonomous weapon systems and the law of armed conflict’ (2013) 4(2) Harvard National Security Journal <https://harvardnsj.org/wp-content/uploads/sites/13/2013/01/Vol-4-Schmitt-Thurnher.pdf> accessed 7 April 2020

[156] Ibid

[157] 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, ‘Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems’ (2019) CCW/GGE.1/2019/3 <https://undocs.org/en/CCW/GGE.1/2019/3> accessed 24 May 2020

[158] ‘The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward’ (2014) (UNIDIR Resources No. 2 <https://www.unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf> accessed 24 May 2020

[159] Heather M. Roff ‘Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons: Briefing Paper for delegates at the Review Conference of the Convention on Certain Conventional Weapons’ (2016) Article 36 <http://www.article36.org/wp-content/uploads/2016/12/Control-or-Judgment_-Understanding-the-Scope.pdf> accessed 24 May 2020

2.3.2 Bias

Algorithmic bias usually refers to the systematic, unfair and harmful biases stemming from the perpetuation of biases inherited from historical data, which has been used to train, test, and eventually shape the outcome of a programme.[160] Hence, much of the research and literature surrounding algorithmic bias have primarily been focused on the societal impact and harm that stems from the development, deployment and use of biased AI technologies. These include studies both on bias perpetuation, and bias mitigation.[161]

Most, if not all discussions and research on the risks surrounding the development, deployment and use of AI technologies will raise the issue of bias. From concerns in education, health, and criminal justice, algorithmic bias is a cross-cutting and overwhelming issue. In education, it is feared that algorithmic bias will perpetuate disparities and inequities that are race-, ethnicity-, gender-based, or even against certain socioeconomic categories and populations with disability.[162] Algorithmic bias is source of worries with regards to sustaining structural and societal disparities and inequities in the provision of healthcare services, as well as in the longer and harsher sentencing of certain ethnicities over others.[163] As such, from as early as the 1990s, tremendous efforts have been dedicated to better understand bias in computer systems, precisely harmful biases in 'computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others.'[164] Algorithmic bias is

[160] Henriette Cramer and others 'Assessing and addressing algorithmic bias in practice' (2018) 25(6) Interactions <https://dl.acm.org/doi/abs/10.1145/3278156> accessed 7 April 2024

[161] Ingvild Bode 'Falling under the radar: the problem of algorithmic bias and military applications of AI' (*Humanitarian Law & Policy*, 14 March 2024) <https://blogs.icrc.org/law-and-policy/2024/03/14/falling-under-the-radar-the-problem-of-algorithmic-bias-and-military-applications-of-ai/> accessed 9 April 2024

[162] Ryan S. Baker, Aaron Hawn 'Algorithmic Bias in Education' (2022) 32 International Journal of Artificial Intelligence in Education <https://link.springer.com/article/10.1007/s40593-021-00285-9> accessed 7 April 2024

[163] See Trishan Panch, Heather Mattie, Rifat Atun 'Artificial intelligence and algorithmic bias: implications for health systems' (2019) 9(2) Journal of Global Health <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875681/> accessed 7 April 2024 and Aziz Z. Huq 'Racial Equity in Algorithmic Criminal Justice' (2019) 68(6) Duke Law Journal <https://heinonline.org/HOL/LandingPage?handle=hein.journals/duklr68&div=33&id=&page=> accessed 7 April 2024

[164] Batya Friedman and Helen Nissenbaum 'Bias in Computer Systems' (1996)14(3) ACM Transactions on Information Systems <https://dl.acm.org/doi/10.1145/230538.230561> accessed 24 May 2020

also of concern in the context of military applications. Concerns range from biased, arbitrary, and potentially unlawful decision-making, to potentially disproportionate and harmful outcomes against specific groups and demographics.[165] In the line of this issue's prominence, it is important to ascertain what algorithmic bias consists of in the first place, its sources, and subsequent (potential) implications in the context of military AI.

A thorough understanding of algorithmic bias, its sources, and the subsequent implications on the performance of a system will also be important in the context of military applications and, ultimately, for these systems' ability to comply with the law. For example, if a computer vision programme exhibiting racial biases in its outputs were to be deployed to assist with target identification in the battlefield. One factor contributing to the system's outputs showing racial biases may stem from the fact that it was exclusively trained and tested using datasets that themselves contained biases; the system will then replicate and perpetuate these biases. The decision to deploy the system can be source of many issues, including legal compliance. If said biases stand in the way of the system's ability to distinguish between fighters and civilians, then there may be a case for establishing the system's incompatibility with the rule of distinction. While Chapter 2 will discuss in more detail the role of data as a key element for legal compliance, it would be useful to quickly note here that identifying and addressing these biases can not only improve mission success; it would also foster legal compliance. There are a number of ways for the identification and mitigation of biases, including the conduct of red-teaming and other adversarial solutions for bias detection, enhancing the transparency and explainability of algorithms, as well as bias audits and impact assessments.[166] Their results will

[165] Philip Alexander 'Exploring Bias and Accountability in Military Artificial Intelligence' (2022) 7 LSE Law Review <https://heinonline.org/HOL/LandingPage?handle=hein.journals/lselr7&div=21&id=&page=> accessed 7 April 2024, see also Katherine Chandler 'Does Military AI Have Gender? Understanding bias and promoting ethical approaches in military applications of AI' (2021) UNIDIR <https://unidir.org/files/2021-12/UNIDIR_Does_Military_AI_Have_Gender.pdf> accessed 7 April 2024

[166] This is an over-simplified, high-level list of suggestions to mitigate risks pertaining to biased datasets. This is not to discount the growing, rich and nuanced discussions and research seeking to address these issues – for example, see: Zaid Khan, Yun Fu 'One Label, One Billion Faces: Usage and Consistency of Racial Categories in

then help developers ascertain whether the system is indeed ready for deployment, or it needs more training to mitigate the biases, or it is not deployable at all and ought to be taken off the market.

There have been attempts at categorising algorithmic bias, for instance based on whether it comes from training data or based on the processing of data.[167] On biased datasets, in addition to reflecting the biases of those creating and generating the data, these datasets will also reflect the biases of those who collect and process the data – from the way it is being gathered, aggregated, processed and selected for the purpose of training the model. In other words, data collection methods can lead to biased datasets. For example, when the sampling process disproportionately includes or excludes certain groups from the datasets, it can result in what is called 'sampling bias'. If the training dataset of a classification algorithm primarily (or even exclusively) consists of images of lighter-skinned individuals, the algorithm trained on said dataset is likely to perform poorly on darker-skinned individuals due to underrepresentation.[168] One concrete example is an experiment ran by AlgorithmWatch back in 2020 showed that Google's Vision Cloud exhibited racial biases that, if used for law enforcement or in conflict, could result in harm and potentially unlawful actions: the algorithm would indeed label an image of a dark-skinned person, holding a thermometer, as holding a 'gun', while a similar image with a light-skinned person would be labelled as an 'electronic device'.[169] If used in an

Computer Vision' (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021) <https://dl.acm.org/doi/abs/10.1145/3442188.3445920> accessed 19 April 2023

[167] For example, see: David Danks and Alex John London 'Algorithmic Bias in Autonomous Systems' (Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, 2017) <https://www.ijcai.org/Proceedings/2017/0654.pdf> accessed 24 May 2020; and 'Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies' (2018) UNIDIR Resources No. 9 < https://unidir.org/publication/algorithmic-bias-and-weaponization-increasingly-autonomous-technologies> accessed 24 May 2020

[168] Seyma Yucer and others 'Measuring Hidden Bias within Face Recognition via Racial Phenotypes' (IEEE/CVF Winter Conference on Applications of Computer Vision, 2022) <https://journals.uc.edu/index.php/isrs/article/view/8006> accessed 12 April 2024

[169] Nicolas Kayser-Bril 'Google apologizes after its Vision AI produced racist results' (*Algorithm Watch*, 7 April 2020) <https://algorithmwatch.org/en/story/google-vision-racism/> accessed 24 May 2020

armed conflict for target recognition, a programme with similar biases could erroneously consider a civilian as directly participating in hostilities and thus, targetable.

Another type of bias corresponds to contextual bias, i.e., where the data is being collected in such way that other information can influence (either consciously or subconsciously) the input.[170] For example, a law enforcement officer recording an individual's behaviour as aggressive can be tainted by subjective criteria, consciously or subconsciously, and that may perhaps be racially influenced. Closely related to contextual bias is selection bias: the latter occurs when the process used to collect data systematically favours certain characteristics or outcomes over others.[171] This can either be due to the surrounding context (e.g., surveys conducted on an online platform would exclude individuals without internet access), and/or due to subjective criteria.

Additionally, labelling bias can also affect the performance and output of algorithms. This occurs when training data is being collected and prepared for use, and is being annotated and labelled by humans (e.g., labelling pictures of cats as 'cats'). These human annotators can introduce biases and, subsequently, lead to biased datasets. One of the most commonly used examples is sentiment analysis algorithms: subjective interpretation of pictures, voice, text and facial expressions can, in fact, reflect the annotator's personal biases and cultural customs.[172]

Finally, bias can also stem from algorithmic feedback loops: biases present in AI algorithms can perpetuate themselves, or even be reinforced through repetitive use, with biased outputs

[170] Forensic Science Regulator (UK) 'Forensic Science Regulator Guidance: Cognitive Bias Effects Relevant to Forensic Science Examinations' (2020) FSR-G-217, Issue 2 <https://assets.publishing.service.gov.uk/media/5f4fc26ce90e074695f80977/217_FSR-G-217_Cognitive_bias_appendix_Issue_2.pdf> accessed 12 April 2024

[171] Alice C. Yu, John Eng 'One Algorithm May Not Fit All: How Selection Bias Affects Machine Learning Performance' (2020) 40(7) RadioGraphics <https://pubs.rsna.org/doi/full/10.1148/rg.2020200040> accessed 12 April 2024

[172] Yi Ding and others 'Impact of Annotator Demographics on Sentiment Dataset Labeling' (Proceedings of the ACM on Human-Computer Interaction CSCW2 Vol. 6, 2022) <https://dl.acm.org/doi/abs/10.1145/3555632> accessed 12 April 2024

feeding into subsequent iterations.[173] This is particularly the case as algorithms are increasingly trained on data from generative models.[174] For example, an image recognition software that consistently misidentifies certain objects due to biased training data will generate mislabelled data, which in turn can reinforce biases in subsequent iterations and uses of it.

2.3.3 Black box

Another characteristic of AI technologies that merits flagging corresponds to the issue often referred to as the 'black box'. i.e., the inability to understand 'how the algorithm is accomplishing what it is accomplishing'.[175] This issue is particularly relevant for deep learning models, generating functions and calculations so complex that they become too opaque, for humans, to actually comprehend the internal workings of the programme to go from the input to the output. Chapter 4 will have entire sections dedicated to describing the black box issue and some of the key legal considerations and approaches to addressing relevant questions. At this stage, however, it is important to note some of the reasons why the black box issue must be discussed.

First, the black box issue lies at the heart of much research and discussions surrounding algorithmic accountability. If decisions are made based on outputs that are difficult to understand, there are concerns that this would not only challenge the accountability and responsibility of the users over the outcome of their actions; this can also raise issues with regards to attribution and forensic evidence. If, for instance, an algorithm is being used for proportionality analyses prior to an attack, and the commander follows the recommended

[173] Nicolò Pagan and others, 'A classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems' (3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Boston, October 2023) <https://dl.acm.org/doi/10.1145/3617694.3623227> acccessed 12 April 2024

[174] Tianwei Chen and others, 'Would Deep Generative Models Amplify Bias in Future Models?' (IEEE / CVF Computer Vision and Pattern Recognition Conference, Seattle, June 2024) <https://arxiv.org/abs/2404.03242> accessed 12 April 2024

[175] Dallas Card 'The "black box" metaphor in machine learning' (*Towards Data Science*, 5 July 2017), <https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0> accessed 11 January 2020

action blindly without much insight as to how the programme came to its results due to a black box. In the case of humanitarian incidents where civilian casualties are heavy and there the proportionality of the attack is being challenged, the opaque nature of the system can raise issues with regards to its ability to inform the commander for their assessments, in compliance with the law, in the first place.[176]

Second, and beyond legal considerations, the black box issue raises issues with regards to trust, confidence, and risk assessments. Trust, by the armed forces, in AI-enabled technologies forms a major concern with regards to the development, adoption and integration of these technologies into military operations.[177] Without clarity and transparency over a system's calculations, end-users may not have full confidence in the systems with regards to its reliability for legal compliance and alignment with mission success. This can also affect the commander's ability to conduct their risks assessment prior to the conduct of, and during, a military operation.

Third, the black box issue can also raise concerns with regards to the procurement of capabilities, by the armed forces, from defence contractors. The procurement of new technologies will, in fact, trigger a number of tests and evaluation from the purchasing armed forces, including a legal review and attribution parameters.[178] The opacity induced by the black box nature of many, if not most AI-enabled systems will stand in the way or at least, complexify the conduct of these tests.

---

[176] AI is being developed for the purpose of enhancing the commander's situational awareness, which in turn can be used to inform proportionality analyses. See for example: Chang-Eun Lee and others 'Deep AI military staff: cooperative battlefield situation awareness for commander's decision making' (2023) 79 The Journal of Supercomputing <https://link.springer.com/article/10.1007/s11227-022-04882-w> accessed 12 April 2024

[177] Christi S. Montgomery 'Trust in the Machine: AI, Autonomy, and Military Decision Making with Lethal Consquences' (Report, Certificate Program in Ethics and Emerging Military Technology, U.S. Naval War College Newport 2019) <https://apps.dtic.mil/sti/pdfs/AD1121116.pdf> accessed 12 April 2024

[178] Magdalena Pacholska 'Many Hands in the Black Box: Artificial Intelligence and the Responsibility of International Organizations' (2023) in Rossana Deplano, Antal Berkes, Richard Collins (eds) *Reassessing the Articles on the Responsibility of International Organizations: From Theory to Practice* (2023, Edward Elgar) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4501072> accessed 12 April 2024

2.4 Data

With the exponentially growing use of, and dependency on, digital technologies across all segments in society, including within the military sphere, an unprecedented amount of data of all sorts is being generated. The pace at which data is being generated and produced is keeps increasing dramatically: A Forbes study from 2018 found that 2.5 quintillion bytes of data were being created each day.[179] By 2025, the tech company IBM estimates more than 150 zettabytes of data that will require analysis.[180] It therefore comes as no surprise that the ability to harness this vast amount of data would present incredibly attractive opportunities and, beyond commercial applications, all fields and disciplines could benefit from this critical mass of data.

In a nutshell, data refers to the information that is used for machines to learn and subsequently produce outputs. Data can take many forms, including text, images, videos, sound, sensor readings, as well as the combination of these. Against the explosion in the number of data generated every day, solutions are being developed to make sense of it: this is what is commonly named as 'big data'. These data can provide insights that, otherwise, would have either been impossible to access, or it would have taken too long for anyone to process such large amounts of data. For example, programmes are now – infamously – used in a widespread manner to enable targeted advertising online.[181] Despite extensive media coverage – and condemnation – on the use of personal data for such applications of AI in the context of the Cambridge Analytica scandal, AI-enabled targeted advertising technologies remain very much present, although awareness and studies on their legal, ethical and societal implications (and

[179] Bernard Marr 'How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read' (*Forbes*, 21 May 2018) <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=471cf14c60ba> accessed 16 March 2023

[180] 'How to manage complexity and realize the value of big data' (*IBM*, 28 May 2020) <https://www.ibm.com/blog/how-to-manage-complexity-and-realize-the-value-of-big-data/> accessed 16 March 2023

[181] Jin-A Choi, Kiho Lim 'Identifying machine learning techniques for classification of target advertising' (2020) 6(3) ICT Express <https://www.sciencedirect.com/science/article/pii/S2405959520301090> accessed 16 March 2023

concerns) are also growing.[182] In the field of medicine and public health, extensive data gathered on COVID-19, on the virus in itself, but also on public responses worldwide against the pandemic, information exchange on best practices, as well as on the measures adopted and their socio-political and security implications will allow for better preparedness for future pandemics of this scale.[183]

The same applies for the military sphere. The ability to generate information from the collection and processing of data of quality and of this quantity would provide advantage at unprecedented levels.[184] Machine learning techniques provide, in fact, 'the ability to predict likely future outcomes, to calculate risks between competing choices, to make sense of vast amounts of data at speed, and to draw insights from data that would be otherwise invisible to human analysts.'[185] For instance, live satellite imagery, along with data generated from sensors and weather information may indicate commanders when would be the best time to launch an attack, as well as how and where on a specific individual target. At a more sophisticated level, AI solutions are being developed, or even used to identify potential targets (e.g., suspected members of non-state armed groups) even based on proxy indicators, which this thesis will discuss in further detail in Chapter 2.[186] Combining heightened situational awareness, intelligence possession, and mission success parameters, these tools will provide, if they work as intended, cutting-edge advantage at the tactical, strategic, and operational levels.

---

[182] On Cambridge Analytica, see: Joanne Hinds, Emma J. Williams, Adam N. Joinson '"It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal' (2020) 143 International Journal of Human-Computer Studies <https://www.sciencedirect.com/science/article/abs/pii/S1071581920301002> accessed 16 March 2023. On a study related to the societal impact of targeted advertisement, see: Veronica Marotta and others, 'The Welfare Impact of Targeted Advertising Technologies' (2022) 33(1) Information Systems Research <https://pubsonline.informs.org/doi/abs/10.1287/isre.2021.1024> accessed 16 March 2023

[183] Aboul-Ella Hassanien, Nilanjan Dey, Sally Elghamrawy, *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach* (1st edn, Springer International Publishing 2020)

[184] George A. Tsihrintzis, Dionisios N. Sotiropoulos, Lakhmi C. Jain (eds), *Machine Learning Paradigms* (1st edn, Springer International Publishing 2019) 2

[185] Ashley Deeks, Noam Lubell, and Daragh Murray 'Machine Learning, Artificial Intelligence, and the Use of Force by States' (2019) 10(1) Journal of National Security Law & Policy <https://jnslp.com/wp-content/uploads/2019/04/Machine_Learning_Artificial_Intelligence_2.pdf> accessed 24 May 2020

[186] Chapter 2, Section 2.3

In addition, and in principle, this unprecedented amount of data also presents opportunities to foster compliance with IHL rules. While the vast majority of research on the law and policy surrounding military AI tends to focus on concerns and risks (and arguably rightly so), one of the key drivers behind the research and development of these technologies is its promise for a more frictionless war. Tools can even be specifically designed for the objective of civilian protection.[187] As such, there is a nascent body of scholarly research looking into the opportunities AI technologies offer for better protection of civilians in conflict.[188] Computer vision programmes can, for example, be used to recognise protected symbols.[189] AI can also be trained specifically for the protection of civilian infrastructure, in accordance with the rule of distinction.[190]

Chapter 2 will unpack in more detail the role and centrality of data for military targeting and the subsequent legal implications. At this stage, however, it would still be useful to note a number of observations. First, there is the commonly held assumption that the more sophisticated and complex an AI programme is, the more it requires training and testing datasets. While it is not the role, nor the objective of our present discussions to reflect on this assumption (and its veracity), this assertion is only partly complete. [191] In fact, quality of the data matters as much, or perhaps arguably even more to reach accuracy and ultimately, reliability in the system. There are a number of factors that characterise high-quality datasets.

[187] Ruben Stewart, Georgia Hinds 'Algorithms of war: The use of artificial intelligence in decision making in armed conflict' (*Humanitarian Law & Policy*, 24 October 2023) <https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict/> accessed 12 April 2024

[188] Anna Rosalie Greipl 'Artificial intelligence in urban warfare: opportunities to enhance the protection of civilians?' (2023) 61(2) *The Military Law and the Law of War Review*, <https://doi.org/10.4337/mllwr.2023.02.03> accessed 12 April 2024

[189] Larry Lewis, Andrew Ilachinski '(U) Leveraging AI to Mitigate Civilian Harm' (2022) Center for Naval Analyses <https://www.cna.org/reports/2022/02/Leveraging-AI-to-Mitigate-Civilian-Harm.pdf> accessed 12 April 2024

[190] Natalya Kandybko and others 'Application of artificial intelligence technologies in the context of military security and civil infrastructure protection' (2023) 2476(1), AIP Conference Proceedings <https://pubs.aip.org/aip/acp/article/2476/1/030019/2891077> accessed 12 April 2024

[191] Holland Michel (n 13)

These factors include, but are not limited to, the diversity of data, the veracity of the data, as well as the use of proxy data and eventual, subsequent concerns from such use. In essence, the data must be well-targeted with regards to the intended application and use of the system under development, and must capture all the nuances sought for in the system's outputs.

In addition, it would also be useful to note that the quality of the training data is critical, developers and users must be aware of these datasets' limitations. This somewhat builds on the types of biases covered in the previous section. For example, in the context of death rate reporting per country during the COVID-19 pandemic, a holistic dataset is not limited to the inclusion of data from across the world (i.e., geographic inclusiveness), other factors that may affect the data, both directly and indirectly, must also be taken into account. These include, for example, intentional or unintentional under-reporting numbers of COVID-19-related deaths; societal factors including inequalities in access to healthcare facilities; geopolitical-induced inequality in access to vaccines, etc.[192]

In the military domain, much concern is given with regards to noise and 'dirty data' that essentially constitutes snake oil and can even make the system a liability more than an asset.[193] Another risk developers and end-users must take into account corresponds to that of 'data poisoning', or in other words, the deliberate interference with the integrity of an AI system through exposure to manipulated or invalid data in order to induce malfunction, and influence the outcome of the system in the desired way.[194] In short, reliable data is key to reliable algorithms.

---

[192] See for example Emil Kupek 'How many more? Under-reporting of the COVID-19 deaths in Brazil in 2020' (2021) 26(9) Tropical Medicine & International Health <https://onlinelibrary.wiley.com/doi/full/10.1111/tmi.13628> accessed 14 April 2023

[193] Husanjot Chahal, Ryan Fedasiuk, Carrick Flynn 'Messier than Oil: Assessing Data Advantage in Military AI' (2020) CSET Issue Brief, Center for Security and Emerging Technology <https://cset.georgetown.edu/wp-content/uploads/Messier-than-Oil-Brief-1.pdf> accessed 12 April 2024

[194] Mark Visger 'Garbage In, Garbage Out: Data Poisoning Attacks and their Legal Implications' (2022) in Laura A Dickinson, Edward W Berg (eds) *Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold* (The Lieber Studies Series, Oxford Academic, 2024)

3 Artificial Intelligence and Military Targeting

Based on the previous section, we will seek to examine and establish more in detail potential AI applications in the military sphere, with a focus on targeting processes. This study will establish a baseline before diving into the specific implications of these issues on legal compliance in subsequent chapters.

3.1 Military targeting

This thesis will focus on military targeting. While there is no single, universally applicable definition of military targeting, it essentially refers to the process of identifying, selecting, prioritising, validating and engaging targets for attack in an armed conflict. In fact, targeting involves a myriad of processes leading to the use of force, including the collection and processing of data for intelligence creation and analysis; strategic and tactical planning in the wake of the operation; decision-making processes surrounding the selected means and methods of warfare to adopt and employ; training of personnel for the deployment and use of the selected means and methods of warfare; etc. These targets can either be individuals, groups of individuals, or objects such as the adversary's military assets (e.g., tanks, command and control centres) as well as key infrastructure (e.g., bridges, communication towers).

Many states have attempted to provide definitions, or at least doctrinal approaches, as to what military targeting consists of in various domains of operations. Some have even defined a specific targeting cycle, which includes all the relevant stages involved in the conduct of military targeting. For example, the United States' Air Force has published their own doctrine in this space, outlining the different stages in their targeting cycle, consisting of six different phases: 1) The definition of the commander's objectives, targeting guidance, and intent, 2) Target development and prioritisation, 3) An analysis of capabilities available, 4) The

<https://academic.oup.com/book/55259/chapter-abstract/428635911?redirectedFrom=fulltext> accessed 22 November 2024

commander takes decisions surrounding the use and assignment of force, 5) Mission planning and force execution, and 6) Combat assessment, providing feedback to inform future operations.[195] The doctrine also considers the different tools available for targeting automation, including those to assist targeteers throughout the cycle, analytical tools, geospatial intelligence tools, as well as collateral damage estimation tools.[196] In addition, the doctrine also differentiates between 'deliberate targeting' and 'dynamic targeting'. The former corresponds to long-planned targets with very specific actions scheduled for their detection, identification, and engagement; while the latter corresponds to targets that were unplanned and unanticipated, and occurs in much more compressed timelines.[197]

NATO has published their own doctrine for 'joint targeting', which the alliance applies as a standard for joint targeting operations at the strategic, operational and tactical levels and across domains of operations and used as reference by the US, UK, and other NATO allies.[198] However, while states and military alliances may have similar approaches, yet fundamental differences will remain. For example, there is bound to be differences in interpretations of what each stage of the targeting cycle constitutes; the philosophy and military culture behind; the weight and value of each stage; who will possess command and control over each stage; etc. Hence, this thesis will not focus on one particular targeting cycle as a reference, but rather will look at specific IHL rules relevant to military targeting given that they are applicable universally.

---

[195] Air Force (US), *Air Force Doctrine Publication 3-60, Targeting*, 7-9 <https://www.doctrine.af.mil/Portals/61/documents/AFDP_3-60/3-60-AFDP-TARGETING.pdf> accessed 26 February 2023

[196] Ibid

[197] Ibid

[198] 'NATO Standard AJP-3.9, Allied Joint Doctrine for Joint Targeting' (2016) North Atlantic Treaty Organization, Edition A Version 1, Allied Joint Publication <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/628215/20160505-nato_targeting_ajp_3_9.pdf> accessed 5 December 2019

One point to note is that there is increased argument for military targeting operations taking place beyond the kinetic realm, i.e., in the context of cyber operations.[199] Cyber operations will not necessarily be excluded from the discussions. These will inevitably constitute an increasingly important dimension of military targeting, whether because some targeting operations may be done via cyber means, or because of potential cyber interferences in targeting operations– and the role AI may play in these cases. [200] As such, given the expected increasing importance of cyber operations in targeting, it will be presumed that the points we are making as part of this thesis also apply to those operations conducted in the cyber realm.

3.2 Military targeting and artificial intelligence

The transformative potential of AI is well-recognised by States. There is, in fact, the understanding that AI is not merely a tool for technological advancement, but a strategic asset that can not only reinforce national security, but also cutting-edge competitiveness in the conduct of warfare. From addressing key, operational challenges to enhancing situational awareness and supporting decision-making, it is clear that harnessing AI capabilities for battlefield applications has its appeals.[201] From enhancing command and control and information management systems, enabling autonomous features in weapons systems, to supporting logistics and trainings, the integration of AI in most, if not all, military tasks presents tremendous opportunities to the adopting armed forces.[202]

---

[199] NATO has, for example, recognised 'cyberspace' as one of their five domains of operations. See: 'Cyber defence' (*NATO*) <https://www.nato.int/cps/en/natohq/topics_78170.htm> accessed 26 February 2023

[200] For example, there have been reports of malware infecting the cockpits of US Predator and Reaper drones, logging every keystroke of pilots – and thus, providing sensitive information on the technical characteristics of these drones, as well as potentially providing insights to the adversary on the US tactics and modus operandi through these keystrokes. See: Noah Shachtman 'Exclusive: Computer Virus Hits U.S. Drone Fleet' (*Wired*, 7 November 2011) <https://www.wired.com/2011/10/virus-hits-drone-fleet/> accessed 9 April 2020

[201] Tate Nurkin, Julia Siegel 'Battlefield Applications for Human-Machine Teaming: Demonstrating Value, Experimenting with New Capabilities, and Accelearting Adoption' (2023) Atlantic Council Scowcroft Center for Strategy and Security, <https://www.atlanticcouncil.org/wp-content/uploads/2023/08/Battlefield-Applications-for-HMT.pdf> accessed 13 April 2024

[202] Grand-Clément (n 1)

To this end, States are dedicating increasing resources for the development of these capabilities, with defence research agencies dedicating teams and projects on artificial intelligence. For example, the United Kingdom's Defence Science and Technology Laboratory (Dstl), the Ministry of Defence's dedicated research agency in science and technology, is supporting the development and delivery of AI solutions as part of the Australia, UK and US (AUKUS) partnership.[203] At the industry level, big tech companies such as Google, OpenAI and defence contractors alike are developing AI solutions for armed forces around the world, ranging from cloud services to battle management software.[204]

3.2.1 (Potential) Uses of AI for military targeting

Today, the deployment and use of artificial intelligence in conflict far transcends the realms of fiction. In 2020, the UN Panel of Experts on Libya reported the deployment of a Turkish-manufactured drone, the STM Kargu-2, in Libya and which 'hunted down and remotely engaged' logistics convoys and retreating members of the Haftar Affiliated Forces.[205] This was, arguably, one of the first documented instances of AI deployment in the battlefield. Fast forward to today, the on-going conflict in Ukraine is often reported to be a 'testing ground' for new capabilities, including AI technologies.[206] In Gaza, the Israeli military is reported to have used AI systems to identify targets: 'Lavender' for individuals, and 'Habsora', or 'The Gospel'

[203] Defence Science and Technology Laboratory 'Dstl AI success with AUKUS' (*GOV.UK*, 15 December 2023) <https://www.gov.uk/government/news/dstl-ai-success-with-aukus> accessed 13 April 2024

[204] See for example, Billy Perrigo 'Exclusive: Google Contract Shows Deal with Israel Defense Ministry' (*TIME*, 12 April 2024) <https://time.com/6966102/google-contract-israel-defense-ministry-gaza-war/> accessed 13 April 2024; Jon Harper 'Anduril, Palantir awarded contracts for Army robotic combat vehicle software system integration' (*DefenseScoop*, 3 April 2024) <https://defensescoop.com/2024/04/03/anduril-palantir-diu-army-contract-robotic-combat-vehicle-software/> accessed 13 April 2024. More recently, see Biddle (n 4)

[205] United Nations Security Council 'Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011)' (2021) S/2021/229, 17 para 63 and 148 Annex 30 <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/037/72/PDF/N2103772.pdf?OpenElement> accessed 19 April 2023

[206] Morgan Meaker 'Ukraine's War Brings Autonomous Weapons to the Front Lines' (*WIRED*, 24 February 2023) <https://www.wired.com/story/ukraine-war-autonomous-weapons-frontlines/> accessed 26 February 2023; see also Vera Bergengruen 'How Tech Giants Turned Ukraine Into an AI War Lab' (*TIME*, 8 February 2024) <https://time.com/6691662/ai-ukraine-war-palantir/> accessed 13 April 2024

for objects.[207] Widespread reports in the mainstream media have sparked vivid and heated discussions on the deployment and use of AI in the military domain, a subject that until recently seemed so far-off to the greater public.[208] Regardless of where one stands on this issue, one thing is certain: AI will increasingly shape the conduct of warfare, and these technologies will only grow in prominence in the conduct of military targeting operations.

As such, it would be important, at this stage, to take stock on the (potential) uses of AI for military targeting. While the present section does not seek to provide a comprehensive taxonomy of all possible applications of AI technologies in military targeting, the following three main applications ought to be listed in the light of our future discussions within this thesis: intelligence support, choice of means and methods of warfare, and weapons systems enhancement.

First, AI can be used to support the intelligence groundwork required prior to the planning stages of a targeting operation through the collection, aggregation, and processing of vast quantities of data. Insights gained through these data analytics capabilities will feed into the commander's analysis and decision-making processes, which will eventually shape the way targeting operations are conducted. While data will be discussed in further detail in Chapter 2, it is worth noting at this stage that intelligence can be collected in two ways. First, there are 'signals intelligence' (SIGINT), which relies on data collected through sensors, radars, but also

[207] See, respectively, Yuval Abraham ''Lavender': The AI machine directing Israel's bombing spree in Gaza' (*+972 Magazine*, 3 April 2024) <https://www.972mag.com/lavender-ai-israeli-army-gaza/> accessed 13 April 2024 and Yuval Abraham ''A mass assassination factory': Inside Israel's calculated bombing of Gaza' (*+972 Magazine*, 30 November 2023) <https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/> accessed 13 April 2024

[208] The Guardian's coverage on the Lavender system has particularly sparked a lot of debates on Israel's use of AI for military targeting and the role of human operators in an AI-informed kill chain, which even prompted the IDF to issue a response. For The Guardian article, see: Bethan McKernan and Harry Davies ''The machine did it coldly': Israel used AI to identify 37,000 Hamas targets' (*The Guardian*, 3 April 2024) <https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes> accessed 13 April 2024; for the IDF's response, see: 'IDF Response as Sent to the Guardian' (*IDF*, 3 April 2024) <https://www.idf.il/189654> accessed 13 April 2024

from communications online.[209] In principle, these could include, for example, data from social medial platforms, but also data collected from digital infrastructures such as satellites, or even Wi-Fi routers. Second, human intelligence (HUMINT) corresponds to intelligence 'collected and provided by human sources.'[210] Their collection would, however, require 'boots on the ground', which may not always be possible depending on the circumstances. Hence, advanced capabilities for intelligence collection, enabled and enhanced by AI, could prove to be particularly useful in places where the user state has limited access and human sources on the ground that would enable the collection of key intelligence.[211] Such capabilities are already in place and being used by a number of states. Perhaps the most prominent example was Google's Project Maven which, in partnership with the US Department of Defense, consisted of using AI to interpret video images in order to improve the targeting of drone strikes.[212] The UK has acknowledged, in its Defence AI Strategy document, that in May 2021, the British Army used AI for an operation specifically to 'process masses of complex data, including information on the surrounding environment and terrain, offering a significant reduction in planning time over the human team; whilst still producing results of equal or higher equality.'[213] More recently, as part of the AUKUS defence agreement between Australia, the UK and the US, the Dstl has

[209] Frank Sauer 'The Military Rationale for AI', in Thomas Reinhold, Niklas Schörnig (eds) *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm* (Springer Nature Switzerland 2022) <https://link.springer.com/chapter/10.1007/978-3-031-11043-6_3> accessed 27 February 2023

[210] Steven A. Stottlemyre 'HUMINT, OSINT, or Something New? Defining Crowdsourced Intelligence' (2015) 28(3) International Journal of Intelligence and Counterintelligence <https://www.tandfonline.com/doi/full/10.1080/08850607.2015.992760> accessed 27 February 2023

[211] Jennifer Gibson 'Death by Data: Drones, Kill Lists and Algorithms' in Alasdair McKay, Abigail Watson, Megan Karlshøj-Pedersen (eds), *Remote Warfare: Interdisciplinary Perspectives* (E-International Relations Publishing 2021) <https://www.e-ir.info/2021/02/18/death-by-data-drones-kill-lists-and-algorithms/> accessed 22 November 2024

[212] Daisuke Wakabayashi and Scott Shane 'Google Will Not Renew Pentagon Contract That Upset Employees' (*The New York Times*, 1 June 2018) <https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html> accessed 27 February 2023

[213] Ministry of Defence (UK), *Defence Artificial Intelligence Strategy* (n 36)

dedicated considerable research for the integration of autonomy into vehicles for land operations.[214]

Second, AI-enabled technologies can assist with the choice of means and methods of warfare, and decision-making processes surrounding the conduct of operations. This is the case for example with battle management software through the processing of large amounts of data, providing commanders with advanced situational awareness on the battlespace, and subsequently help them 'identify and prioritize warfare resource and course of action options.'[215] In other words, the software will enable the commander to obtain a holistic view over the situation in the intended area of operations against, on the one hand, considerations related to mission success and legal boundaries, and on the other hand data with regards to internal resources available. Through such comprehensive overview, this will enable the commander to improve their decision-making and ultimately make the best use of their resources while increasing chances at mission success and while operating within the boundaries of the law. Another relevant example could be related to helping with the protection of cultural heritage and the environment in times of armed conflict through the identification of high-risk locations where target engagement may go against the duty to preserve such sites, especially as both areas are increasingly gaining traction and the level of awareness in their protection is rising.[216]

Third, AI can be embedded within weapons systems to increase their capabilities. This can be done, for example, through autonomous navigation capabilities, which Lockheed Martin has

[214] Defence Science and Technology Laboratory 'AUKUS trial advances AI and autonomy collaboration' (*GOV.UK*, 5 February 2024) <https://www.gov.uk/government/news/aukus-trial-advances-ai-and-autonomy-collaboration> accessed 13 April 2024

[215] Geoffrey B. Grooms 'Artificial Intelligence Applications for Automated Battle Management Aids in Future Military Endeavors' (Thesis, Naval Postgraduate School 2019) v <https://apps.dtic.mil/sti/citations/AD1080249> accessed 27 February 2023

[216] See: 'Protection of cultural property: military manual' (2016) UNESCO <https://unesdoc.unesco.org/ark:/48223/pf0000246633> accessed 5 December 2019

been developing autonomous underwater vehicles for over 35 years[217], while Airbus' unmanned aircraft systems including with autonomous flight technology have already been supporting military forces for nearly a decade and deployed for missions in combat zones such as Afghanistan and Mali.[218] Other possible applications pertain to hardwiring computer vision capabilities that would enable target identification and, possibly, even systems programmed with target engagement capabilities.

3.2.2 Potential issues from the use of AI for military targeting

In principle, if the programme works as intended, AI holds the potential of increasing mission success in military targeting while also fostering compliance with IHL. The use of AI may result in less personnel deployment in the battlefield in addition to conferring significant advantage over the adversary with enhanced situational awareness and decision-making processes. These promises, however, are understandably met with a lot of criticisms and scepticism, pushing a number of states and civil society organisations to advocate for the prohibition of developing, deploying and using AI technologies for military applications.

Yet, while it is not the objective of this thesis to discuss whether indeed AI technologies should be prohibited for such uses, it is important to note that a blanket ban would neither be feasible, nor desirable. In fact, nuance must be added to the assumption that the development and use of autonomous technologies for military applications would, in and of themselves, be problematic and/or unlawful.[219] In addition to the potential opportunities these technologies may bring from an operational perspective (i.e., for mission success), they may even help, under specific circumstances, increase compliance with the law (e.g., data analytics to support proportionality assessments). A key determining factor with regards to their legality pertains to the intended

[217] 'A-Size Autonomous Underwater Vehicles' (*Lockheed Martin*) <https://www.lockheedmartin.com/en-us/products/a-size-autonomous-underwater-vehicles.html> accessed 11 November 2019
[218] 'Unmanned Aircraft Systems' (*Airbus*) <https://www.airbus.com/defence/uav.html
[219] Slijper (n 55)

applications and uses; hence it is critical that developers and the military work together on hardwiring legal considerations from the earlier stages in the technology's lifecycle 'upstream', in order to ensure that only those systems compliant 'by design' are eventually deployed and used in the battlefield. While the following chapters will unpack more in detail the issues and considerations that may arise with regards to the development (and eventual use) of AI for military targeting, the present section seeks to provide a snapshot of those key question marks.

The first area of issues pertains to data, which we will explore in more detail in chapter 2. Not only does it consist of issues and limitations from the quantity of data, but also their quality, which arguably corresponds to the core, main issues surrounding data and military AI applications. Regarding the former, there is a general assumption that the more training data there is, the better it is to improve the system's accuracy and overall performance. There may, however, be issues with regards to the limited amount of data available for the training stage due to secrecy issues and the sensitivities the military realm is deeply shrouded with. These issues can translate through concealed records, or even through the absence of records due to their deliberate destruction. With regards to the latter (i.e., the quality of data), a number of issues may arise and would require closer examination, which, again, we will do later in chapter 2. Perhaps the most prominent concern with regards to data is risks pertaining to harmful biases. Biases can either stem from the data in itself – as in, the dataset merely reflects larger, deeply rooted societal biases; biases coming from the programmers – i.e., through collective and subjective biases such as groupthink and assumptions; and bias can also manifest itself from the algorithm. This may not only result in the adoption of a course of action that do not necessarily reflect the commander's intent; biases can also lead to unlawful courses of action and outcomes. For example, we will see in more detail in chapter 2 how the absence of an adequate training sample that is not representative of the targeted group of persons can lead to inaccurate results and the engagement of unlawful targets. Take, for example, the controversy

in 2020 surrounding Google's Vision Cloud's computer vision programme, which had been discussed earlier.[220] While Google did end up rectifying the error, it raises concerns or at least, questions with regards to the company's training datasets that seemingly reflect larger, societal issues (i.e., racism). More generally, this incident raises concerns regarding the reliability of classification programmes akin to this one to assist with military targeting operations. More discussions surrounding the quality of data – and the subsequent legal implications – will be elaborated in more detail in chapter 2.

The second area of issues pertains to the design and development processes, in themselves, of AI-enabled programmes. More specifically, beyond data, another key consideration while developing AI technologies 'in the lab' relates to the very idea that machine learning is trained and tested for anti-personnel targeting. One of the arguments made by advocates for a ban in the early days of discussions and deliberations of military AI is that decisions involving life-and-death matters should not be made by AI technologies; as otherwise such decisions would go against existing moral and ethical standards. How does this issue translate with regards to international humanitarian law? Chapter 3 will cover this question more in detail, in addition to presenting an approach challenging the human-centric focus of discussions and deliberations so far; notably in the light of discussions bringing 'human control' to the fore as a silver bullet to legal and ethical issues surrounding military AI. The third key question on this topic chapter 3 will look at pertains to the ever-learning nature of certain machine learning technologies, and whether legal review mechanisms, notably those established within the framework of Article 36 API, are fit for purpose.

The third and final area of issues this thesis will look at pertains to the question of uncertainties and the legal implications of knowns versus unknowns. This question is not only of high

[220] Kayser-Bril (n 169)

relevance in the light of the messy, unpredictable and dynamic nature of war – bound to be shrouded with uncertainty some describe as the 'fog of war'. It is also of relevance with regards to black box models and the legal implications of 'not knowing' the processes on how AI-enabled programmes reach their final results. On the latter, some possible approaches to addressing this opacity could be with regards to commander's responsibility, as well as the legal framework surrounding the conduct of investigations of alleged violations of IHL. In the light of these uncertainties coming not only from black box but also other factors chapter 4 will be covering in more detail, it leads us to the question of whether there is, under IHL, indeed a duty for accurate targeting and, more broadly, a legal duty for due diligence as a way for us to address the implications of such uncertainties.

Chapter II – Datasets

1 Introduction

Data plays a critical role in shaping the performance and outputs of AI technologies. From the earliest stages of the technology's lifecycle, i.e., the testing, evaluation, verification, and validation (TEVV) processes, data constitutes AI's lifeblood.[221] Data's importance stretches beyond these stages and is as decisive in the deployment and eventual use of the technology; however, in the light of the present thesis' focus on the pre-deployment stages of AI technologies, the present chapter will not discuss in much extent the post-deployment phases in their lifecycle.

The quality and quantity of datasets used to develop and train AI-enabled programmes will indeed bear great influence not only over their performance (i.e., how the system achieves its designated task based on a number of metrics such as speed and efficiency), but also the quality and reliability of their final output (i.e., the extent to which the output is useful, factually correct, and may be evaluated against a set of qualitative metrics – e.g., compliance with international law and ethical requirements, mission success rate, etc.)[222] While the latter will not exclusively depend on training data as the sole determining factor, it is often asserted that

[221] 'Testing, Evaluation, Verification and Validation', often referred to by its acronym (TEVV), corresponds to the set of processes AI systems go through between the development stages and its deployment: TEVV processes are key to ensure the system's reliability and performance, and will inform the decision as to whether or not it is ready for deployment, or whether it needs to undergo further training. In addition to technical reliability, possible evaluation metrics include: whether the system performs as intended; whether the system provides operational advantage; whether the system meets the legal and ethical requirements set by the testing armed forces; etc. Beyond the technical methods and solutions for TEVV, the latter involves an entire ecosystem of processes, infrastructure and workforce that underpin their execution. See: Michèle A. Flournoy, Avril Hailes, Gabrielle Chefitz 'Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems' (2020) WestExec Advisors <https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf> accessed 10 May 2024 and Jan Maarten Schraagen 'Responsible use of AI in military systems: prospects and challenges' 66(11) Ergonomics <https://doi.org/10.1080/00140139.2023.2278394> accessed 10 May 2024

[222] Department of Defense (US), Research & Engineering, Autonomy Community of Interest (COI), Test and Evaluation, Verification and Validation (TEVV) Working Group, *Technology Investment Strategy 2015-2018* (2015) <https://defenseinnovationmarketplace.dtic.mil/wp-content/uploads/2018/02/OSD_ATEVV_STRAT_DIST_A_SIGNED.pdf> accessed 18 November 2020

an algorithm's output would only be as good as the input data.[223] In other, colloquial words, the impact of training datasets can be summarised in four words as 'garbage in – garbage out'.[224]

In the study of data's role in the development, training, testing, evaluation, verification and validation of AI technologies, two specific and essential aspects ought to be dissected: Quantity, and quality. While both quantity and quality are strongly interconnected and, even, interdependent; it is worth unpacking both aspects in detail, especially in the context of military targeting and, subsequently, their legal implications, which this chapter will seek to do later. At this stage, however, it is important to provide an introduction as to why the quantity and quality of training datasets are important – with more detailed discussions to follow.

First, there is widespread acknowledgment that, in order for AI to perform reliably, the quantity of training datasets will be key. One notable example in which AI-enabled technologies are of increasing use is the medical field, specifically classification algorithms that would enable the detection and/or confirmation of diagnostics, such as diabetes.[225] To this end, large volumes of training data that have been curated and annotated are necessary to train the programmes; yet, developers of such technologies often report suffering from the insufficiency of such datasets.[226] While developers are actively trying to find methods to circumvent and address

[223] S Neelamegam, E Ramaraj 'Classification algorithm in Data mining: An Overview' (2013) 3(5) International Journal of P2P Network Trends and Technology <http://www.ijpttjournal.org/volume-3/issue-8/IJPTT-V3I8P101.pdf> accessed 19 November 2020

[224] M F Kilkenny, K M Robinson 'Data quality: "Garbage in – garbage out"' (2018) 47(3) Health Information Management Journal <https://journals.sagepub.com/doi/pdf/10.1177/1833358318774357> accessed 18 May 2023

[225] Mehedi Hassan, Swarnali Mollick, Farhana Yasmin 'An unsupervised cluster-based feature grouping model for early diabetes detection' (2022) 2 Healthcare Analytics <https://www.sciencedirect.com/science/article/pii/S2772442522000521> accessed 8 July 2024

[226] This issue was raised by many participants to a roundtable held by Chatham House on 8 May 2023 on 'Harnessing Digital Technologies for Universal Health' in Kuala Lumpur, Malaysia, which the author attended. Due to the roundtable being held under the Chatham House rule, such claims cannot be attributed to participants. More generally speaking, the insufficiency of annotated training data for machine learning applications in the medical field is a recognised issue in literature. See for example: Misgina Tsighe Hagos, Shri Kant 'Transfer Learning based Detection of Diabetic Retinopathy from Small Dataset' (2019), arXiv <https://arxiv.org/abs/1905.07203> accessed 26 May 2022

issues related to this scarcity, it is however important to point out that without sufficient training data, AI-enabled programmes will not be able to run and produce accurate and reliable outputs. In many cases, aside from ethical and legal considerations, the sheer lack of training data stands in the way of developing technologies viable for commercialisation and subsequent deployment and use due to the likely low level of accuracy resulting from this scarcity.

Second, in addition to the quantity of training datasets, its quality also plays an important role in shaping the performance of a programme and its outputs. For example, there has been considerable research undertaken in the field of natural language processing use for online abuse detection on online platforms, for which the adequate training datasets would be needed to develop such tools.[227] While a number such datasets are available and ready for use, they also differ from one another (e.g., annotation method employed, availability in open source) and thus, each comes with their own shortcomings when, left unaddressed, may lead to a 'garbage in, garbage out' situation.

In the same vein, the quantity and quality of datasets to train AI-enabled technologies for military targeting will equally be critical to ensure their ability to comply with international humanitarian law. This chapter will, later on, unpack in further detail each of both components.[228] First, however, our discussions would benefit from a general overview of what data consists of, how it can be defined, and establish in what ways data could be relevant to military targeting before a deep dive into the key legal issues.

1.1. Definition

In a general sense, data may be defined as 'information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electric

[227] B Vidgen, L Derczynski 'Directions in abusive language training data, a systematic review: Garbage in, garbage out' (2020) 15(12) PLOS ONE, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0243300> accessed 26 May 2023
[228] See Sections 2 and 3 of Chapter 2.

form that can be stored and used by a computer'.[229] Building on this general definition, data may be considered as pieces which, put together, will result in a fuller picture of the subject in question. For example, one individual's age is a piece of data which, in and of itself, does not provide much information on the individual other than their age. However, coupled with other data from the same individual (e.g., their gender, ethnicity, and live location), these data can provide information on this specific individual – which one can then use to make correlations and draw eventual hypotheses and conclusions on this specific individual. These insights could, in this instance, be related to the individual's exercising habits and how they correlate vis-à-vis their professional background. Or, hypotheses could be drawn related to the probability that the individual subject will use public transportations, what for, and what time of the day.

Data collection – and more generally the collection of information beyond computerized data – has always been of interest and necessary both for commercial purposes and by public bodies. For example, data gathering was a key activity for Western colonial administrators, companies and topographers to exercise influence and power over South-East Asia in the 19th Century.[230] In the same vein, historical literature on warfare and the military has always recognized the importance of information for obtaining military advantage over the adversary.[231] In fact, put together, data could constitute key intelligence that would enable parties to an armed conflict to obtain such advantage. For example, during the Second World War, the English scientist Alan Turing played a key role in devising techniques to break the German Enigma ciphers, which were used by the German military to encipher and exchange top-secret messages.[232]

---

[229] 'Data' (*Cambridge Advanced Learner's Dictionary & Thesaurus*), <https://dictionary.cambridge.org/dictionary/english/data> accessed 18 November 2020

[230] Farish A. Noor, *Data-Gathering in Colonial Southeast Asia 1800-1900: Framing the Other* (Amsterdam, University Press, 2019) <https://www.aup.nl/en/book/9789463724418/data-gathering-in-colonial-southeast-asia-1800-1900> accessed 28 May 2023

[231] See: Sun Tzu, *The Art of War* (5th century BC)

[232] 'What Alan Tuirng means to us' (*The Alan Turing Institute*, 23 June 2019) <https://www.turing.ac.uk/blog/what-alan-turing-means-us> accessed 29 May 2023

Access to this critical intelligence played a key role in securing Allied victory in the Battle of the Atlantic.[233]

Today, with the increasing digitization of society and our growing dependency on digital technologies, tremendous amounts of information are produced daily in the form of data. These types of information are two-fold: 1) The digitalization of previously analogue records (e.g., there are efforts worldwide to push for the digitization of land registry)[234]; and 2) The generation of new data from the emergence of new technologies, from smartphones, the internet of things, or even from the use of generative AI accessible to all, akin to ChatGPT. Coupled with rapid and ground-breaking progress in AI technologies and, more generally, computer science, this technological ecosystem would enable the collection and processing of massive amounts of data. Given the omnipresence of data across all segments in society, harnessing these technologies would enable access to advanced predictive analytics capable of assisting and even changing day-to-day decision-making. These data analytics tools are central to certain commercial applications, such as for example targeted advertisement, notably put in the spotlight in 2018 with the infamous case of Cambridge Analytica.[235]

Beyond the private, for-profit sector, public bodies are also increasingly using – or even depending – on AI technologies to harness the vast amount of data available for the exercise of public functions. For instance, the Office of the United Nations High Commissioner for Refugees (UNHCR) released, in 2021, a paper which looks at the use of big data sources for the modelling and prediction of forcibly displaced persons, along with possible avenues for the Agency to include these insights into their work.[236] However, in the light of its high

[233] S Budiansky 'German vs. Allied Codebreakers in the Battle of the Atlantic' (2002) 1(1) International Journal of Naval History, <https://www.ijnhonline.org/wp-content/uploads/2012/01/pdf_budiansky1.pdf> accessed 29 May 2023

[234] Aparajitya Goyal 'Benefits of Land Registry Digitization' (*World Bank Blogs*, 17 April 2012) <https://blogs.worldbank.org/en/digital-development/benefits-of-land-registry-digitization> accessed 9 July 2024

[235] Hinds, Williams, Joinson (n 182)

[236] 'Big (Crisis) data for predictive models' (2021) UNHCR <https://www.unhcr.org/media/big-crisis-data-predictive-models-literature-review> accessed 10 July 2024

stakes, especially as the Agency's work has a direct impact on highly vulnerable populations and individuals, there is also the recognition that the use of such models must be carefully evaluated against their 'relevance, accuracy, and contribution of big different data sources'.[237]

Similarly, during the COVID-19 pandemic, the UK's Department for Education came under close scrutiny due to their contentious use of AI to award students' A-level grades as a result of the cancellation of teaching and examinations.[238] This move has been heavily criticized and even resulted in many protests by professionals in the education sector and students alike which, ultimately, led to the Department for Education abandoning this approach.[239] Cities are increasingly using AI technologies to help with crowd management and the anticipation of unsafe situations through data analytics.[240] Yet, the use of predictive analytics for law enforcement is heavily criticized – especially for 'predictive policing', where models are being used to predict when and where crimes are likely to happen.[241]

The security and defence sector are far from being spared by this data frenzy. In fact, AI-enabled data analytics play a central role in enabling the conduct of today's military operations, as alleged more recently in the context of the on-going Russian invasion on Ukraine.[242] Palantir, a data-analytics firm known to provide AI solutions including for defence, is publicly known to support Ukraine's defence. Its software is reported to analyse 'satellite imagery,

---

[237] Andrea Pellandra, Geraldine Henningsen 'Predicting refugee flows with big data: a new opportunity or a pipe dream?' (*UNHCR Blogs*, 6 January 2022) <https://www.unhcr.org/blogs/predicting-refugee-flows-with-big-data-a-new-opportunity-or-a-pipe-dream/> accessed 10 July 2024

[238] Elliot Jones, Cansu Safak, 'Can algorithms ever make the grade?' Ada Lovelace Institute, 18 August 2020, <https://www.adalovelaceinstitute.org/blog/can-algorithms-ever-make-the-grade/> accessed 27 May 2022

[239] Jon Porter 'UK ditches exam results generated by biased algorithm after student protests' (*The Verge*, 17 August 2020) <https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications> accessed 10 July 2024

[240] 'Public Eye: an open-source crowd monitoring solution' (*City of Amsterdam*) <https://www.amsterdam.nl/innovation/mobility/public-eye/> accessed 29 May 2023

[241] Will Douglas Heaven 'Predictive policing algorithms are racist. They need to be dismantled.' (*MIT Technology Review*, 17 July 2020) <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/> accessed 10 July 2024

[242] George Grylls, 'Kyiv outflanks analogue Russia with ammunition from Big Tech' (*The Times*, 24 December 2022) <https://www.palantir.com/assets/xrfr7uokpv1b/1Fw2bFxYXmu3RWX7FvssB9/e64d19b6f042bda3d2a61e4fc43ea6ec/TheTimes.pdf> accessed 29 May 2023

open-source data, drone footage, and reports from the ground to present commanders with military options'; and it is allegedly 'responsible for most of the targeting in Ukraine'.[243] The following section will dissect in more detail what data is of interest to the military, how it is collected, and how this plays out in the context of targeting. At this stage, however, it is useful to note that, more generally, the sector is trying to grapple with the sheer amount of data generated and available in the light of the possible opportunities and challenges.[244]

Additionally, open-source data has notably put into questioning commonly-held assumptions on the secrecy of certain information.[245] For example, data shared on a fitness app, Strava, collected from smart watches has unveiled the presence, and the map, of secret military bases.[246] Data from that same app has reportedly been used to track and target a Russian submarine commander while he went on a run; although the Ukrainian armed forces has sought to downplay speculations that Kyiv may have been behind this operation.[247] A flashcard app has unearthed information pertaining to US nuclear weapons stationed in Europe, which otherwise is shrouded in secrecy.[248] These examples show that in today's context, the diversity of data sources is such that data from fitness and learning apps can reveal the most unexpected, classified information – and AI is only going to disrupt this space even further.

---

[243] Bergengruen (n 206)

[244] The UK Army has, for example, a plan document specifically dedicated to this. See: Army (UK) The Army Digital & Data Plan 2023-2025: A guide to help you deliver the Army's Digital Transformation, Edition 1.0 (2023) <https://www.army.mod.uk/media/21608/2_295200-mod_addp_review-file_05.pdf> accessed 29 May 2023

[245] For more on the implications of open-source intelligence on security, see: Ardi Janjeva, Alexander Harris, Joe Byrne (2022) 'The Future of Open Source Intelligence for UK National Security', RUSI, <https://rusi.org/explore-our-research/publications/occasional-papers/future-open-source-intelligence-uk-national-security> accessed 16 June 2022

[246] Alex Hern 'Fitness tracking app Strava gives away location of secret US army bases' (*The Guardian*, 28 January 2018) <https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases> accessed 1 February 2022

[247] Mariya Knight 'Russian commander killed while jogging may have been tracked on Strava app' (*CNN World*, 12 July 2023) <https://edition.cnn.com/2023/07/11/europe/russian-submarine-commander-killed-krasnador-intl/index.html> accessed 10 July 2024

[248] Foeke Postma 'US Soldiers Expose Nuclear Weapons Secrets Via Flashcard Apps' (*Bellingcat*, 28 May 2021) <https://www.bellingcat.com/news/2021/05/28/us-soldiers-expose-nuclear-weapons-secrets-via-flashcard-apps/> accessed 22 November 2024

1.2. Relevance to military targeting

In the context of military targeting, the collection and processing of data is essential in all stages of targeting cycles, ranging from the intelligence collection and processing stages to the deployment and engagement of the target.[249] As alluded to just earlier, this data and their sources are incredibly – and unprecedently – diverse. From 'classic' data such as maps and satellite imagery to more 'trivial' sources such as the aforementioned cases of fitness and flash card apps, it is undeniable that whoever is able to deploy the appropriate technologies to leverage this data will obtain cutting-edge advantage, over adversaries, due to the unique insights these data can reveal. This is even further facilitated through the 'internet of battlefield things', a concept that encapsulates the fact that sensors, wearables and connected devices in the battlefield tend to all be part of a same network, thus constituting a key enabler in the collection and use of such data for military targeting.[250]

First, AI would enable the collection, aggregation, processing and collation of this kaleidoscope of data altogether; thus enabling better live situational awareness for the commander. For example, an AI-enabled programme could collect and assemble, altogether, data related to the mapping of the battlefield (e.g., roads, presence of protected buildings such as hospitals, residential areas, etc.), the area's topography, the live weather on site, as well as indicators of

[249] Each state has identified, shaped and implements their own military targeting cycles and practices vary from one state to another. There may be instances where states may form an alliance and harmonise their military targeting cycle – for instance, in the case of NATO allies, or even form one specific and common joint task force as part of an inter-state coalition for a specific mission whereby military efforts against a common adversary are coordinated – for instance, the Operation Inherent Resolve with military personnel from 30 countries. See: 'About CJTF-OIR' (*Operation Inherent Resolve*) <https://www.inherentresolve.mil/About-CJTF-OIR/> accessed 18 November 2020

[250] For more on the internet of battlefield things, see for example the work of the US Army, in collaboration with a consortium of research partners, including the University of Massachusetts and Carnegie Mellon University: 'Internet of Battlefield Things (IoBT) CRA' (*DEVCOM*) <https://www.arl.army.mil/business/collaborative-alliances/current-cras/iobt-cra/> accessed 27 May 2022. For a literature review summarising the key dimensions, see: Lin Zhu, Suryadipta Majumdar, Chinew Ekenna 'An invisible warfare with the internet of battlefield things: A literature review' (2021) 3 Human Behaviour & Emerging Technologies <https://onlinelibrary.wiley.com/doi/epdf/10.1002/hbe2.231> accessed 27 May 2022

civilian presence (e.g., from the metadata of civilian mobile phones).[251] In fact, one could even argue that there is 'too much' data; hence, AI-enabled tools will enable the collection and processing of these data in a way that is useful for the end-users. Put together, with the caveat that the technology works as intended and functions reliably, not only will the commander's situational awareness be improved; which subsequently increases chances of mission success. Such AI applications would also foster, in principle, compliance with IHL – for example through proportionality assessments and ultimately decreasing risks of civilian harm.

Second, the collection and processing of open-source data could reap serious tactical and strategic benefits. For example, again, in the context of the Russian invasion of Ukraine, a heatmap of Russian SIMs roaming and connected to the network in Ukraine provided insights on the presence and concentration of Russian forces across the country.[252] If used by the Ukrainian armed forces, such creative approach to using open-source data could indeed shape their strategic and tactical decision-making, and feed into subsequent actions and operations, including lethal targeted strikes and other offensive campaigns against Russian forces.

Another documented use of metadata from mobile phones for military targeting relates to the Skynet surveillance programme by the US National Security Agency (NSA). By using metadata on the location and call logs from mobile phones in a certain, pre-defined area, the programme would collect and subsequently aggregate this data to identify 'suspicious' patterns

[251] These are all data that, separately, are instrumental in enabling military targeting operations – even more so when put together to increase situational awareness. In fact, the use of AI for such applications is of interest even since the 20th century – See for example on using AI to integrate weather-related data for decision-making: D. Ballard, L. Owsley, 'Artificial intelligence in the helicopter cockpit of the future' (1991) *IEEE/AIAA 10th Digital Avionics Systems Conference*, Los Angeles, USA, <https://ieeexplore.ieee.org/abstract/document/177154> accessed 29 May 2023, and more recently Michael O'Donnell and others, 'Roadmap to Implement Artificial Intelligence in Course of Action Development & Effect of Weather Variables on UH-60 Performance' (2021), *Proceedings of the Annual General Donald R. Keith Memorial Conference*, West Point, USA, <http://www.ieworldconference.org/content/WP2021/Papers/GDRKMCC_21_61.pdf> accessed 29 May 2023. And in principle, programmes akin to the one used by the City of Amsterdam, for example, to live monitor crowdedness could be used for the purpose of monitoring civilian movement (n 240)

[252] Mike Martin (@ThreshedThought) 'Russian SIMs roaming in Ukraine. Gives you a good idea of the concentration of their forces.' (*Twitter*, 12 May 2022 11:48 AM) <https://twitter.com/ThreshedThought/status/1524687923676954624> accessed 9 June 2022.

that could indicate someone's affiliation to Al-Qaeda.[253] Assuming that this programme is indeed reliable, this information would then help feed into the commander's decision-making – including for targeted lethal strikes.

Third, beyond aiding the very conduct of military targeting operations, data also plays a central role in training and planning. In fact, another way in which the combination of data with AI could be leveraged is through the development of advanced, predictive models that will feed into planning and subsequent decision-making processes. These models could in fact help, for example, predict the likely effects and consequences of an attack by drawing on past, historical data combined with information surrounding the contextual environment in which the attack is to be carried out.[254] Armed forces and software companies are indeed developing 'synthetic environments', drawing on historical and live data, and even possibly synthetic data, to re-create and model the battlefield and help decision-makers not only simulate and estimate the effects of an attack, but also its likely effects in the overall environment of operation and on the agents present.[255] In fact, endeavours to develop such models have been documented for over 30 years now, with literature tracing back to the 1990s, with the US in particular having been eager to develop such tools to obtain a three-dimensional view of the environment with active agents ('moving and stationary vehicles') and obtain different views of the battlefield.[256]

---

[253] Cora Currier, Glenn Greenwald, Andrew Fishman 'U.S. Government Designated Prominent Al Jazeera Journalist as "Member of Al Qaeda' (*The Intercept*, 8 May 2015) <https://theintercept.com/2015/05/08/u-s-government-designated-prominent-al-jazeera-journalist-al-qaeda-member-put-watch-list/> accessed 19 November 2020; see also Kim Zetter 'So, the NSA Has an Actual Skynet Program' (*WIRED*, 8 May 2015) <https://www.wired.com/2015/05/nsa-actual-skynet-program/> accessed 19 November 2020

[254] Murray D, Public evidence for the Artificial Intelligence in Weapons Systems House of Lords Select Committee, 'AI in Weapon Systems Committee' (London, 23 March 2023) <https://committees.parliament.uk/oralevidence/12931/pdf/> accessed 29 May 2023

[255] For example, see the US Army Research Laboratory DEVCOM's report: Rao, De Melo, Krim (n 151); also 'What is a synthetic environment?' (*Improbable Defence*) <https://defence.improbable.io/what-is-a-synthetic-environment/> accessed 29 May 2023,

[256] Rex G. Haddix II, 'An Immersive Synthetic Environment for Observation and Interaction with a Large Volume of Interest' (Thesis, Air Force Institute of Technology 1993) <https://apps.dtic.mil/sti/citations/ADA262547> accessed 30 May 2023

In this sense, it is undeniable that data plays a key role not only in intelligence gathering and forming, but also in the actual conduct of targeting operations. And its role will only grow increasingly critical in the light of the new and emerging AI-enabled tools that are under development by armed forces and defence companies across the world. Ultimately, reliable data is indeed key to reliable programmes – which, ultimately, would not only enable mission success from a technical perspective, but also constitutes a form of guarantee that the programme has been developed with legal compliance in mind. Yet, as alluded to earlier, a programme will only be as good as the data it has been and is being fed with.[257] Hence, a number of key issues inherently associated with the quality and quantity of data must be covered, in order for us to then dive deeper into their legal implications.

2. Quality of data

A number of factors can affect the quality of training and testing datasets and, subsequently, the performance of an algorithm and its final output. First, the veracity of these data is key to ensuring the quality of training and testing datasets. In other words, if the data is not accurate, nor adequate for the algorithm's intended purpose, its output will inevitably be affected and ultimately decrease their reliability for use. Second, the diversity of training and testing datasets, i.e., their sheer richness, can also affect their quality. In other words, it is a question of whether or not these datasets are representative enough of the possible nuances and variables of the environment in which the programme is expected to operate. Third, the use of proxy data – and their quality – for the training and testing of programmes can also shape their output. This section will dissect, in detail, these three factors – what they mean, why they are important, and how they are relevant to military targeting – before inviting the reader for deeper reflections on their legal implications.

[257] Neelamegam, Ramaraj (n 223)

2.1. Veracity of data

The first factor that can affect the quality of training and testing datasets pertains to the veracity of the data. Before diving deep into the relevant legal considerations, we will first examine, in detail, the relevant dimensions to the veracity of data. The latter would indeed depend on a number of considerations – namely, accuracy in time and space; human involvement; reliability of the source; and risks of data poisoning. What each of these considerations mean and how they are relevant to military targeting will be central to our discussions.

2.1.1. Relevant dimensions to the veracity of data

Accuracy in time and space

First, accuracy in time and space is central to the veracity of data. While historical data is important, measures must be in place to ensure that the data remains indeed relevant and true through time, but also with regards to location. Data can, for example, be incorrect/false simply because it is outdated. From datapoints on demographics to those on the socio-cultural context, they must be updated regularly in order to be true. In predominantly Muslim cultures, for examples, major gatherings are to be expected around specific days that follow the Lunar calendar – thus the calendar dates will vary from year to year.

Veracity in time and space is particularly relevant for maps – as the disposition and layout of roads may be different, property developments may be absent, and along with other infrastructures such as bridges, water reservoirs, and the presence of parks and other green spaces. For example, in the context of the Russian invasion on Ukraine in February 2022, some Russian commanders were allegedly relying on outdated topographic maps from 1969.[258] The

[258] See 'Russian military received outdated maps of Ukraine prior to invasion' (*Militarnyi*, 18 July 2022) <https://mil.in.ua/en/news/russian-military-received-outdated-maps-of-ukraine-prior-to-invasion/> accessed 9 June 2023; 'СБУ встановила, що російські командири вторглися в Україну, керуючись картами з минулого століття' [SBU found that Russian Commanders Invaded Ukraine using Maps from the Last Century] (*Security*

outdated maps failed to take into account residential developments in the city of Kharkiv, as well as water reservoirs, both of which were built in the 1970s. Not only did this outdated information resulted in losses to the Russian Brigade using these maps; there are also risks pertaining to compliance with international humanitarian law. Failure to account for the residential developments may have led the Russian forces to unknowingly target an area with a high concentration of civilians. While eventual civilian casualties may have been accidental – because the maps were so outrageously outdated, being 53 years old, and because there are technologies available today that would have enabled the Russian forces to obtain an updated mapping of the area, there could be a case made for their failure to comply with the rule of precautions in attack. Had they taken the appropriate precautionary measures, including to obtain and use up-to-date maps, this hypothetical loss of civilian lives and damage to civilian infrastructure would have been prevented.

In this sense, the use of outdated maps to train and test AI-enabled technologies designed to assist with military targeting will shape their outputs. These outputs would not only be unreliable and pose risks to mission success; they could also lead to preventable civilian casualties, had the programme been trained with data that is up to date. One thing to note is – developers have been trying to tap into the availability, today, of high-quality satellite imagery by integrating the latter into their programme. This is the case of Palantir, for example, with their 'AI Platform for Defense', helping commanders decide on the appropriate courses of action in the battlefield based on a number of parameters – including the time required, the availability of personnel and vehicles, assets at hand, and live satellite imagery providing an up-to-date overview of the battlefield.[259] This would enable the programme to produce outputs

Service of Ukraine, 18 July 2022) <https://ssu.gov.ua/novyny/sbu-vstanovyla-shcho-rosiiski-komandyry-vtorhlasia-v-ukrainu-keruiuchys-kartamy-z-mynuloho-stolittia> accessed 9 June 2023

[259] 'AIP for Defense' (*Palantir*) <https://www.palantir.com/platforms/aip/> accessed 9 June 2023

based on up-to-date data, generate better recommendations, and ultimately minimising risks of accidental civilian harm.

Human involvement

Training and testing datasets must be collected, curated, labelled when appropriate (e.g., for supervised learning settings), and processed before even being fed into the programme.[260] While, today, data collection is increasingly automated, subsequent processes would still need the involvement and intervention of humans. Humans are indeed needed in instances where, for example, training data needs to be labelled, or certain data would be undesirable for the training and testing of the programme in question and therefore need to be filtered out manually – to avoid a 'garbage in, garbage out' situation as described earlier. This was the case for example with OpenAI's ChatGPT, which infamously required Kenyan workers to detect, label and filter out toxic content from the programme's training datasets.[261] This manual labour done by human labellers was meant to ensure that ChatGPT would produce outputs that are safe for public usage. Similarly, through project Maven, the US DoD sought to create an automated target recognition (ATR) system, i.e., an AI-enabled programme that, as its name suggests, would identify targets in a set environment. The training of project Maven's algorithms is reported to have required over 100,000 hand-labelled images for each object the system ought to recognise, involving 'a large group of people – sophisticated analysts and engineers' to manually go through the training data and clean it up.[262]

---

[260] For more on data in the context of military AI, see: Yasmin Afina, Sarah Grand-Clément 'Bytes and Battles: Inclusion of Data Governance in Responsible Military AI' (2024) Centre for International Governance Innovation <https://www.cigionline.org/publications/bytes-and-battles-inclusion-of-data-governance-in-responsible-military-ai/> accessed 22 November 2024

[261] Chloe Xiang 'OpenAI Used Kenyan Workers Making $2 an Hour to Filter Traumatic Content from ChatGPT' (*VICE*, 18 January 2023) <https://www.vice.com/en/article/wxn3kw/openai-used-kenyan-workers-making-dollar2-an-hour-to-filter-traumatic-content-from-chatgpt> accessed 16 June 2023

[262] Cheryl Pellerin 'Project Maven to Deploy Computer Algorithms to War Zone by Year's End' (*DOD News*, 21 July 2017) <https://www.defense.gov/News/News-Stories/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/> accessed 16 June 2023, see also Gavin S. Hartnett and others, 'Operationally Relevant Artificial Training for Machine Learning: Improving the Performance of Automated

And while the intervention of humans in the learning of AI programmes can indeed lead to desirable outputs, it is also not the ultimate silver bullet for perfectly functioning AI. In fact, human error and subjective experience can affect the veracity of training data – and ultimately lead to inaccurate outputs.[263] In law enforcement, the subjectivity of police work and its subsequent impact on predictive policing technology faces growing scrutiny.[264] If the same issue arises in the military domain and subjective biases are such that they affect the veracity of ATR systems' outputs, this could not only have operational repercussions but also possibly of legal order.

Another way in which human involvement could affect the veracity of data relates to high observer variability. In medicine, for example, this issue relates to situations where one same medical sample are diagnosed differently by one or more doctors and thus, affecting the way the data is presented, used and interpreted.[265] In other words, one same MRI scan can for example be interpreted differently, depending on who is looking at and analysing the scan, subsequently leading to different – and incoherent – diagnoses. If used to train an AI-enabled programme, these data will have a certain degree of variability, subsequently affecting its overall veracity.

In the military realm, high observer variability can also be present. For example, in the context of the armed conflict between Israel and Hamas in May 2021, Israel has targeted and launched a strike on the al-Jaala Tower in Gaza, which houses the offices of a number of media outlets, including the Associated Press News (AP), as well as Al-Jazeera. Israel had justified the

---

Target Recognition Systems' (2020) RAND <https://www.rand.org/pubs/research_reports/RRA683-1.html> accessed 17 June 2023

[263] Arthur Holland Michel 'Known Unknowns' (2021) UNIDIR <https://unidir.org/known-unknowns> accessed 18 May 2021

[264] Ajay Sandhu, Peter Fussey 'The 'uberization of policing'? How police negotiate and operationalise predictive policing technology' (2021) 31(1) Policing and Society <https://www.tandfonline.com/doi/full/10.1080/10439463.2020.1803315> accessed 10 July 2024

[265] See for example: Zoran B. Popović, James D. Thomas 'Assessing observer variability: a user's guide' (2017) 7(3) Cardiovasc Diagn Ther <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440257> accessed 14 July 2022

targeting of the building on the basis that 'Hamas military intelligence was operating inside the building', while the group would use the journalists as human shields.[266] In a Twitter thread, the Israel Defense Forces (IDF) further justified their attack on the tower, on the basis that the building was used to manufacture weapons and 'positioned equipment to hamper IDF operations'.[267]

The AP and journalists from Al Jazeera, however, both claim that they had been operating in the building for years and had no indication that there were indeed intelligence activities undertaken by Hamas in the building.[268] Despite calls for evidence on Hamas activities in the building from a number of States, including the US, as well as media outlets, Israel has so far not provided such evidence.[269] Here, what is supposed to be a factual reality, i.e., the presence of Hamas activities in the Al-Jaala Tower), is highly contested. Each party has their own, competing version of the facts; and depending on who is in the right and who is in the wrong, this would heavily affect the extent to which that particular strike was in compliance, or not, with IHL.

If, for example, Israel were to use this historical data to train a decision support system (e.g., akin to the Habsora or Lavender, widely reported to be used in the on-going conflict in Gaza).[270]

---

[266] Fares Akram, Lee Keath 'Israel strikes Haza home of Hamas leader, destroys AP office' (*AP*, 16 May 2021) <https://apnews.com/article/israel-west-bank-gaza-middle-east-israel-palestinian-conflict-7974cc0c03897b8b21e5fc2f8c7d8a79> accessed 21 May 2021

[267] Israel Defense Forces (@IDF) 'Yesterday, we targeted an important base of operations for Hamas' military intel in Al Jala Tower in Gaza. The base gathered intel for attacks against Israel, manufactured weapons & positioned equipment to hamper IDF operations. (1/3)' (*Twitter*, 16 May 2021 07:14 AM) <https://twitter.com/IDF/status/1393797067273867266> accessed 21 May 2021

[268] Ibid, see also: 'Gaza tower housing Al Jazeera office destroyed by Israeli attack' (*Al Jazeera*, 15 May 2021), <https://www.aljazeera.com/news/2021/5/15/building-housing-al-jazeeera-office-in-gaza-hit-by-israeli-strike> accessed 21 May 2021

[269] Hope Yen, 'Blinken hasn't seen any evidence on AP Gaza building strike' (*AP*, 17 May 2021), <https://apnews.com/article/middle-east-israel-business-israel-palestinian-conflict-government-and-politics-abd641af1607fbae7f49e1cce7dbc49e> accessed 21 May 2021

[270] The +972 magazine has released reports on both Habsora and Lavender, see: Abraham (n 207). The IDF has published their own statement with regards to Habsora and Lavender, see: 'The IDF's Use of Data Technologies in Intelligence Processing' (*IDF*, 18 June 2024) <https://www.idf.il/en/mini-sites/idf-press-releases-regarding-the-hamas-israel-war/june-24-pr/the-idf-s-use-of-data-technologies-in-intelligence-processing/> accessed 10 July 2024

This historical interpretation of the facts, and their official legal justification of the strike, could have an impact on the programme's outputs (e.g., legal assessments of future attacks), despite how contested, in the public domain at least, the veracity of the facts is. Even if such systems' calculations would rely just partly on this historical data and combine it with dynamic data (i.e., data collected live from the battlefield), their training will already be shaped and influenced by data, the veracity of which can be contested thus affecting their reliability.

Of course, this is not to discount the discussions and debates that may have happened, behind the scenes, within the IDF while planning for this attack and assessing its legality. However, unless the nuances and contestations on the legality of the strike have been recorded, datafied, and that those layers of discussions are integrated into the AI programme in question, it would be limited to learning from and replicating the official justification provided by IDF on the attack. This ultimately leads to the question of who gets to label training data in the light of their influence over the programme's outputs.

Faulty source

As technological progress is proceeding apace, armed forces across the globe are growing increasingly dependent on advanced sensor capabilities and collection platforms. Sensing technologies are broad in range, from space-based assets (e.g., satellites to collect imagery), complex networks of radar stations (e.g., to detect the presence, movement and velocity of adversary missiles), to wearable devices and surveillance systems (e.g., cameras with thermal sensing).[271] This complex ecosystem of technologies can collect and provide tremendous amounts of raw data which, put together by AI-enabled programmes, can provide unprecedented insights through advanced data analytics. But with this opportunity also comes

[271] Military radars deployed back during the Cold War by the Soviet Union could, for example, detect planes and project their direction. See: Patricia Lewis and others, 'Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy' (2014) Chatham House <https://www.chathamhouse.org/2014/04/too-close-comfort-cases-near-nuclear-use-and-options-policy> accessed 15 November 2024

a risk: commanders' decision-making processes will indeed be heavily dependent and influenced by these data and thus, the sensing technologies providing these data must be reliable and unfaulty.

Faulty sensors will lead to faulty data. In fact, in circumstances where there are 'common faults in the sensors themselves or in the source of data' (e.g., faulty sensors because they have been damaged in the battlefield, or simply because they have aged), the data will not be as reliable.[272] The issue of faulty sensors is actually not new and exclusive to AI. In fact, concerns surrounding the sensors' effectiveness and reliability pre-date the age of AI with issues arising from the Cold War and even tracing back to World War II.[273] Faulty data can result from technical issues, but also from very 'simple' issues. For example, a smudge of mud on an optical sensor could interfere with its ability to detect variations in light energy and thus affecting the data it produces. In 1983, Soviet early warning satellites erroneously signalled an incoming attack of US nuclear missiles due to sunlight reflected from clouds, despite the fact that the satellites did not have any particular issues themselves.[274] Considering the 1980s were considered as the height of the Cold War where tensions between the US and the Soviet Union peaked, this incident shows the extent to which faulty data – even from perfectly functioning sensors – could have high repercussions, including of nuclear magnitude.

An added layer of attention to the issue of sensors' reliability and effectiveness has emerged over the past few years in the light of increased dependency on the 'internet of things'.[275] But

[272] Holland Michel (n 263)
[273] See Simon Bennett 'System Reliability: A Cold War Lesson' in Gitanjali Adlakha-Hutcheon, Anthony Masys, *Disruption, Ideation and Innovation for Defence and Security* (Springer 2022) <https://link.springer.com/chapter/10.1007/978-3-031-06636-8_2> accessed 15 November 2024 and Louis Brown 'Significant effects of radar on the Second World War' in Oskar Blumtritt, Hartmut Petzold and William Aspray (eds) *Tracking the History of Radar* (1994, Institute of Eletrical and Electronics Engineers, Inc.) <https://ethw.org/w/images/0/0e/Tracking_the_History_of_Radar.pdf> accessed 15 November 2024
[274] Lewis and others (n 271)
[275] See for example, Jan Chudzikiewicz, Janusz Furtak, Zbigniew Zielinski 'Fault-tolerant techniques for the Internet of Military Things' (IEEE 2nd World Forum on Internet of Things (WF-IoT) Milan, Italy, 2015) <https://ieeexplore.ieee.org/abstract/document/7389104> accessed 20 June 2023

the impact faulty sensors has further grown, considerably, in the light of AI technologies: Not only can faulty data affect key decision-making processes, but it can also bear negative influence over the performance of AI-enabled programmes relying on these data both for training and testing. In other words, the quality of training and testing data will also depend on their source – i.e., the sensors collecting the data in question.

Resilience against data poisoning and disinformation

The factual accuracy of the training data must be ensured against risks of data poisoning. Data poisoning specifically refers to the deliberate interference on training data to influence, or even gain some control over the model's learning process.[276] Data poisoning attacks can thus affect the model's outputs, their overall reliability and veracity. Additional nuance, however, must be conferred for models specifically designed and developed to perform on factually incorrect data, e.g., to generate disinformation campaigns. Once deployed, this would also entail the need for the programme to be resilient, in its performance, to adversarial attacks and potential misinformation campaigns deliberately launched by the adversary - or even non-deliberately due to its continuous exposure to false or inaccurate information.[277]

2.1.2. Precautions in attack, precautions in data

In the context of AI-enabled technologies for military targeting, their outputs bear extremely high stakes. Beyond considerations such as mission success and potential reputational damage both from the perspective of the contractor and the military forces in question, the life and death of combatants and civilians are also at stake – and, ultimately, compliance with

[276] Antonio Emanuele Cinà and others, 'Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning' (2023) 55(13) ACM Computing Surveys <https://dl.acm.org/doi/full/10.1145/3585385> accessed 10 July 2024

[277] An AI-enabled system's performance highly depends on the data it has been trained and tested with, but also on the data it is processing in real-life applications: it is not because a system was trained and tested with factually correct data that it will be resilient to 'untrue' information - especially when those machine learning systems are based on reinforcement learning, where they continuously learn, which subsequently affects how they perform in the long run. See for example: Karen Hao 'How Facebook got addicted to spreading misinformation' (*MIT Technology Review*, 11 March 2021) <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/> accessed 21 May 2021

international humanitarian law. Hence, it would be in everyone's interests – from developers to investors and end-users – to ensure that the data used for training and testing of data is of high quality given the tremendous influence it can exercise on the final outputs. In the following sections, we will reflect as to how there would be a legal obligation (beyond ethical and economical) to ensure such quality from the earlier stages of an AI technology's lifecycle. Specifically, we will explore the rule of precautions in attack as a legal basis for such obligations – and that these precautionary measures must extend beyond attack to the design, development and testing phases of the technology in question.

The IHL rule of precautions in attack and its temporal applicability

The rule of precautions in attack is of customary nature, applicable both in international and non-international armed conflict, and the essence of which is codified in Article 57 of the 1977 Additional Protocol I to the 1949 Geneva Conventions, as well as to a certain extent in Article 13(1) of the 1977 Additional Protocol II to the 1949 Geneva Conventions.[278] In essence, 'constant care must be taken to spare the civilian population, civilians and civilian objects. All feasible precautions must be taken to avoid, and in any event to minimise, incidental loss of civilian life, injury to civilians, and damage to civilian objects.'[279]

From a temporal perspective, Article 57(1) API does not provide a specific timeframe, other than 'in the conduct of military operations', as to when such 'constant care' should start exactly. However, the second part lists a series of precautionary measures to adopt 'with respect to attacks': in addition to the Article's title in itself, the first – obvious – assertion we can make is the applicability of these precautionary obligations to the attack in itself. In international

[278] Protocol additional to the Geneva Conventions of 12 August 1949, and relating to the protection of victims of international armed conflicts (Protocol I) (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 (AP I) art 13 and 57. For the ICRC, this requirement requires compliance with the need for precautions in attack - so as to confer this general protection to the civilian population and individual civilians. See: 'Rule 15: Principle of Precautions in Attack' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule15> accessed 21 May 2021

[279] Ibid

humanitarian law, an 'attack' is narrowly defined in Article 49(1) API specifically as 'acts of violence against the adversary, whether in offence or in defence.'[280] Extensive academic research has been specifically dedicated to interpreting and applying this definition in today's context, notably for cyber operations, looking at adjacent concepts such as 'violent consequences' and the value of 'danger' in these discussions.[281] Other scholars have argued that the definition provided by Article 49 is not enough and does not factor in the notion of an 'attack' in the broader, Geneva Law context.[282] Regardless of where one stands in this debate – a fascinating discussion that would however fall beyond the scope of this thesis – what is important to reiterate is the applicability of precautionary obligations during the conduct of an attack, including when they employ AI-enabled means and methods. It is argued here that precautionary measures must extend to the development stages of AI technologies for military targeting for a range of practical and legal reasons. From a purely technical and practical perspective, as established throughout this thesis, the design and development stages of AI technologies will shape in a significant and decisive manner their performance and decisions leading to their subsequent use.

In addition, Article 57(2) includes, in its scope, a series of obligations for 'those who plan or decide upon an attack'.[283] This inclusion of those at the planning and decision-making stages suggests that the rule of precautions extends beyond the attack in itself as narrowly defined, but also to the events and processes building up to the attack. The ICRC's 1987 Commentary would in fact argue that the term 'military operations' should be understood as enshrining 'any movements, manoeuvres and other activities whatsoever carried out by the armed forces with

[280] API, art 49(1)
[281] See for example Michael N Schmitt '"Attack" as a Term of Art in International Law: The Cyber Operations Context' (4th International Conference on Cyber Conflict, 2012) <https://ccdcoe.org/uploads/2012/01/5_2_Schmitt_AttackAsATermOfArt.pdf> accessed 26 June 2023
[282] Christof Heyns, Stuart Casey-Maslen, Thomas Probert 'The Definition of an "Attack" under the Law of Armed Conflict' (*Articles of War*, 3 November 2020) <https://lieber.westpoint.edu/definition-attack-law-of-armed-conflict-protection/> accessed 26 June 2023
[283] API, art 57(2)(a)

a view to combat'.[284] In other words, the obligation for precautionary measures is not only applicable to the events leading up to target engagement (e.g., deployment of weapon system in question; navigation towards, in and around the intended area of attack) but also to intelligence collection and planning. This is particularly supported by subsequent provisions in Article 57, extending the obligation to 'take all feasible precautions in the choice of means and methods of attack'.[285] In addition, Article 57 API further states that this precautionary duty extends to 'those who plan or decide upon an attack'; thus confirming the extension of such duty to the operational planning stage before the attack has even be set in motion. Hence, the commander and those involved in the decision-making process while planning the attack are bound by this duty to undertake the necessary precautions.

Beyond Article 57 API, military planning would generally also entail the identification and consideration of alternative pathways/courses of action, and subsequently the legal implications of each of those alternatives.[286] To this end, situational awareness will be critical. Awareness on the landscape, the demographics, the urban planning in and around the targeted area - both historical and live data on these information will be critical for parties to an armed conflict to plan their military operations and subsequent attacks. With the increasing assistance by – and even dependency on digital technologies to feed into this situational awareness (e.g., through live sensors present in the battlefield, live satellite imagery and surveillance footage from ISR drones), there is a growing need to use AI systems that will allow the collection and processing of all the data to feed into military decision-making in a short period of time.

[284] 'Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977: Commentary of 1987' (1987) ICRC <https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=D80D14D84BF36B92C12563CD00434FBD> accessed 21 May 2021

[285] API, art 57(2)(a)(ii)

[286] Yishai Beer 'Humanity Considerations Cannot Reduce War's Hazards Alone Revitalising the Concept of Military Necessity' (2015) 26(4) European Journal of International Law <https://academic.oup.com/ejil/article/26/4/801/2599602> accessed 21 May 2021

Such systems can even use all of the data collected and processed for situational awareness against more 'operational' data from the armed forces, e.g., a live count of available delivery methods, munitions, their calibre, etc. This will help planners and decision-makers to not only plan the military operation with the most adequate timing and location, but also help in the choice of the right means and methods of warfare against the set objective (e.g., choice of weapon system, how to conduct and launch the attack, what to do after the launch of the attack, etc.) Coming back to the example of Palantir's 'AI Platform for Defense', the software is capable of generating a set of courses of action based on a number of parameters, including the adversary's attack battalion; the assets and armament available and the time it would require to engage the target; distance to target; fuel level; personnel available and required; etc.[287] The software combines these considerations with live data from sensors stationed in and around the battlefield – providing for example a detailed terrain analysis with slope angles, passable areas, the appropriate vehicle type for such landscape, etc. The processing of live data will add in veracity and precision, adding to mission success in addition to a more comprehensive picture of what to target, or not. The integration of such functions – and a user-friendly interface allowing the commander to make decisions easily with clear information – can thus serve as a good example of measures enabling compliance with the rule of precautions in attack.

The use of such technologies rests on the assumption that the technology in question was designed and tested for legal compliance already. Arguably, this would mean that the obligation to comply with the rule of precautions in attack extends even beyond the operational planning stage of a military operation and its subsequent attacks: this rule also applies to the development, training and testing of the technologies that are expected to assist and be used by those involved in the planning and decision-making processes. If their judgment and decision-making, which will need to comply with the rule of precautions, will depend on the calculations

[287] 'AIP for Defense' (n 259)

and results provided by these technologies, it then means by extension that there is an obligation for these technologies to be developed, trained and tested with compliance with the rule of precautions in attack in mind.

Veracity of data as a precaution in attack

Building up on the previous section, we can establish a certain duty, by States, to mandate their developers and contractors in ensuring the veracity of data by virtue of the rule of precautions in attack. This is particularly due to the consequences of inaccurate data which could, potentially, lead towards the violation of international humanitarian law. Outdated datapoints on the location of military objectives, for example, can lead to the targeting of civilian buildings mistaken as lawful targets; and increasing automation and dependency on autonomous systems to engage targets with little (to no) oversight by humans can only exacerbate risks of such incidents happening. As such, if these risks can be prevented from the early stages of the technology's lifecycle, including by ensuring the veracity of its training data in the light of its decisive and influence over the system's output, a case can be made for such measures to constitute part of compliance efforts with the rule of precautions in attack.

One particular aspect, from the rule of precautions, worth dissecting further is the 'feasibility' dimension. As this rule prescribes the adoption of 'feasible' precautionary measures to minimize or, at best, avoid the incidental loss of civilian life, what that concretely means with regards to the developers' role in ensuring the veracity of training data must be examined. Many States would generally approach and interpret this concept of 'feasibility' in precautions as 'being limited to those precautions which are practical or practically possible, taking into account all circumstances ruling at the time, including humanitarian and military considerations.'[288] This interpretation crafts a careful balance between humanitarian

[288] 'Rule 15: Principle of Precautions in Attack' (n 278)

considerations on the one hand, and military necessity on the other hand – which, generally, serves as a key reference in all IHL rules to ensure that it protects while also taking into account realities in the battlefield.

In the light of the above, the rule of precautions would not only require that the States procuring such capabilities do everything 'practically possible' to foster compliance from the training and testing of the system, which includes mandating the veracity of its training datasets. In practice, this means for example, that the developers must adopt measures to clean the training datasets by verifying and filtering out errors, inconsistencies and inaccuracies. Once the data has been cleansed, a number of processes must be in place to maintain the veracity of the training datasets for the system's intended purpose and throughout its lifecycle.[289] In other words, data hygiene processes must be in place in order for the procuring State to comply with the rule of precautions in attack, in the light of the influence the training data's veracity can exercise over the system's outputs and subsequent ability to comply with the law. This assertion is particularly important in the light of the hefty costs and time-consuming nature of establishing these processes.[290] As such, it may be tempting for developers to reduce attention and resources dedicated towards data hygiene practices. Bypassing these processes, however, will not only have implications on the system's overall performance; this may also have legal ramifications for the procuring State with regards to the rule of precautions in attack – especially if these processes were deemed as 'feasible' but deliberately overlooked. This example demonstrates that while the responsibility of implementing IHL obligations do remain within the hands of States, this may have subsequent implications on third parties (e.g., developers) in specific contexts such as procurement where, further down the line, IHL will very much be relevant.

[289] Afina, Grand-Clément (n 260)
[290] Insights from a roundtable the author attended held under the Chatham House rule, as part of a United Nations Institute for Disarmament Research (UNIDIR)'s event held in March 2024 in Bellagio, Italy.

It is important to add, however, some nuance to our discussions. The obligation to undertake all feasible precautions is absolute; however there may be limitations as to what could be achieved especially under certain circumstances. There is thus the understanding that the veracity of training datasets is not expected to reach one hundred percent, and that ultimately the output would not be one hundred percent accurate. This is particularly important due to a number of limiting factors regarding training datasets for military targeting systems, including the dynamic and inherently messy nature of warfare, the sheer unavailability of data, or even due to the inability to ascertain the veracity of the data (e.g., due to technical limitations, inaccessibility, etc.) The assessment as to what is the minimum level of accuracy required will need to factor in these limiting factors, and more generally reflections on the relationship between probability and feasibility. As such, it may be argued that as long as the procuring State can demonstrate they have done everything they would reasonably be expected to do within their powers and spheres of responsibility, including mandating specific practices and measures to the developers and contractors in the procurement of AI-enabled capabilities, they are not necessarily in violation of the rule of precautions in attack.[291]

2.2. Diversity of data

2.2.1. What 'Diversity of Data' means and why it is important

A discussion on the need for diversity of data inevitably leads to establishing, in the first place, what 'diversity' means in the context of data and AI, and how this diversity can be measured – if it can, at all. One general way of defining 'diversity' of data is to expose a machine learning programme, in its training, 'with more diverse values of its training data', so that 'the trained

[291] As it has been said in their piece: 'An examination of feasibility is a question of reasonableness.' See: Jeffrey S. Thurnher 'Feasible Precautions in Attack and Autonomous Weapons' (2018) in Wolff Heintschel von Heinegg, Robert Frau, Tassilo Singer (eds) *Dehumanization of Warfare* (Springer, 2018) <https://doi.org/10.1007/978-3-319-67266-3_6> accessed 15 July 2022

model could cover more future unseen cases'.[292] In other words, 'diversity of data' would entail the exposure of the programme in training to a wider, diverse and rich set of data, variables and situations to which it would be expected to be exposed to and make decisions on once deployed. In the context of datasets pertaining to human subjects for example, this would translate into the following definition of diversity of data: 'variety in the representation of individuals in an instance or set of instances, with respect to sociopolitical power differentials (gender, race, etc.)'[293] This approach, coupled with other notions such as inclusivity, ultimately strive to decrease discriminatory biases in the programme's calculations and results.

The diversity of training and testing datasets (or the lack thereof) is generally shaped by the source, i.e., where this data is being collected from. The inherent limitations, such as biases, of the sources will be reflected in the data. This is an extremely important consideration for our discussions, given how the lack of diversity in training and testing datasets will inevitably shape the system's outputs. The latter can end up inappropriately biased and thus, inaccurate and harmful to mission success. This can have a direct impact on legal compliance and the life or death of combatant, other fighters and civilians on the ground, e.g., through false positives and inaccurate assessments over certain protected characteristics.[294]

For example, many machine learning-based facial recognition programmes use training data scraped from digital platforms (e.g., Flickr) to obtain samples of human face images.[295] These

[292] Hyontai Sug 'Performance of Machine Learning Algorithms and Diversity in Data' (MATEC Web of Conferences 210, GSCC, 2018) <https://www.matec-conferences.org/articles/matecconf/pdf/2018/69/matecconf_cscc2018_04019.pdf> accessed 5 March 2021

[293] Margaret Mitchell and others (2020), 'Diversity and Inclusion Metrics in Subset Selection' (2020) ArXiv <https://arxiv.org/pdf/2002.03256.pdf> accessed 5 March 2021

[294] For a detailed discussion on biases in machine learning models and their implications in international humanitarian law, see: Nema Milaninia 'Biases in machine learning models and big data analytics: The international criminal and humanitarian law implications' (2021) 102(913) International Review of the Red Cross, <https://international-review.icrc.org/articles/biases-machine-learning-big-data-analytics-ihl-implications-913> accessed 18 March 2021

[295] One of the instances most covered by the press was the use of Flickr without the consent of its users. See: Olivia Solon 'Facial recognition's 'dirty little secret': Millions of online photos scraped without consent' (*NBC News*, 12 March 2019) <https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921> accessed 17 June 2022. IBM for example is one of the companies that trained

datasets are inherently and inevitably limited: One of the limitations pertain to the fact that these data tend to only reflect a certain demographic composition.[296] Pictures scraped from Flickr would for example tend to represent mostly the demographics of Flickr users, i.e., of a certain age, from a handful of countries, and with access to the internet. The pictures uploaded on the platform are bound to reflect the demographic reality of those directly connected to these users. The identification of these limitations on representation of this dataset, i.e., the biases in the dataset and their subsequent limitations, must thus be taken into account when it is used to develop facial recognition technologies – and addressed accordingly. If the population represented in the training and testing dataset is primarily white/Caucasian, and if the system has only been exposed to that particular dataset, it would be inadequate to deploy and use that same system in areas where the primary subjects would not be white/Caucasian because the outputs may not be accurate. The limitations that come from this training dataset would render the programme inadequate for deployment and use in another geographic region where the demographic reality is significantly different from those represented in those training and testing datasets; unless the programmer compensates and addresses this limitation accordingly.

Some have proposed, as workaround, the scaling of models and training datasets: the more data they are trained on, the less likely they would be prone to biases.[297] In practice, there is a growing amount of research indicating the opposite. A recent study, for example, shows that as the training data increased, misclassification rates of human offensive classes such as criminal records increased.[298] Of the fourteen visio-linguistic models evaluated in this study,

---

their facial recognition programme on Flickr, See: 'IBM used Flickr photos for facial-recognition project' (*BBC*, 13 March 2019) <https://www.bbc.co.uk/news/technology-47555216> accessed 21 May 2021

[296] Eun Seo Jo, Timnit Gebru 'Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning' (Conference on Fairness, Accountability and Transparency (FAT*), 27-30 January 2020) <https://dl.acm.org/doi/pdf/10.1145/3351095.3372829> accessed 17 June 2022

[297] Hoan Ton-That 'The Myth of Facial Recognition Bias' (*ClearviewAI*, 28 November 2022) <https://www.clearview.ai/post/the-myth-of-facial-recognition-bias> accessed 17 November 2024

[298] Abeba Birhane and others, 'The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models' (FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency 2024) <https://dl.acm.org/doi/abs/10.1145/3630106.3658968à> accessed 17 November 2024

'the probability of predicting an image of a Black man and a Latino man as criminal increases by 65% and 69%, respectively'.[299] This does not necessarily mean that scaling the volume of training datasets is not important, nor necessary; it is just critical that these issues are inherently complex, and (novel) solutions are not always necessarily immune to them.

Additionally, this last example also shows that diversity of data is one of the concrete examples where the quality and quantity of data are intrinsically linked and inter-dependent. The previous example demonstrated that not only it is important for datasets to be representative and reflective of the reality; it is also about having enough different datapoints in order to obtain a fuller and thus, more informed picture. For example, weather forecasting requires a number of datapoints, from the temperature to atmospheric pressure and variables, humidity levels, cloud formation, as well as speed and direction of wind. A combination of these datapoints, all diverse in nature and at scale, enables today's models to be incredibly accurate – including Google DeepMind's GraphCast, capable of predicting with high levels of accuracy the weather for up to 10 days.[300] Similarly, in the military domain, if a system is designed to identify a command centre, a combination of diverse datapoints (e.g., satellite imagery, electronic signature, human intelligence reports, communication patterns etc.) will provide a fuller and more informed picture on the structure in question and whether it would, indeed, constitute a military objective or not.

2.2.2. Relevant military applications

There are a number of AI applications designed to support military operations where the diversity (or the lack thereof) of training data will have a significant impact on the final output. Such applications include those for military targeting, including for intelligence collection and processing, target identification, as well as target engagement. In this sense, it is important to

[299] Ibid
[300] Remi Lam and others, 'Learning skillful medium-range global weather forecasting' (2023) 382(6677) Science <https://www.science.org/stoken/author-tokens/ST-1550/full> accessed 17 July 2024

unpack some of the specific applications where the diversity of training and testing datasets is critical – not only for mission success but, ultimately, for legal compliance too.

First, there is strong appetite from armed forces around the world to adopt facial recognition technologies to enable the identification and selection of targets. The US DoD, for example, has awarded, in 2022, RealNetworks with over USD $700,000 to provide facial recognition capabilities for autonomous unmanned aircraft systems (i.e., drones), with the view of using this function for 'special ops, ISR, and other expeditionary use-cases.'[301] In the light of our discussions earlier, it is clear that diversity in the training and testing data (or the lack thereof) will have profound implications on how such facial recognition capabilities will perform and their final outputs.

Training and testing datasets lacking diversity may for example not account for the demographic realities in which the technology is intended for deployment in. Or, the lack of diversity may stem from the lack of training data accounting for all possible scenarios in which the facial recognition technology would be deployed in (e.g., with different lightings). In fact, the United States' Army Research Laboratory is reported to have used a dataset 'containing 500,000 images from 395 people' for the training of a programme designed to identify faces 'under both ordinary visible-light conditions and with heat-sensing thermal cameras in low-light conditions.'[302] In order for such technology to work, this would require a diverse set of training data that would be representative of all possible traits of different faces both during daytime but also in the dark, as the (lack of) light can considerably change the relevant features of people's faces.

[301] 'Implement Face Recognition on Autonomous sUAS for Identification and Intelligence-Gathering' Solicitation Number X21.1, Contract FA864922P0006 (*Department of Defense*) <https://www.sbir.gov/node/2217727> accessed 27 June 2023

[302] Jeremy Hsu 'Army Trains AI to Identify Faces in the Dark: The U.S. Army Research Laboratory has developed a dataset of faces to train facial recognition that works in darkness' (*IEEE Spectrum*, 9 March 2021) <https://spectrum.ieee.org/army-trains-ai-to-identify-faces-in-the-dark#toggle-gdpr> accessed 17 June 2022

More recently, Clearview AI has reportedly provided free access to the Ukrainian Ministry of Defence with facial recognition capabilities to support their operations – including to 'identify Russian operatives', although the exact purpose and extent to which Ukraine is actually using these technologies remains unclear.[303] The programme underpinning this technology must have been exposed to training datasets representative of the demographic realities representing Russian operatives, which obviously would be different from those for example of Western operatives. It has been reported that Clearview attempted to address this issue by scraping over 2 billion images from VKontakte, a Russian social media platform – which, in principle, would enable Clearview AI to ensure its training datasets were diverse enough for it to work reliably in the context of the Russo-Ukrainian war.[304]

While these facial recognition technologies would probably be deployed more in the context of dynamic targeting in the battlefield, another relevant application relates to intelligence collection and forming to inform the commander's decision-making. Israel's security agency, Shin Bet, is reportedly using generative AI to detect and foil threats to national security; the programme is said to 'flag anomalies in surveillance data and sorting through "endless" intelligence', in addition to help feed into decision-making processes.[305] While Shin Bet is reportedly using this programme for law enforcement, it does not mean that IHL does not necessarily apply especially in a context of occupation. In occupied territories for example, AI can be used to enable mass surveillance – as Israel reportedly does in the occupied West Bank.[306] AI also enables ATR capabilities, typically to detect and identify targets from vast

[303] Paresh Dave, Jeffrey Dastin 'Exclusive: Ukraine has started using Clearview AI's facial recognition during war' (*Reuters*, 14 March 2022) <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearview-ais-facial-recognition-during-war-2022-03-13/> accessed 17 June 2022

[304] Paresh Dave 'Exclusive: Ukraine has started using Clearview AI's facial recognition during war' (*Reuters*, 14 March 2022) <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearview-ais-facial-recognition-during-war-2022-03-13/> accessed 27 June 2023

[305] 'Israel's Shin Bet spy service uses generative AI to thwart threats' (*Reuters*, 27 June 2023) <https://www.reuters.com/technology/israels-shin-bet-spy-service-uses-generative-ai-thwart-threats-2023-06-27/> accessed 28 June 2023

[306] See: 'Automated Apartheid: How Facial Recognition Fragments, Segregates and Controls Palestinians in the OPT' (2023) Amnesty International, <https://www.amnesty.org/en/documents/mde15/6701/2023/en/> accessed

amounts of imagery obtained from a wide range of sensors, including satellites, semi-active radars, and TV cameras.[307] These programmes can be used in a centralised way – where vast amounts of data are collected from many sensors and aggregated by one same programme; or AI can also be integrated into systems such as drones specifically deployed in the battlefield for intelligence, surveillance, target acquisition and reconnaissance (ISTAR).[308] It is critical that these AI applications have been trained and tested using datasets that are adequately diverse and will ensure appropriate representation of realities in the areas they are meant to be deployed in. For instance, on the latter example of AI integration into ISTAR drones – if these drones are meant to be deployed in areas where civilian Bedouin tents are commonly found, the software must have been trained with enough data to make sure that these structures are not misidentified as makeshift military command centres and thus, potential targets.

Finally, another AI application under development – if not deployed already – in armed conflict situations is predictive analytics. In other words, these are AI-enabled tools used to predict specific elements such as human behaviour, individual membership to certain groups, and trends within specific groups of individuals. Again, for such applications, diversity in the training and testing datasets are critical. Predicted patterns are often, if not all, heavily influenced by cultural norms and realities that, if not being taken into account and contextualized against a wider set of data, can lead to misinterpretations and, consequently, the misidentification of a target. For example, such technologies could be used to 'map and understand recurrent conflict patterns and forecast potential crises.[309] Such programmes would

29 June 2023, and Adam Satariano and Paul Mozur 'Facial Recognition Powers 'Automated Apartheid' in Israel, Report Says' (*The New York Times*, 1 May 2023) <https://www.nytimes.com/2023/05/01/technology/israel-palestine-facial-recognition.html> accessed 29 June 2023

[307] Gary P. Beckett 'Leveraging Artificial Intelligence and Automatic Target Recognition to Accelerate Deliberate Targeting' (Thesis, Air War College, Air University 2020) <https://apps.dtic.mil/sti/pdfs/AD1107486.pdf> accessed 29 June 2023

[308] Mark Palmer 'AI's crucial role onboard drones for ISR capability improvement' (*Shield AI*, September 2022) <https://sentientvision.com/ais-crucial-role-onboard-drones-for-isr-capability-improvement/> accessed 29 June 2023

[309] Eleonore Pauwels 'Artificial Intelligence and Data Capture Technologies in Violence and Conflict Prevention: Opportunities and Challenges for the International Community' (2020) Global Center on Cooperative Security

generate predictions by taking into account a diverse set of data points, including parties involved in said conflicts, crowdedness, the presence of arms, patterns of violence, means and methods of warfare employed, etc. Such predictions would allow armed forces to anticipate areas of conflict, identify potential changes in the adversary's strategy on the battlefield, and plan accordingly to maintain advantage.[310] Yet, without a comprehensive and diverse training dataset reflecting local customs and realities on the ground, such predictions can be erroneous and potentially lead to unlawful targeting. Certain tribes in Saudi Arabia and Yemen, for example, shoot gunfire on celebratory occasions such as weddings or political victory.[311] The concentration of a crowd, the bearing of firearms and gunfire could easily be mistaken, by the programme, as the beginning of an outbreak of violence; paving the way for potential targeting because the training datasets were not diverse enough to factor in the possibility that certain tribes see firearms as exceeding sheer instruments of violence.

On a micro level, these tools could be used on individuals in the battlefield with the aim to predict their behaviour and anticipate accordingly. This can be done, for example, to predict the status of an individual in the battlefield (i.e., a lawful target – as a combatant or other targetable fighter, or an unlawful target – as a civilian, or a soldier *hors de combat*) based on sentiment analysis. The latter can be done, for example, through image recognition by analysing the individual's body language and their facial expression; speech analytics can also be used to record, translate if needs be, analyse and confirm the output, e.g., by expressing

<https://www.globalcenter.org/wp-content/uploads/2020/10/GCCS_AIData_PB_H.pdf> accessed 26 November 2020

[310] Joe Saballa 'US Army Seeking AI System That Predicts Enemy Actions' (*The Defense Post*, 11 July 2023) <https://www.thedefensepost.com/2023/07/11/us-army-ai-system/> accessed 14 July 2023

[311] Marie-Christine Heinze 'On 'Gun Culture' and 'Civil Statehood' in Yemen' (2014) 4(1) Journal of Arabian Studies <https://www.tandfonline.com/doi/full/10.1080/21534764.2014.920190?needAccess=true> accessed 5 July 2023

intent to surrender.[312] Once confirmed as surrendering, they would be considered as *hors de combat* and thus benefit from protection under IHL.

The reliability of such predictive technologies based on the human behaviour is, in itself, controversial and highly contested. In fact, beyond military applications, more generally, these computer vision-based technologies designed to predict and identify certain behaviours and emotions are being developed and deployed as surveillance and assessment means in a number of contexts, ranging from employment to schools.[313] But many argue that these technologies are not reliable and some go as far as seeing such technologies as inherently unethical to develop and deploy. While such discussions would fall outside of the scope of this thesis, it would be useful to note at this stage that the diversity of training and testing datasets also plays an important role for the performance of the programme.[314] The lack of diversity can indeed affect the detection of certain emotions (again, based on the assumption that these technologies can be reliable); thus shaping certain decision-making processes that may partly be based on a subject's emotions and intent such as surrendering.

2.2.3. Implications on the rule of distinction

The rule of distinction, of customary nature, dictates who and what can and cannot be targeted in an armed conflict situation.[315] This rule is crystallised in the 1977 Additional Protocol I to the 1949 Geneva Conventions, precisely in Articles 48, 51(2) and 52(2). Furthermore, the ICJ has specifically consecrated the rule of distinction as a 'cardinal principle' – in fact, the ICJ

[312] For example, studies have been undertaken to develop and test behavioural analytics to detect users in need of mental health care through an app analysing voice recordings and phone usage – a model that has been backed by DARPA. A similar voice-based analytics tool, combined with machine translation, could be used to detect a combatant's intent to surrender in the battlefield, for example.
[313] China in particular is known to be more aggressive in the development and deployment of such technologies without scrutiny measures in place – see Nikki Sun 'Workplace AI in China' (2024) Chatham House <https://www.chathamhouse.org/2024/07/workplace-ai-china/about-author> accessed 22 November 2024
[314] De'Aira Bryant, Ayanna Howard 'A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity' (AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019) <https://dl.acm.org/doi/abs/10.1145/3306618.3314284> accessed 5 July 2023
[315] 'Principle of Distinction' (*ICRC*) <https://casebook.icrc.org/law/principle-distinction> accessed 27 November 2020

has stated that 'States must never make civilians the object of attack and must consequently never use weapons that are incapable of distinguishing between civilian and military targets.'[316] The latter, adjacent rule the Court refers to consists of the prohibition of indiscriminate attacks and, subsequently, the prohibition to use means and methods of warfare that would be indiscriminate in their effects. In fact, referring to nuclear weapons, the Court has described them as incapable of drawing 'any distinction between the civilian population and combatants, or between civilian objects and military objectives, and their effects, largely uncontrollable, could not be restricted, either in time or in space, to lawful military targets.'[317] In other words, parties to an armed conflict must ensure that the effects resulting from any attack would be restricted in time and space to those lawful military targets.

As discussed in the previous couple of sections, diversity of data (or the lack thereof) plays a key role in shaping the final outputs and can bear great influence over critical decision-making processes related to military targeting. These processes being key to implementing the rule of distinction, we can therefore ascertain that diversity in the training and testing datasets are critical for legal compliance. In this sense, there are two particular considerations for us to discuss.

First, as discussed in the previous section, training and testing datasets must be adequate and representative enough of the socio-cultural realities surrounding not only the adversary but also the overall environment of operations. This diversity in training data is indeed critical, as ascertained earlier, to ensure the veracity of outputs generated by programmes and, ultimately, inform decision-making in compliance with the law. Coming back to our example of a combatant deciding to surrender (and thus, be granted *hors de combat* protection). AI-enabled programmes designed and deployed to differentiate targetable from non-targetable individuals

[316] *Legality of the Threat or Use of Nuclear Weapons* (n 69) [78]
[317] Ibid [92]

must be dynamic – as in, they must be able to change their assessment depending on the circumstances and thus, be capable of identifying for example a combatant expressing the intent to surrender and change their status from targetable to non-targetable. This could include computer vision capabilities to analyse the combatant's body language, and/or speech analytics to recognise words expressing surrender. These programmes must have been trained and tested with training datasets that not only include gestures and speech assumed to be universal for surrendering (e.g., putting one's hands up; abandoning firearms; white flag). They must also have taken into account local jargons, sayings and local 'ways' of surrendering – in other words, cultural and sociolinguistic nuance from the ground.[318] Failure to incorporate these into the training of the programmes will inevitably result in misidentifying a surrendering combatant as such, which could potentially lead to their unlawful targeting – as the rule of distinction prohibits the deliberate targeting of combatants attempting to surrender.

Second, diversity of data is essential to represent the demographic realities of the environments in which the programme is expected to operate. Failure to account for appropriate demographic representation in the training data sample can lead to faulty outputs which, in turn, would have been expected to inform key decision-making processes in military targeting. For example, Face++ and Microsoft AI have both developed facial recognition programmes aimed at analysing and predicting emotions based on facial traits, and subsequently help with decision-making processes.[319] Yet, in addition to sentiment analysis generally being a highly controversial field, as discussed earlier, a study conducted on these exact programmes revealed that both exhibited racial biases in their outputs: both would indeed tend to interpret black

[318] A. Stevie Bergman, Mona T. Diab 'Towards Responsible Natural Language Annotation for the Varieties of Arabic' (ACL Findings 2022) <https://arxiv.org/abs/2203.09597> accessed 7 July 2023

[319] Although Microsoft has decided, in June 2022, to stop developing facial recognition and analysis technologies in response to concerns, notably surrounding biases and unreliability. See: Kashmir Hill 'Microsoft Plans to Eliminate Face Analysis Tools in Push for 'Responsible A.I.' (*The New York Times*, 21 June 2022) <https://www.nytimes.com/2022/06/21/technology/microsoft-facial-recognition.html> accessed 1 July 2022

subjects as exhibiting more negative emotions (e.g., anger or contempt) than white subjects.[320] Such faulty and problematic outputs generally result from the programmes inheriting limitations in the training datasets reflecting historical racial biases.

Taking back the example from the first consideration of AI use to identify targetable combatants: The US has expressed interest in developing such technologies through, for example, their Urban Reconnaissance through Supervised Autonomy (URSA) project. The latter is said to develop autonomous systems to help detect hostile forces through 'human behaviors, autonomy algorithms, integrated sensors, multiple sensor modalities, and measurable human responses to discriminate the subtle differences between hostile and innocent people.'[321] In other words, the programme is designed to differentiate targetable from non-targetable individuals based on sentiment analysis. If the programme faces the same limitations as those developed by Face++ and Microsoft AI, combatants of colour will be more at risk of being falsely identified as exhibiting negative emotions and thus, subject to direct targeting when, in fact, they may be surrendering – or simply not targetable at that moment. But because the programme would be more prone to associating them with hostile/negative emotions, they may not be recognised as protected – thus increasing risks of violations of the rule of distinction where they could end up being targeted due to negligence from the developers' side to expose the programme to a diverse set of training data.

2.3. Proxy data

2.3.1. What 'proxy data' means and its importance

In a nutshell, proxy data, alternatively referred to as proxy variables or 'inferences', do not hold explicitly, and directly, the information needed to answer the question the model is tasked

[320] Lauren Rhue 'Racial Influence on Automated Perceptions of Emotions' (2018) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765> accessed 27 November 2020

[321] Bartlett Russell 'Urban Reconnaissance through Supervised Autonomy (URSA)' (*Defense Advanced Research Projects Agency*) <https://www.darpa.mil/program/urban-reconnaissance-through-supervised-autonomy> (accessed 1 December 2020)

to examine, but rather indicators that indirectly provide the information needed.[322] Proxy data are generally used not only to complement direct indicators where proxy data adds further, secondary context. In many instances, the data will simply be unavailable and thus, proxy data is the only way to obtain certain information. Big data solutions pave the way to harness proxy data and identify critical correlative patterns that otherwise would be unattainable. In this sense, proxy data are a natural occurrence in a big data context, and constitute an inherent and critical part of data analytics.

For example, a programme can infer someone's age range based on large datasets that, albeit not directly indicating these individuals' age, tap into an ecosystem of proxy indicators which, put together, could possibly indicate said age range. These proxy indicators could, for example, be the person's address, education, smoking and drinking habits, spending patterns, credit score, and pet ownership. Another example could relate to measuring living standards in developing countries. Due to the limited number (or simply lack) of data from demographic surveys in these countries on household incomes or consumption expenditures, researchers aiming to assess living standards in these countries 'have had little alternative but to rely on simple proxy indicators', with the hope that these set of indicators will, taken together, somehow provide some insights on living standards.[323] In this context, proxy variables could for instance be access to clean water, certain household items (e.g., refrigerator, radio, TV), toilet facility, but also for example children's schooling, fertility and child mortality rates - which, put together, would provide insights on a country's living standards and compensate for

[322] George Čevora 'How Discrimination occurs in Data Analytics and Machine Learning: Proxy Variables' (*Towards Data Science*, 5 February 2020) <https://towardsdatascience.com/how-discrimination-occurs-in-data-analytics-and-machine-learning-proxy-variables-7c22ff20792> accessed 16 April 2021

[323] Mark R. Montgomery and others 'Measuring Living Standards with Proxy Variables' (2000) 37(2) Demography <https://link.springer.com/article/10.2307/2648118> accessed 16 April 2021

the lack of data directly pertaining to consumption expenditure, the generally preferred metric used to measure a country's living standards.[324]

Proxy data may also be used as a tool to help overcome potential biases associated with certain indicators. An algorithm developed to assess and evaluate an organisation's employees' performance at work may, for example, be trained to use more 'impartial' indicators such as the level of outreach and speaking engagements and the number published outputs, as opposed to indicators that may be more prone to obvious biases such as the employees' level of education and gender (e.g., a predictive system could be biased into correlating men and/or those with higher levels of education, with higher performance without taking into account privileges and other, adjacent datapoints) – although it is arguable that no indicator is truly neutral/would not be subject to potentially harmful biases: working patterns for example may be affected by someone's need for commuting, lack of childcare, etc.[325]

2.3.2. Relevant military applications

Proxy data plays a critical role in informing and shaping decision-making processes in the military space. In fact, due to the non-straightforward nature of warfare, proxy indicators are sometimes the only available solution to ascertaining certain critical information such as, for example, someone's membership to a non-state armed group (NSAG) and thus, their status as to whether or not they'd potentially be targetable, i.e., whether they may be considered as directly participating in hostilities.[326] There is, in fact, a growing body of research dedicated to identifying 'predispositions' of members of NSAGs based on the socio-cultural context, political landscape, and relationship networks surrounding them – in other words, proxy indicators of their membership to certain armed groups.

---

[324] Ibid

[325] Andrew Burt 'How to Fight Discrimination in AI' (*Harvard Business Review*, 28 August 2020) <https://hbr.org/2020/08/how-to-fight-discrimination-in-ai> accessed 23 April 2021

[326] For a more detailed discussion on direct participation in hostilities, see later in Chapter 2, Section 3.1.2. and Chapter 4, Section 3.3

For example, interviews have been conducted to identify different motivators for joining non-state armed groups in Colombia, including individual senses of social justice, security and validation; political affiliation; and political climate.[327] Research has also been conducted to draw the sociological and operational profile of members of the Provisional Irish Republican Army (IRA) where they identified 39 variables across 30 years, including age, location, roles, employment status, and factored into the analysis Britain's changing counter-terrorism policy and changes in environmental conditions.[328] These studies provide a wide, diverse range of data that, once aggregated, can provide comprehensive and invaluable intelligence on members of the NSAGs in question and identify proxy indicators that would allude to their membership. Taking the example of profiling IRA members, their identification can be based on the abovementioned indicators which, in themselves, do not necessarily mean they are members of the IRA but, taken together, may indicate their membership to a certain degree of probability – and it may be that using these proxy data is one of the only ways of identifying them.

Beyond benefitting from such knowledge on the strategic level (i.e., to devise strategies based on one's understanding of the adversary), such body of research can be used to feed into the development of programmes tasked with their identification in conflict. In fact, proxy data constitutes a key enabler in intelligence collection, processing and building – and big data solutions will only increase their potential to feed into decision-making processes.[329] One prominent example of such proxy data analytics is the NSA's Skynet surveillance programme, which does exactly this. The programme collects metadata from mobile phones in a specific, set area, more specifically location and messaging logs. Skynet will then, based on

[327] Mauricio Florez-Morris 'Joining Guerrilla Groups in Colombia: Individual Motivations and Processes for Entering a Violent Organisation' (2007) 30(7) Studies in Conflict & Terrorism <https://www.tandfonline.com/doi/abs/10.1080/10576100701385958> accessed 3 February 2022

[328] Paul Gill & John Horgan 'Who Were the Volunteers? The Shifting Sociological and Operational Profile of 1240 Provisional Irish Republican Army Members' (2013) 25(3) Terrorism and Political Violence < https://www.tandfonline.com/doi/full/10.1080/09546553.2012.664587> accessed 3 February 2022

[329] Ashley S. Deeks, 'Predicting Enemies' (2018) 104(8) Virginia Law Review <https://www.jstor.org/stable/26790717> accessed 17 July 2024

commonalities and patterns in the data, draw inferences as to whether someone's activities would indeed indicate their membership to Al-Qaeda, a non-state armed group, more precisely as a courier. Taken out of context, location and messaging logs in themselves are not, in themselves, directly indicative of membership to Al-Qaeda. However, given the secretive nature of their work, it would be practically impossible to obtain any direct indicators that would identify the couriers as such. Hence, proxy indicators may very well be the only way to identify these couriers – thus making these data central to intelligence collection and analysis.

In addition to intelligence collection and processing, proxy data can also be used for AI systems integrated into weapons and other means and methods of engaging with targets. This is for instance the case of integrating autonomous capabilities in weapons systems for navigation and target identification. These systems will need to be trained and tested in environments similar to those where they are expected to be deployed (e.g., systems meant to be deployed in the desert must be trained and tested for such environments). Yet, despite the wealth of data being generated and available today, sometimes there is not enough data available to re-create realistic training and testing environments where all variables will be accounted for. This insufficiency can be due to a number of reasons, including the sheer absence of relevant data, and/or the lack of access to such data, and/or because such data would be expensive, to cite a few. This gap in data can be addressed by using synthetic data, i.e., data that has been generated artificially and not collected from real-world events through sensors.[330] However, the use of synthetic data could be met with resistance for a number of (valid) reasons, such as the sheer reliability of synthetic data. The US DoD's Joint Artificial Intelligence Center (JAIC), now integrated within the Chief Digital and Artificial Intelligence Office (CDAO), is known to not use synthetic data and only use operational data for testing.[331] Yet again, there is only so much

[330] Sergey I. Nikolenko, *Synthetic Data for Deep Learning* (Springer, 2022) 4
[331] This information has been obtained during a table top exercise the author attended in April 2021, held under the Chatham House rule; thus the author may use it without however attributing this information.

operational data it can access to, which could potentially lead them into a data shortage issue. As such, an alternative to the absence of 'direct' data is the use of proxy data that, taken together, will enable the re-creation of realistic testing environments.

2.3.3. Proxy data and their legal implications

(Il)legality of proxy indicators?

The first key question that needs addressing relates to whether the use of proxy data is, in itself, unlawful. The present section will particularly look at compliance with the rule of distinction: this choice is motivated by the fact that, in the author's view, proxy data use is most central for intelligence collection and processing which, in turn, will feed into decision-making processes surrounding the identification of targets. As such, these processes will be a key determinant to the subsequent legal assessment as to whether or not the identified target is, indeed, a lawful target. Due to the inherent, profound impact proxy data use will have on distinction assessments, this section will therefore focus exclusively on the latter and not other IHL rules. In addition, it is important to note at this stage that our discussions will not pertain to compliance vis-à-vis the proxy data in and of themselves; but rather, the ability for programmes to use proxy data. In the end, if a proxy data-based programme is able to perform in compliance with the law, there should be no compliance issues with their use *per se*.

Human Rights Watch (HRW) has argued, in the early ages of military AI discussions, that 'fully autonomous weapons' 'would not be able to make such a determination when combatants are not identifiable by physical markings' because of inadequate sensors, the adversaries' ability to 'trick these robots', and these weapons systems 'would not possess human qualities necessary to assess an individual's intentions, an assessment that is key to distinguishing targets'.[332] In other words, fully autonomous weapons would not be capable of being used in

[332] Human Rights Watch (n 75)

compliance with the rule of distinction. Yet, if anything, proxy data may constitute the key for compliance with the rule of distinction, addressing the abovementioned concerns raised by HRW.

First, physical markings are not the only way to ascertain a combatant's status as such. Of course, these markings would constitute obvious indicators of somebody's status such as, for example, uniforms. But, as established earlier, there are other ways of establishing – or at least, inferring – somebody's status through proxy indicators. As discussed in the previous section, the assessment of an individual's status will depend on a host of variables, both direct and indirect, and the more datapoints there are more likely it will be accurate or at least, able to capture the bigger picture and context. In the intelligence collection and planning stages, this could be done for example by tapping into socio-cultural indicators that could, potentially, increase the probability of somebody's belonging to certain NSAGs: In Latin America, for example, States generally perceive the likelihood of someone's belonging to an NSAG as heightened if they fall below the threshold of poverty.[333] In the battlefield, as controversial as such solutions can be, computer vision programmes could be used to detect hostile intent and, subsequently, contribute at least partly to assessments as to whether or not the subject in question is indeed targetable.[334] It is however important to note, at this stage, that the extent to which intent would even be considered for identifying civilians directly participating in hostilities as such remains generally subject to much debate. In fact, most commentators would argue that intent does not affect the targetability of civilians directly participating in

[333] This insight is from track-1 regional consultations held under the Chatham House rule the author has attended in Santiago, in June 2024, on responsible AI in the military and wider security domains.
[334] Although, again, whether AI-enabled programmes to measure and/or predict one's intention is technically feasible, reliable and ethical, or not, is another question that would fall beyond the scope of this thesis. For an example of studies on the prediction of human intent, see: Steven Holtzen and others 'Inferring human intent from video by sampling hierarchical plans' (IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2016) <https://ieeexplore.ieee.org/abstract/document/7759242> accessed 4 February 2022

hostilities.[335] The ICRC is generally aligned with this view by centring its assessment on the objective purpose of the act in question, noting that 'the subjective motives driving a civilian to carry out a specific act cannot be reliably determined during the conduct of military operations and, therefore, cannot serve as a clear and operable criterion for "split second" targeting decisions'.[336] Interestingly, and more recently, the ICRC has raised the issue of 'intent' in its position on autonomous weapons systems, specifically noting that 'effectively protecting combatants/fighters who are placed *hors de combat* and civilians who are not, or no longer, taking a direct part in hostilities calls for difficult and highly contextual, conduct-, intent- and causality-related legal assessments by humans in the context of a specific attack'.[337]

Regardless of where one stands in the debate with regards to intent, establishing the status of a fighter, i.e., as to whether they are indeed a civilian directly participating in hostilities, is the very opposite of a straightforward assessment and is highly contextual. In fact, it would be useful to remind ourselves that, ultimately, the rule of distinction is not exclusively based on markings uniforms. Members of a state's armed forces, despite their uniform, can still be protected from direct attack under the rule of distinction if, say, they are *hors de combat*.. As such, if anything, proxy data use constitutes an inherent part of distinction assessments and is therefore not unlawful in itself.

In addition, we could argue that proxy data cannot be separated from direct indicators – especially in armed conflict where, as established previously, facts and information are not as

[335] See for example Kevin Jon Heller, 'The Concept of 'the Human' in the Critique of Autonomous weapons' (2023) 15 Harvard National Security Journal <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342529> accessed 3 June 2025

[336] Melzer 'Interpretative Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law' (2009) ICRC <https://shop.icrc.org/interpretive-guidance-on-the-notion-of-direct-participation-in-hostilities-under-international-humanitarian-law-pdf-en.html> accessed 13 July 2022, footnote 150. Although an in-depth discussion on this fascinating issue would fall beyond the scope of this thesis, the Interpretative Guidance provides for a much more detailed discussion on direct participation in hostilities including the question of intent.

[337] ICRC, 'International Committee of the Red Cross (ICRC) position on autonomous weapon systems' (2022) IRRC No. 915, <https://international-review.icrc.org/articles/icrc-position-on-autonomous-weapon-systems-icrc-position-and-background-paper-915> accessed 5 June 2025

straightforward. As such, proxy data is a natural part of the ecosystem of intelligence collected and needed for military decision-making; and not using these data could actually be detrimental not only for mission success but also for legal compliance. Coming back to the statement made by HRW, the organisation further argued that machines are inherently incapable of assessing one's intentions. As HRW pointed out, intention can be one of the key, deciding factors as to whether someone is targetable or not – echoing the point just made in the previous paragraph. Combatants may indeed intend to surrender and by actively doing so, they are *hors de combat* and become protected under the rule of distinction.. But a combatant's intent to surrender may not be enough to establish their status as *hors de combat*: for example, the ICRC defines surrender as a 'unilateral act', thus implying the need for the active act of surrendering (e.g., with physical actions by putting their hands up, throwing away their weapons, etc.) and thus, the sole mental element of intent would not suffice.[338] This example of intent shows that, whether or not an indicator is indeed 'direct' or 'proxy' is often not as black or white. And in the context of big data – and intelligence collection and processing more generally, again, proxy data is a natural occurrence and a critical component. Thus, its use must be treated similarly as direct data, and as such cannot be determined as inherently unlawful.

Proxy indicators: A legal, precautionary necessity?

It is arguable that integrating proxy indicators into training and testing datasets are necessary for compliance with the rule of precautions. As established in prior sections, developers will have a decisive role in ascertaining whether a State has implemented and complied with its obligations to undertake all 'feasible' precautions to minimize (or, at best, to avoid) incidental loss of civilian life, injury to civilians, and damage to civilian objects – and this includes through the adoption of the appropriate technical solutions that would decrease such risks. This section will argue that proxy data constitutes a key solution to decreasing risks of adversarial

[338] 'Surrender' (*ICRC*) <https://casebook.icrc.org/glossary/surrender> accessed 13 July 2022

attacks; and their integration may very well constitute a legal obligation in the spirit of the rule of precautions.

As many scholars have pointed out, AI-enabled technologies for military targeting are not immune to adversarial attacks, i.e., deliberate disruptions to an algorithm with the aim to change its final outputs.[339] This concern was also expressed by HRW, as seen in the previous section, with the ability of the adversaries to 'trick these robots' by interfering with their sensors. This is a problem the AI field grapples with even beyond the military realm. Perhaps the most notorious documented example of adversarial disruption on AI-enabled systems is the slight, subtle alterations on road signs, whereby graffiti and stickers were enough to cause a neural network to 'misclassify stop signs as speed limit 45 signs or yield signs'.[340] Such adversarial attacks could be highly dangerous in situations where, for example, someone with malicious intent puts these otherwise harmless stickers on stop signs, and subsequently disrupts the performance of a self-driving car - which in turn may lead to potentially lethal road accidents because the car did not stop (but instead, accelerated its speed to up to 45 miles per hour because it thought the - compromised - road sign indicated it was allowed to do so).

It is easy to imagine an analogous incident in the battlefield. Say, for example, adversaries who find a way to disrupt the computer vision programme integrated in armed UAVs, which rely on the algorithm for navigation, target identification, selection and engagement. They conduct an adversarial attack on the system and 'trick' it into believing that a hospital is, in fact, a targetable military compound by making slight alterations to the hospital's otherwise 'obvious'

[339] See for example: Edward Hunter Christie and others 'Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability' (2024) 4 AI and Ethics <https://link.springer.com/article/10.1007/s43681-023-00261-0> accessed 14 July 2023, and Peter A. Milani 'Autonomous Weapon Systems for the Land Domain' (2020) Submission to the Concept for Robotics and Autonomous Systems 2040 <https://cove.army.gov.au/sites/default/files/autonomous_weapon_systems_for_the_land_domain.pdf> accessed 14 July 2023

[340] Evan Ackerman 'Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms' (*IEEE Spectrum*, 4 August 2017) <https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms> accessed 4 May 2021

signs indicating its nature (e.g., putting a sticker next to a red cross emblem, akin to the stop sign example). This would result in a situation where, without human oversight, the AI-enabled system would then deliberately target the hospital – protected in IHL – and thus lead to its violation. Although, obviously, the actual extent of risk that such adversarial attacks would actually happen is questionable (e.g., how would those malicious actors know of the 'vulnerabilities' that would trigger this malfunction in the adversary's AI-enabled system?), one cannot discount the existence of such risks and it is important to reflect on these risks against the ability of such AI-enabled systems use to comply with the rule of distinction.

This is where proxy data steps in, providing resilience and reliability to the system against those adversarial attacks. In fact, proxy data is, arguably, going to be central to nearly all algorithmic decision-making: In environments as messy and uncertain as the battlefield, it is rare to have direct evidence especially with regards to dual-use military objectives and membership to non-state armed groups. If, based on the previous example, the AI-enabled image recognition system did not exclusively rely on the presence of 'straightforward' indicators to assess an object's nature (e.g., does the object present a red cross and/or red crescent emblem? If yes - it is non-targetable; if not - it may be targetable). Rather, the system would also use, in its assessment, proxy, 'non-straightforward' variables (e.g., the number of vehicles rushing in/around/out of the building, potentially indicating the presence of emergency vehicles and thus, indirectly indicating the protected status of that building; although this must assumption be contextualized as traffic could also indicate a busy military post – hence the number and diversity of datapoints will be particularly important here) that would help feed into its calculations on whether or not the scanned object is indeed targetable. This way, even if the image recognition system is faced with an adversarial attack, the fact that it does not exclusively rely on direct variables but also proxy data to inform its calculations would make

the system much more resilient against risks from these adversarial attacks, and thus more reliable from a legal compliance perspective.

Of course, the opposite possibility must also not be excluded, where proxy data could lower the threshold for targeting – in other words, certain proxy indicators would be leveraged by the adversary to mislead the weapon system towards engaging certain targets. For example, a civilian structure turned into a military command and control centre can, purposefully, be surrounded by what would indicate its civilian status (e.g., the presence of playgrounds, busy traffic of non-military vehicles) and confuse the adversary. Putting aside the fact that this may be considered as use of human shields, a prohibited practice under statutory and customary international humanitarian law, such situation cannot be excluded from happening in reality.[341] This is where developers must ensure that they find the appropriate, technical solutions to ensure that proxy data are indeed bringing more benefits to accuracy, mission success and, subsequently, legal compliance and decrease risks of it being a liability.

These examples therefore demonstrate that the use of proxy data in AI-enabled systems should, in principle, not violate compliance with the rule of distinction *per se*. It is highly dependent on the system's performance, for which proxy data can be considered as a necessity in order to ensure its resilience against risks from adversarial attacks and, more generally, for the system's accuracy and performance. This, however, is obviously dependent on a number of technical and contextual factors - if, in the end, the use of proxy data would actually weaken the system's performance and its ability to act in compliance with the rule of distinction, obviously the system's developers would need to re-evaluate and re-calibrate the use of such proxy data in a way that would make the system in compliance with the rule of distinction.

[341] On the prohibition of using human shields, see 'Rule 97: Human Shields' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/Rule97> accessed 22 November 2024

Reliability of proxy indicators

Proxy variables are, as their name indicates - proxy. They may indicate certain elements, but this is limited to 'indications' providing a likelihood and must not be seen as oracles replacing the 'direct' data/variables sought to be identified by the AI-enabled programme. They may complement an assessment, but it must not be to the point where these proxy variables are taken as direct indicators of something. In fact, most tools will not be giving a direct indicator, only a probability – even for applications as straightforward as the image recognition of a tank. While assessing the standards of living in a certain population, one cannot exclusively rely on proxy indicators individually, such as access to the bathroom or certain electronic appliances, without the wider context in which these data are being used.

Such use of proxy indicators, however, must be exercised with caution in order to ensure compliance with IHL – particularly with the rule of distinction. As established in the previous section, the use of proxy indicators in themselves is not unlawful; what matters is the end-result and the role these proxy indicators play in influencing the programme's calculations and results which may or may not be unlawful. To this end, the decision to use proxy data must be informed and shaped by the law from the programme's development and testing stages. This would include, for example, making sure that the programme has been adequately trained and tested to factor in these proxy data in its calculations. The NSA's Skynet programme was trained to use proxy data, i.e., GSM metadata and more precisely call logs and location data. Once deployed, it should not be expected to function reliably to enable decision-making using other kinds of data, such as the content of the texts sent from mobile phones (thus, implying the use of language models to analyse these texts) – unless it has also been trained and tested to do so. Additionally, its outputs should also be taken in the wider context in which the programme is designed to be deployed in – this would be where human commanders must remain in the loop,

in order to ensure that decision-making processes are indeed done in compliance with the law with the necessary assessments and appropriate judgment.

Beyond proxy data, this is also applicable in general for all programmes: end-users must have been clearly informed of the programme's limitations and capabilities from its training and testing stages, including the type of data it would be able to process. This is particularly important if the end-user is meant to maintain a certain degree of supervision, or even control over the use of programmes that use proxy data. Not only would the programme be designed with an interface designed in such way that would enable the end-user to exercise their prerogatives; they must also have the necessary degree of technical literacy to be able to understand and use the programme in the intended way.[342]

In addition, one thing to note is that, beyond big data applications, the use of proxy indicators in warfare is actually not new. And while such use may have been necessary at times, there had been missed opportunities combined with error of human judgment, which subsequently led to incidents resulting in civilian loss. For example, in 1991, the United States conducted two aerial, laser-guided bombing attacks on the Al-Amiriyyah air-raid shelter in Iraq. The Pentagon erroneously considered the civilian bunker as a military command center due to the presence of electromagnetic pulse (EMP)-based equipment in addition to the presence of chain link and barbed wire fences.[343] Representatives of the United States notably tried to justify the attack with the military communications intercepted from within the facility, alongside aerial and satellite imagery indicating the presence of military vehicles and personnel.[344] Yet, there were also reports of obvious signs indicating that the compound was, in fact, a civilian shelter, with

[342] Although it is also important to note that, even with the required level of technical knowledge, the exercise of supervision and/or control by human operators does not constitute a silver bullet and will inherently carry risks related to human error; which we will discuss in further detail in Chapter 3.
[343] 'The Bombing of Iraqi Cities: Middle East Watch Condemns Bombing Without Warning of Air Raid Shelter in Baghdad's Al Ameriyya District on February 13' (1991) Human Rights Watch <https://www.hrw.org/reports/1991/IRAQ291.htm> accessed 4 February 2022
[344] Ibid

television footage showing signs marking 'shelter' both in Arabic and English – and even the designated serial number of 'Civilian Defense Public Shelter No. 25' on a sign at the building's entrance.[345] Coverage from the ground actually contradict the US' version of facts, with journalists visiting the building post-bombings and reporting indications that it had only been a shelter with no sign that it had been used as a military command centre.[346]

This incident demonstrates that while proxy indicators can be useful or even necessary at times, commanders should ensure that their intelligence is as comprehensive as possible. In the case of Al-Amiriyyah, the United States should have not exclusively relied on aerial surveillance and try to obtain, as much as possible, information on the shelter from the ground as well or at least, from other sources (e.g., human intelligence) attainable in the light of the absence of US military presence in Baghdad at the time. This is particularly due to the fact that the targeted compound was known to be located in a neighbourhood in Western Baghdad with a nursery school, a supermarket, and a mosque in the immediate vicinity of the bunker: if anything, these facilities should have been proxy indicators that the compound was, in fact, likely of civilian nature. Additionally, it was relatively common knowledge by those following the conflict at the time, that surrounding cities had used similar structures as civilian shelters at night. In the light of this knowledge, the US should have been held to a higher standard of precautionary measures to ensure their operations are done in compliance with the law.

In fact, in such 'high risk' areas where civilians are very likely to be present in and around the area, a comprehensive approach to intelligence collection is critical. Civilian objects such as the shelter are, by virtue of the rule of distinction, protected from direct attack – and not only has this rule been crystallised in international treaties, it has also been asserted and re-affirmed

[345] Ibid
[346] Robert D. McFadden 'WAR IN THE GULF: Iraq; IRAQIS ASSAIL U.S. AS RESCUE GOES ON' (*The New York Times*, 15 February 1991) <https://www.nytimes.com/1991/02/15/world/war-in-the-gulf-iraq-iraqis-assail-us-as-rescue-goes-on.html> accessed 16 July 2023

by the ICJ.[347] The law particularly holds, to a high standard, the protection of civilian objects and those with dual-use functions are only targetable in very specific situations limited in time, nature and purpose.[348] In case of doubt, where an object is normally dedicated to civilian purposes, the law adopts a very protective stance and the onus is on the attacking party to demonstrate that the object is being used to make an effective contribution to military action.[349] As such, even if the intelligence collected by the US suggest that the compound could have been used for military communication and/or as a command centre, they should demonstrate certainty that at the time of the bombings, the compound was used as such: proxy indicators, such as satellite and aerial footage of wired fences would not have been enough, especially in the light of the common knowledge described above. In other words, even if a military objective is deemed as targetable, steps must be undertaken to verify and validate its status as continuously targetable to the extent it is feasible, leveraging both proxy and direct indicators made available to the commander and decision-maker, as mandated by the rule of precautions. Human Rights Watch has in fact raised concerns as to whether the United States took all precautionary measures to minimise risks of incidental loss of civilian lives and damage to civilian objects.[350] Both customary and statutory, this obligation for precautions in attack would have mandated the US to not exclusively rely on proxy indicators.

Ultimately, the Al-Amiriyyah bombings show that while proxy indicators can be useful to collect intelligence on the civilian and potential military activities in designated areas, they must be taken in concert with many, and diverse, datapoints to contextualise these indicators and verify their veracity – especially when the standard is set higher, by the law, to justify the

[347] See for example API, art 48; Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3 (Rome Statute) art 8(2)(b)(i); *Legality of the Threat or Use of Nuclear Weapons* (n 69) [78-79]
[348] API, art 52(2)
[349] API, art 52(3)
[350] 'The Bombing of Iraqi Cities: Middle East Watch Condemns Bombing Without Warning of Air Raid Shelter in Baghdad's Al Ameriyya District on February 13' (n 343); 'Rule 15: Principle of Precautions in Attack' (n 278)

legality of attacks on the target in question. It is clear that today, with the increasing digitization of society, and access to greater amounts of data, big data solutions can help facilitate the collection and processing of this tremendous amount of data to feed into decision-making. This AI-enhanced capability could be key in meeting the high standard set, by the law, to ensure the legality of attacks: if anything, more data will lead to a more comprehensive picture of the situation on the ground where both proxy and 'direct' indicators would be taken altogether in order to avoid incidents akin to Al-Amiriyyah. These considerations are key from the development and testing stages of such AI solutions if, ultimately, these technologies were to foster compliance with the law.

3. Quantity of data

3.1. The quantity/volume of data

3.1.1. The more, the better?

There is a commonly-held assumption that the greater the volume of training data, the better it is for the performance of the programme in question.[351] Some may even be tempted to go as far as saying that even the most complex problems in AI 'may be solved by simple statistical models trained on massive datasets'.[352] For proponents of such belief, data is seen as the ultimate enabler for high-performing AI technologies capable of solving even the most complex, deeply convoluted issues.

As alluded to in our earlier discussions in this chapter, too little data can indeed lead to performance-related issues. For example, the development and use of synthetic environments for military training would need to have been exposed to enough training data to represent all

[351] Laura Galindo, Karine Perset, and Francesca Sheeka (2021) 'An overview of national AI strategies and policies' (2021) OECD, Going Digital Toolkit Note, No. 14 <https://goingdigital.oecd.org/data/notes/No14_ToolkitNote_AIStrategies.pdf> accessed 21 July 2022

[352] Alon Halevy, Peter Norvig, and Fernando Pereira 'The Unreasonable Effectiveness of Data' (2009) 24(2) IEEE Intelligent Systems <https://ieeexplore.ieee.org/document/4804817> accessed 21 July 2022; as cited in Xiangxin Zhu and others 'Do We Need More Training Data?' (2016) 119 International Journal of Computer Vision <https://link.springer.com/article/10.1007/s11263-015-0812-2> accessed 21 July 2022

variables and eventualities to be considered in said environment – including details such as weather conditions.[353] The weather can exercise a lot of direct and indirect influence on military decision-making, and simply not having enough training data to factor in these perhaps small but extremely impactful details can have serious consequences not only on the programme's reliability, but also on the military's training and performance in general. In a way, this is closely related to our earlier discussions on the need for diversity of data: the latter is essential to ensure representation and an accurate depiction of reality. This ultimately leads to the question of access to large volumes of training data sufficiently representative of the environment and context in which the machine is expected to operate the context surrounding the problem at hand; and that would be aligned with the purpose for which the programme was developed.

On the flipside of the coin, one could also argue that too much data can lead to too much noise. Key parameters to target, collect, aggregate and use the right set of training and testing data are critical, as too much data could 'distract' the programme from the issues it is ultimately designed to solve. In fact, too much data carries risks of not being targeted and focused enough; and may even bring in irrelevant data that would cause (preventable) irregularities and anomalies in the datasets. This would result in too much of a high variability above the necessary threshold, and the programme will become too unpredictable and, ultimately, unreliable. This noise can either come from the data in itself (e.g., irrelevant data), or even from the labelling of the data that may, or may not, be as accurate and relevant. As such, the quantity and quality of the training and testing datasets are intimately linked; and it is critical that

[353] Nae-hyun Cho, Jong-chul Park, Man-kyu Kim 'The Application of Distributed Synthetic Environment Data to a Military Simulation' (2010) 19(4) Journal of the Korea Society for Simulation <https://koreascience.kr/article/JAKO201016450102376.page> accessed 21 July 2022

developers exercise caution on a number of considerations pertaining to the quantity, given the effect it may have on the overall performance and accuracy of the programme.[354]

There are, in fact, a number of considerations to factor in when assessing the appropriate quantity of training data required. In addition to accessibility and the availability of data, the following must be taken into account: the problem the programme has been designed to solve (e.g., its complexity, who and what is involved in the problem, the complexity of the context surrounding the problem); the programme's own complexity (e.g., what the developer wants it to do, how, the number of parameters considered and the relevant data to each parameter, etc.) These are issues that need to be addressed, and the solution to which extend far beyond the mere volume of training data. Due to space limitations, this thesis will not unpack these considerations in detail; they would, however, make for a compelling study in the context of examining the intersection between data and AI in the military domain.

3.1.2. Large datasets and the legal requirement for contextualisation capabilities

Large volumes of data would constitute key enablers for contextualisation capabilities in AI technologies. Here, contextualisation capabilities are to be understood as the system's ability, through its programming and training, to take into account various types of datapoints that are different in nature in order to paint the bigger picture, i.e., the context in which the problem at hand is being assessed. Of course, in passing, a separate discussion can be held as to whether AI is capable of understanding context – and there is a rich (and highly controversial) body of literature and research dedicated to revisiting the notion of 'intelligence' in the light of technological progress in the AI space.[355] Aside from the technicalities and wider, fascinating

[354] Sainbayar Sukhbaatar and others 'Training Convolutional Networks with Noisy Labels' (2014) ArXiv <https://arxiv.org/abs/1406.2080> accessed 21 July 2022

[355] Across a wide range of AI applications, researchers have asserted the importance of context, are debating as to whether or not 'contextual AI' is indeed a thing, and cast cautionary tales around the hype on the narratives surrounding these technologies. See for example in the legal field and on transparency requirements against contextual concerns: Heike Felzmann and others 'Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns' (2019) 6(1) Big Data & Society <https://journals.sagepub.com/doi/full/10.1177/2053951719860542> accessed 16 July 2023; in the context of

discussions as to whether AI can indeed possess the ability to understand context, it would be useful for the purpose of this thesis to simply assert that developers are, today, capable of designing and training programmes to solve highly complex issues.

AI programmes are indeed able, today, to 'digest' (very) large training datasets and parameters, enabling holistic and comprehensive assessments that makes them practically able to contextualise. It would however be important to establish that in the present discussion, 'contextualisation' must be understood as not necessarily equating to the ability to grasp context in such depth that it equates to a human agent's capabilities but rather, the ability to connect a wide range of data surrounding the problem at hand in a complex, holistic, and comprehensive way. For example, in the context of large language models, it has been argued that these programmes are not, in themselves, capable of understanding the context surrounding the training data and the generated text; rather, they are able to process vast amounts of datasets that, altogether, form a wider context in which it produces its outputs and adapt accordingly.[356] However, these programmes could potentially be capable of, if developed and trained to do so, processing large volumes of data providing contextual information to perform better: for large language models, this would be for example to identify and factor in 'better correlations between different semantics' for complex tasks such as word sense disambiguation.[357]

---

employment Yuan Pan and others, 'The adoption of artificial intelligence in employee recruitment: The influence of contextual factors' (2021) 33(6) The International Journal of Human Resource Management <https://www.tandfonline.com/doi/abs/10.1080/09585192.2021.1879206> accessed 16 July 2023; in the medical field W. Nicholson Price II 'Medical AI and Contextual Bias' (2019) 33(1) Harvard Journal of Law & Technology <https://heinonline.org/HOL/LandingPage?handle=hein.journals/hjlt33&div=6&id=&page=> accessed 16 July 2023; and more generally see also from an philosophical and morality perspective Niels van Berkel and others 'Human-centred artificial intelligence: a contextual morality perspective' (2020) 41(3) Behaviour & Information Technology <https://www.tandfonline.com/doi/abs/10.1080/0144929X.2020.1818828> accessed 16 July 2023

[356] Large language models generate text by, bluntly and grossly defined, simply putting strings of words together in a way that it makes sense grammatically and semantically. It does not mean, however, that the machine understands, means, or feels any of the text it generates. See: Emily M. Bender and others 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' (FAccT 3-10 March 2021, Virtual Event, Canada, 2021) <https://dl.acm.org/doi/10.1145/3442188.3445922> accessed 22 July 2022

[357] Erik Cambria, Bebo White 'Jumping NLP Curves: A Review of Natural Language Processing Research' (*IEEE Computational Intelligence Magazine*, 11 April 2014) <http://sentic.net/jumping-nlp-curves.pdf> accessed 22 July 2022. For a more general discussion on AI and context, see: Gadi Singer 'Advancing Machine Intelligence:

Contextual targetability

Contextualisation lies at the heart of many, if not all assessments of legal compliance. For example, in the context of the rule of distinction, and as provided in Article 52(2) of API, the definition of a military objective takes into account both (1) the characteristics of the object in question (i.e., 'the nature, location, purpose, or use make an effective contribution to military action') and (2) the context surrounding the object (i.e., the implications of the 'total or partial destruction, capture or neutralization' of the object in question against 'the circumstances ruling at the time').[358] As established by a number of scholars, these elements are 'not necessarily static, but rather dynamic'.[359] In other words, it is not because an object is indeed targetable at a specific point in time that it will continue to be so at a later time: the status of an object does not only depend on its nature but also its function and use. As such, not only does the rule of distinction mandate the conduct of contextual assessment; the rule of precautions in attack would also mandate through all 'feasible' means the verification and validation of a target's status.[360]

Similar to dual-use objects, context also plays a central role in discussions surrounding the targetability of civilians directly participating in hostilities. In fact, while civilians are generally protected from direct attacks, international humanitarian law prescribes the loss of their immunity when they are established as directly participating in hostilities.[361] What the latter means is still subject to much scholarly debate; and one particular notion that came out of the

Why Context Is Everything' (*Towards Data Science*, 10 May 2022, <https://towardsdatascience.com/advancing-machine-intelligence-why-context-is-everything-4bde90fb2d79> accessed 22 July 2022

[358] API, art 52(2)

[359] Wagner M 'Autonomy in the Battlespace: Independently Operating Weapon Systems and the Law of Armed Conflict', in Dan Saxon (ed) *International Humanitarian Law and the Changing Technology of War* (2013) (Martinus Nijhoff Publishers, Leiden, Boston 2013) <https://brill.com/edcollbook/title/21680?language=en> accessed 22 November 2024

[360] For more discussons on 'feasibility', see earlier in Chapter 2, Section 2.1.2.

[361] Melzer (n 336)

discussions is the 'revolving door', as in a civilian's targetability could change depending on the context.[362]

For example, in the context of the Russian invasion on Ukraine, Ukrainian civilians are using apps on their smartphones to provide key information and to generally assist in the war effort. This is done in a range of ways, most notably by helping locate and report the movement of Russian troops to the Ukrainian Armed Forces through an app (Diia, originally done for public services in peacetime allowing the payment of taxes, access identity documents, access to citizens' Covid pass, etc.) that has been re-purposed with a new feature called E-Enemy.[363] As noted in one analysis, this is a typical case study where new technologies are blurring the line regarding the civilians' status and whether they would be considered as directly participating in hostilities by sending these data to the Ukrainian Armed Forces.[364] In response, the adversary (Russian forces) could target Ukrainian civilians using this app, with the argument that they are directly participating in hostilities at that very moment and thus, making them targetable under the law. This hypothesis could make sense from a legal perspective vis-à-vis the rule of

[362] There are a number of studies dedicating to unpacking this notion of 'revolving door' and its very application is also controversial and subject to much debate. While fascinating, such discussions would fall beyond the scope of this thesis. For reference, see for example: Michael N. Schmitt '„Direct Participation in Hostilities" and 21st Century Armed Conflict' in H Fischer (ed) *Crisis Management and Humanitarian Protection* (BWV 2004) <https://www.uio.no/studier/emner/jus/humanrights/HUMR5503/h09/undervisningsmateriale/schmitt_direct_participation_in_hostilties.pdf> accessed 16 July 2023; Bill Boothby 'And for Such Time as: The Time Dimension to Direct Participation in Hostilities' (2009-2010) 42 N.Y.U. J. Int'l L. & Pol <https://heinonline.org/HOL/LandingPage?handle=hein.journals/nyuilp42&div=27&id=&page=> accessed 16 July 2023; and Alessandro Silvestri 'The Revolving Door of Modern Warfare: Civilian Direct Participation in Hostilities' (Thesis, The University of Western Australia 2022) <https://api.research-repository.uwa.edu.au/ws/portalfiles/portal/223429021/THESIS_DOCTOR_OF_PHILOSOPHY_SILVESTRI_Alessandro_2022_Part_1.pdf> accessed 16 July 2023. These discussions have even extended to those in the context of cyber operations, see for example: Tassilo Vp P. Singer 'Update to revolving door 2.0: The extension of the period for direct participation in hostilities due to autonomous cyber weapons' (9th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2017) <https://ieeexplore.ieee.org/abstract/document/8240332> accessed 16 July 2023

[363] See: Yaroslav Druziuk 'A Citizen-like chatbot allows Ukrainians to report to the government when they spot Russian troops – here's how it works' (*Business Insider*, 18 April 2022) <https://www.businessinsider.com/ukraine-military-e-enemy-telegram-app-2022-4?r=US&IR=T> accessed 22 July 2022; Tim Judah 'How Kyiv was saved by Ukrainian ingenuity as well as Russian blunders' (*Financial Times*, 10 April 2022) <https://www.ft.com/content/e87fdc60-0d5e-4d39-93c6-7cfd22f770e8> accessed 22 April 2022

[364] Lukasz Olejnik 'Smartphones Blur the Line Between Civilian and Combatant' (*WIRED*, 6 June 2022) <https://www.wired.com/story/smartphones-ukraine-civilian-combatant/> accessed 22 July 2022

distinction, notably if Russia manages to justify such use of the app as meeting the three cumulative criteria defined by the ICRC on establishing a civilian's status as directly participating in hostilities.[365] In fact, the ICRC's interpretative guidance argues that the transmission of tactical intelligence could constitute one of the acts with direct causation of harm – one of the cumulative criteria set to establish a civilian's direct participation in hostilities.[366] In a way, a separate, more elaborate conversation could be held on whether these criteria would be applicable to Ukrainians 'merely' sending this data via the Diia app – especially with regards to the threshold of harm and the need for a direct causation.[367] But what is important to note here is that, if Russia were to deploy an AI-enabled solution to directly target users of the Diia app, simply identifying and targeting the civilian users would not be enough. It would need to do so against an array of contextual considerations, including as to whether or not the data sent by the targeted user is consequential enough to classify them as directly participating in hostilities and thus, targetable.

Thus, in both instances for objects and persons, how the law will apply will highly depend on the context. AI-enabled solutions that pretend being capable of informing decision-making processes surrounding military targeting in compliance with the law must be capable of factoring in the wide range of parameters and variables relevant to the context in which the operation is undertaken. In this sense, the more relevant data there is to provide context, the better for legal compliance. This is particularly relevant given the inherently complex, messy and dynamic nature of warfare and the context surrounding military targeting operations. This complexity is particularly pronounced today with the use of everyday technologies by civilians that further blur the line between the targetable and the non-targetable. To this end, large

[365] Melzer (n 336)
[366] Ibid
[367] This would however fall beyond the scope of this thesis, although such discussions are important to have in the light of increasing digitisation of warfare.

training datasets would be necessary – and could even be argued as a legal obligation in the light of the rule of precautions, if one were to argue that such large training datasets could indeed reduce risks of incidental harm and collateral damage on civilians and civilian objects.

Proportionality

It also goes without saying that contextualisation capabilities are also necessary for proportionality calculations, which strikes the balance between the expected level of 'incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof' on the one hand, and 'the concrete and direct military advantage anticipated' on the other hand.[368] Again, such assessments are highly contextual and depend on a number of considerations both from the ground and at the strategic and tactical level. This is relevant, for example, at the planning stage of a military targeting operation, for the assessment on the expected harm that will be caused by the attack, as well as on 'how reasonably foreseeable' would it be to expect such harm.[369] Proportionality assessments are also very much relevant, among others, in the context of dual-use targeting, i.e., those against objects that can serve both civilian and military purposes (e.g., a bridge, communications facilities, etc.)[370]. As such, the more data there is on the battlefield in question, the better and the more comprehensive the assessment would be and the greater situational awareness would be; thus enabling compliance with the rule of proportionality. In a way, one could even argue that in the light of increased data generated in and around battlefields, belligerents may have a certain duty to collect and process such data if

[368] API, art 51(5)(b)

[369] Emanuela-Chiara Gillard 'Proportionality in the Conduct of Hostilities: The Incidental Harm Side of the Assessment' (2018) Chatham House <https://www.chathamhouse.org/sites/default/files/publications/research/2018-12-10-proportionality-conduct-hostilities-incidental-harm-gillard-final.pdf> accessed 7 May 2021

[370] See Maurice Cotter 'Military Necessity, Proportionality and Dual-Use Objects at the ICTY: A Close Reading of the Prlić et al. Proceedings on the Destruction of the Old Bridge of Mostar' (2018) 23(2) Journal of Conflict and Security Law <https://academic.oup.com/jcsl/article-abstract/23/2/283/5069317> accessed 16 July 2023; and Marco Sassòli 'Legitimate Targets of Attacks under International Humanitarian Law' (Background Paper prepared for the Informal High-Level Expert Meeting on the Reaffirmation and Development of International Humanitarian Law, Cambridge, January 27-29, 2003) <http://humanrightsvoices.org/assets/attachments/documents/Session1.pdf> accessed 16 July 2023

it is indeed known to reinforce proportionality assessments and, ultimately, enable better compliance. Such argument could be made in relation to the duty to respect and ensure respect, in addition to the rule of precautions if such data collection would minimize incidental loss of civilian life, injury to civilians and damage to civilian objects. This is particularly important given the dynamic and ever-changing nature of warfare and thus, constant change in the presence and weight of relevant factors in the proportionality calculation.[371] A couple of caveats must, however, be added. First, belligerents may have different access to technologies due to limited financial, human and technical capacity, particularly States from the Global South whose access to cutting-edge sensors and data analytics tools may be limited. Should an argument be made that States have a duty to collect and process as much data as possible to support and reinforce proportionality assessments and, more generally, foster compliance, expectations must be commensurate and measured against the respective capabilities belligerents may have in the first place, as this will have subsequent implications on the quantity and quality of data at hand, in addition to their overall usefulness for legal compliance and even for operational support. Second, the collection of such data at scale, while in principle critical to enhance proportionality assessments, must also be done in accordance with the belligerents' obligations under international human rights law, particularly with regards to the right to privacy.[372] While IHL may take precedence on the basis of *lex specialis*, how international human rights law applies in armed conflict and their implications on belligerents' data collection and processing practices in times of armed conflict must not be forgotten.

This presupposes the need for large training datasets for the purpose of training the programme, both at the pre-deployment level ('in the lab') and even post-deployment in the case of

[371] Gary D. Solis, *The Law of Armed Conflict: International Humanitarian Law in War* (1st edn Cambridge University Press, 2010) 273

[372] Yasmin Afina, 'Regional Perspectives on the Application of International Humanitarian Law to Lethal Autonomous Weapon Systems' (2025) UNIDIR <https://unidir.org/publication/regional-perspectives-on-the-application-of-international-humanitarian-law-to-lethal-autonomous-weapon-systems/> accessed 3 June 2025

reinforcement learning programmes, which 'continuously' learn even post-deployment based on rewards (e.g., mission success). The more and larger training datasets are, the more the programme will 'know' of possible variables and parameters. The more comprehensive its assessments will be, the more it will be able to comply with the rule of proportionality. For example, armed forces can use AI-enabled synthetic environments to simulate an attack and plan accordingly. Such simulations would enable to not only estimate the effects and advantage of the planned attack, but also the anticipated collateral damage both on civilians and objects surrounding the target – thus helping decision-makers in estimating the planned attack's compliance (or non-compliance) with proportionality obligations and adapt accordingly. Yet, in order to develop such environments enabling accurate simulations and calculations, large amounts of contextual data would be necessary – not only with regards to the simulated environment in question (e.g., number and flux of civilians in/around the area of operation; accurate census of civilian infrastructure and residential areas, etc.). Mission-relevant training data will also be necessary in order to simulate, for example, the likely effect of a missile's trajectory and blast on certain types of buildings at a specific time in the day/night and under specific weather conditions (e.g., windy and rainy conditions can change the trajectory of a missile).[373] And these data must be relevant to the context in which the simulation would be carried out – as discussed in earlier sections of this chapter with regards to the veracity and adequacy of training data. As such, the more relevant training data and parameters the programme is exposed to, the more comprehensive its assessments will be. The more accurate its simulations will be, the more it will be reliable to support proportionality assessments.

3.2. Descriptors

The present section will delve into descriptors due to the added nuances they bring to considerations surrounding the need for training and testing data at scale. In fact, as discussed in the first half of this

[373] Rao, De Melo, Krim (n 151)

chapter, beyond the assumption that 'the more data the better', this data must be of high quality: Descriptors act as a 'bridge' between quantity and quality, as they quantify the very attributes and properties of the data that will determine their quality. Further details surrounding the definition and relevance of descriptors will be provided in the first sub-section. Design choices surrounding descriptors may have implications on IHL application further down the line, for example with regards to the positive identification of targets as then discussed in the second sub-section below.

3.2.1. Definition and relevance

Descriptors, or 'feature descriptors', are pieces of information that describe the features of the data being analysed. In other words, descriptors quantify the attributes and properties of that data. In this sense, descriptors are particularly (but not only) relevant for tasks like classification and clustering, as they provide a way of quantifying and representing complex data in a form that the algorithm can process.

This is the case, for example, for computer vision, i.e., AI applications that involve the processing of images, such as 'object tracking, 3-D imaging, mobile augmented reality, image matching, object detection and image registration'.[374] These descriptors can be 'global', as in they provide information about the entire image at a greater scale, or they can also be more targeted, 'local', focusing on a specific part of the picture in question.[375] For example, in a picture that captures a snapshot of a city centre, global descriptors would provide information about the entire picture of the city (i.e., it is a picture of a city centre); whereas local descriptors would focus on certain, specific sections of the picture that may be of interest (e.g., local descriptors about a specific building could relate to the colour and size of the windows, their shape, etc.) The choice, and number, of descriptors bear great influence on the accuracy of the

[374] Khushbu Joshi, Manish I. Patel 'Recent advances in local feature detector and descriptor: a literature survey' (2020) 9 International Journal of Multimedia Information Retrieval <https://doi.org/10.1007/s13735-020-00200-3> accessed 28 July 2022
[375] Ibid

programme and thus, its reliability for use. As such, the quantity of feature descriptors would be important, as the greater the number of descriptors there is, the more the description of the image in question will be accurate. A large volume of descriptors would for example enable the detection of even small objects in a highly complex environment/setting. This, however, comes with a number of caveats.

First, the volume of descriptors should indeed be adequately reflecting the volume of information contained in the image in question, and the overall objective of the programme. For example, a computer vision programme has been developed to identify, through surveillance footage, e.g., from a network of closed-circuit television (CCTV) in a city, individuals likely to be members of non-state armed groups blending in a crowd of civilians. The programme must have been trained on key, relevant descriptors that would indeed be relevant for the problem at hand. Such relevant descriptors could, for example, be patterns of movement and trajectories; carrying openly firearms; and/or certain behavioural patterns that would, potentially, indicate somebody's membership to an armed group (with, of course, the caveat that these must be taken in the wider context as discussed earlier in the chapter). Irrelevant descriptors would, for example, be the size and colour of windows in surrounding buildings; or the algorithm would 'lose' its focus and become obsolete for the purpose it was originally programmed for. In addition, descriptors would also need to meet a certain number of quality-related criteria, such as its repeatability (i.e., the programme would be able to detect the same object/features presented under different viewing conditions); the descriptors should also be accurate; efficient; local; and distinctive.[376] This presupposes that a large volume of descriptors is not enough and not a solve-all solution.

[376] Tinne Tuyterlaars and Krystian Mikolajczyk 'Local Invariant Feature Detectors: A Survey' (2007) 3(3) Foundations and Trends in Computer Graphics and Vision <https://homes.esat.kuleuven.be/~tuytelaa/FT_survey_interestpoints08.pdf> accessed 28 July 2022

Another important point to stress is that while human programmers could pre-set some key descriptors and 'teach' the algorithm to detect objects with matching features, for a lot of applications, the detection of features/descriptors is done by an algorithm. In fact, for these applications, automating the identification of descriptors lies at the heart of the programme's functionality and purpose, as programmers may see an added-value in having an algorithm identify descriptors instead of humans who would otherwise not be able to detect the same target objects/features. For example, a programme could be designed to find secret military bases by processing large volumes of satellite imagery over the adversary state's territory. The programme could work on the basis that these military bases have certain, distinct features/descriptors that the programme has detected and deemed as indicative of a military compound, based on complex calculations that would require great compute power and which may be so complex that humans would not even be able to run them by themselves. This kind of application presents great potential for ISR purposes and thus, to inform, influence and even shape military targeting operations. Yet, as alluded earlier, the process behind the detection and definition of these descriptors can be so complex and convoluted that it may not always be possible to understand and explain the calculations – i.e., a black box problem and its subsequent legal issues, which we will further discuss in Chapter 3.

3.2.2. The case for a legal obligation of targets' positive identification

As established in the previous section, the automated definition of descriptors can prove to be extremely helpful for intelligence, surveillance and reconnaissance, which would subsequently inform and shape military targeting operations. This can, for example, be the case of programmes developed to identify hidden military compounds of the adversary based on satellite imagery. This can also be done for anti-personnel targeting – for example, through computer vision programmes connected to city-wide CCTV networks and flag individuals, whose patterns of movement and behaviour may suggest their membership to certain non-state

armed groups. Such capabilities can also be integrated into weapons systems to enable high-precision targeting on a specific individual and/or object based on key descriptors that would correspond to those of a desired target. These solutions are, in principle, possible as long as the programme has been adequately trained and tested on large volumes of datasets with the appropriate quality and descriptors.

While this application would indeed be interesting from the perspective of intelligence collection and mission success, more generally, our discussion leads us to the question of whether their development and deployment would be feasible in compliance with IHL. Perhaps more broadly, even beyond applications with automated descriptors definition, the question of whether all forms of ATR technologies could be developed and deployed in compliance with the law arises. The answer will obviously very much depend on a number of factors and technicalities for each programme, including their set parameters, their training and testing datasets, where they come from, how it has been trained and tested, the level of oversight and control exercised by human operators not only post-deployment but also in the development and testing stages, etc. This will also depend on a number of additional considerations, such as where the technology is being integrated in (e.g., does the AI solution process data from a network of surveillance assets at the command and control level, or is it integrated within a weapon system?); when the ATR process happens (e.g., is it done at the intelligence collection/planning level, or in dynamic targeting settings?); the very purpose for which the technology is being developed; etc.

Against this backdrop, with the myriad of possibilities in developing ATR solutions, a comprehensive, one-size-fits-all application of the law is practically impossible, nor desirable. In fact, it is practically impossible because, as we are establishing throughout the present thesis, there are a number of legal considerations that apply depending on how AI technologies for military targeting are being developed and tested, and these considerations may vary from one

application to another. The design, testing and evaluation of ATRs for anti-material (e.g., identification of hidden bunkers in the desert based on satellite imagery) will be inherently different to those of anti-personnel (e.g., identification of non-state armed group members through GSM mobile data) both from a technical and legal perspective. One would operate on the basis of classification of a specific type of data (e.g., imagery) while the other would operate on the basis of inferences drawn from patterns and correlation, e.g., drawn from call logs and global positioning system (GPS) tracking. From a legal perspective, although in the same spirit, assessment of their targetability will also be different, i.e., the definition of military objectives is distinct to the definition of civilians directly participating in hostilities because of their inherently different nature. As such, there is a need for clarity and granularity in terms of how the law applies to each of these applications, and how relevant IHL requirements could be translated into technical recommendations (e.g., benchmarks to evaluate an anti-personnel ATR's compliance with the rule of distinction). These clarifications would ultimately ensure that existing international law is understood well enough so that not only would there be less risks of 'grey areas' and confusion once these technologies are deployed. Clarity on how the law applies would also ensure that these technologies are being developed, from the outset, as compliant with the law 'by design' – which ultimately constitutes the driving force of this thesis.

In the midst of (necessary) nuances in applying the law, one common assertion that we could make is – if ATR solutions are indeed being developed through the automated identification of descriptors, these programmes must exclusively be limited to the positive identification of lawful targets. This assertion is particularly underpinned by two rules: the rule of distinction, and the rule of precautions in attack.

On the rule of distinction, the law differentiates individuals and objects that would make for lawful targets on the one hand (i.e., combatants, civilians directly participating in hostilities

and military objectives), and those that would be protected from direct attacks (i.e., civilians and civilian objects). These persons and objects are established as targetable through a set of positive, cumulative criteria these targets must meet in order to make any direct attack on them compliant with the rule of distinction. Conversely, civilians and civilian objects are defined negatively, as anyone not taking part in combat and any object that is not military objective. In other words, the assumption from the law is that all objects and persons are non-targetable unless proven otherwise (i.e., by meeting the cumulative criteria for combatants, civilians directly participating in hostilities, or military objectives). In addition, in case of doubt, persons and/or objects in question fall into the civilian realm; thus overall making IHL's approach incredibly protective.[377]

Subsequently, when it comes to developing ATRs, these programmes must be designed in a way that aligns with this protective approach by the law – i.e., they must exclusively be programmed to conduct the positive identification of targetable persons and/or objects, functioning with the assumption that they are protected against direct attack unless proven otherwise with the highest degree of confidence. This prerogative would be applicable to all ATRs, both for individual and material targets, and help safeguard a minimum threshold where, when applied across the board, risks of non-compliance will be minimised, and these technologies align with the law's protective stance 'by design'. Cognisant of limited resources, this would enable programmers to focus testing and evaluation processes on very focused descriptors of targets if the rest is assumed to be non-targetable anyways – instead of both having to assess descriptors of targets on the one hand, and those of non-targets on the other hand. In addition, this presumption of non-targetability (i.e., presumption of negatives) would

[377] Melzer (n 336); see also Michael N. Schmitt, 'Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics, (2013) 4 Harvard National Security Journal <https://harvardnsj.org/2013/02/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/> accessed 1 August 2022

indeed align with IHL's approach with regards to doubts surrounding someone or an object's status and targetability.

In addition to aligning with the rule of distinction, standardising all ATRs as, by default, programmed to undertake the positive identification of lawful targets only would also align with the rule of precautions. As established earlier in this thesis, the applicability of this rule also extends to the early stages in the technology's lifecycle, i.e., from 'within the lab'. As such, developers have an obligation to undertake all feasible precautionary measures to minimise risks of harm both on civilians and civilian objects. In this sense, if ATRs operate on the basis that all subjects are, by default, non-targetable unless proven otherwise, the positive identification of targets must be set with the highest level of confidence attainable. In other words, the programme must be exhibit high rates of confidence in its calculations and the high probability that the identified target is, indeed, positive and accurate. Studies have demonstrated that for those programmes set to operate on high levels of confidence with thresholds over 90 percent, they are more likely to produce false negatives.[378] In the context of ATRs, false negatives would correspond to situations where a lawful target has been misidentified as unlawful target. While this may have an impact on mission success, from a legal perspective, this would be a much more protective approach than the opposite (where false positives would be more likely) as a false negative will likely lead to inaction (i.e., not engaging with the target).

In fact, false positives are more problematic vis-à-vis legal compliance, i.e., where the programme falsely identifies a target as lawful when they are actually protected from direct

[378] Department of Defense (US), Office of the Under Secretary of Defense For Acquisition, Technology and Logistics, Defense Science Board Task Force 'Munitions System Reliability (2005) 20301-3140 <https://permanent.fdlp.gov/lps72288/ADA441959.pdf> accessed 4 August 2022; as cited by Alan Backstrom, Ian Henderson 'New capabilities in warfare: an overview of contemporary technological developments and the associated legal and engineering issues in Article 36 weapons reviews' (2012) 94(886) International Review of the Red Cross <https://international-review.icrc.org/sites/default/files/irrc-886-backstrom-henderson.pdf> accessed 22 November 2024

attacks. False positives could happen for a number of reasons which, beyond military applications, is the subject of much research.[379] In the context of ATRs, false positives could result, for example, from biased training data, as discussed earlier in this chapter, but also issues stemming from the descriptors. They could either be inaccurate, i.e., falsely associating certain attributes to targets, or there could also be too many descriptors, lowering the threshold and widening the cast for positive targets (and thus, increasing risks of false positives and beating the very purpose of restricting ATRs to the positive identification of lawful targets only for protective reasons).

While we will discuss in further detail legal issues surrounding uncertainties surrounding AI-enabled decision-making in military targeting in Chapter 4, a quick discussion on the legal implications of false positives from ATRs is necessary. The use of ATR tools, as its name indicates, would imply delegating certain decision-making processes to AI-enabled programmes, with the assumption that these technologies are indeed reliable for use within the boundaries of the law. As such, false positives emanating from ATRs, while unfortunate, are not always necessarily unlawful.

Even prior to the AI era (or at least, today's situation where AI is in the process for wide integration across the board), false positives in target identification were of concern. For example, in 1999, a NATO aircraft mistakenly fired several missiles at the Chinese Embassy in Belgrade, which resulted in the death of 3 Chinese citizens in addition to an estimated of 15 injured, in addition to damage done to the Embassy's building and those surrounding it. Despite

[379] False positives are of particular concern (and the subject to much research) in the medical field due to their profound implications on healthcare delivery, see for example: Hyo-Eun Kim and others, 'Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study' (2020) 2(3) Lancet Digital Health <https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30003-0/fulltext?amp=1> accessed 18 July; the education system is also grappling with this issue especially in the light of mass use of generative AI, see for example: Doraid Dalalah, Osama M. A. Dalalah 'The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT' (2023) 21(2) The International Journal of Management Education <https://www.sciencedirect.com/science/article/abs/pii/S1472811723000605> accessed 18 July 2023

establishing the civilian nature of the buildings and individuals injured or deceased as a result of the bombing, the International Criminal Tribunal for the former Yugoslavia (ICTY)'s Office of the Prosecutor (OTP) still decided not to investigate this incident because the aircrew and senior leaders involved in the attack were given the wrong information.[380] In a hypothetical, but analogous case where the false positive identification of a target as such results from the outputs of an ATR system. How the law applies may follow the same reasoning as the ICTY's OTP – as in, the attack was carried on the basis of wrong information provided by the programme, and those involved in the incident did not know of the information's nature (a false positive). Thus, they cannot be held liable for relying on a false positive – as long as the appropriate, feasible precautionary measures have been undertaken to check the veracity of the programme's outputs. Some, most notably proponents of a ban treaty on autonomous weapons systems, have been wary of such situations and have even shared concerns of a potential legal vacuum.[381]

The very existence of such a gap in the law is, however, open to question. Narratives asserting the supposed existence of a lacuna in accountability not only risk being wrongfully exploited but also challenge the established principle that accountability will always be established in the light of the IHL obligation to investigate.[382] Moreover, the use of AI-enabled technologies in military targeting inherently involves and carries risks, given that such systems can never achieve one hundred percent accuracy. The degree of uncertainty that ensues could, at worst,

---

[380] International Criminal Tribunal for the Former Yugoslavia, *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia* (13 June 2000) <https://www.icty.org/x/file/Press/nato061300.pdf> accessed 21 November 2024

[381] A coalition of civil society organisations (the Campaign to Stop Killer Robots) and a number of countries are particularly calling for a new international treaty banning lethal autonomous weapons systems due to the potential for an accountability vacuum. A number of academic literature has also been shedding light on this question, see for example: Ilse Verdiesen, Filippo Santoni de Sio, Virginia Dignum 'Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight' (2021) 31 Minds and Machines <https://doi.org/10.1007/s11023-020-09532-9> accessed 6 August 2022

[382] Lubell N, Public evidence for the Artificial Intelligence in Weapons Systems House of Lords Select Committee, 'AI in Weapon Systems Committee' (London, 23 March 2023) <https://committees.parliament.uk/oralevidence/12931/pdf/> accessed 18 July 2023

lead to the inadvertent targeting of protected objects or individuals. Nevertheless, uncertainty in technological reliability is not new and not exclusively inherent to those that are AI-enabled: even non-autonomous technologies grapple with risks of malfunction or misuse, for example due to 'traditional' cybersecurity vulnerabilities.[383] Even human decision-making will never be at one hundred percent reliable, for example due to habits, biases and cognitive limitations. An extra layer of complexity comes to light when evaluating the difference between a system's limitations over the certainty of its performance on the one hand, and the user's perception and reliance on the technology on the other hand. For example, this issue would arise if a commander is operating a system that can only operate at maximum 80% of certainty based on its testing and evaluation, yet they follow its outputs and recommendations thinking the technology is operating at 100% certainty. There will always be a degree of uncertainty, the implications of which will be discussed in further detail in Chapter 4. It is however important to note, at this stage, that this uncertainty does not necessarily equate to a legal vacuum in accountability and criminal responsibility. This applies to AI-enabled technologies as well – even in situations where, for example, civilians have been inadvertently targeted due to convoluted and too-complex descriptors. Establishing the criminal responsibility of those involved in the decision-making process and eventually the development and testing phases of the technology is a critical area of the law in dire need for clarification – especially under which circumstances, for which technology, at what stage of the targeting cycle, etc.; however, such fascinating discussion would fall beyond the scope of this thesis.[384]

[383] Government Accountability Office (US) 'Weapon Systems Cybersecurity: DOD Just Beginning to Grapple with Scale of Vulnerabilities' (2018) GAO-19-128 <https://www.gao.gov/assets/700/694913.pdf> accessed 6 August 2022

[384] A team in the Graduate Institute in Geneva has, for example, a project dedicated to analysing 'the challenges of ascribing criminal responsibility for war crimes raised by the advent of increasingly autonomous weapon systems and human-machine shared decision-making in the targeting process (so-called mixed-initiative systems).' See: 'LAWS & War Crimes Project: Algorithmic Warfare and Accountability', <https://lawsandwarcrimesdotcom.wordpress.com/> accessed 6 August 2022

Chapter III – Design and Development of AI-Enabled Systems

1 Introduction

1.1. Scope

This third chapter will look closely at select elements from the design and development stages of AI for military targeting. Specifically, this chapter will cover three aspects: 1) Key considerations surrounding machine learning use for anti-personnel targeting; 2) Legal implications on the crowd of actors involved in the design, development, testing and evaluation of these technologies; and 3) The role of legal review mechanisms and possible limitations or, at least, open-ended questions surrounding their conduct in the context of military AI. These three sections are grouped together considering they raise important and decisive questions related to the design and development of AI while excluding data, which has been discussed in detail in the previous chapter.

The first section will discuss the contentious nature of machine learning use for anti-personnel targeting and whether the sense of urgency shared by many actors, States and civil society alike, to prohibit these systems is necessarily commensurate with legal implications that are distinctive of systems inducing lethality. From the embryonic stages of discussions on AI in the military domain and up to this day, high levels of attention are dedicated to the use of these technologies for anti-personnel targeting. This sense of urgency, reflected in various policymaking fora, creates pressure, for many, to push for governance frameworks that would prohibit altogether the development of these systems. It may even be that anti-personnel targeting would, at the development stages of machine learning systems, have certain unique technical considerations compared to those of anti-material targeting. Yet, this first section will argue that, from a legal standpoint, most of the same issues will arise both in anti-personnel and in anti-material targeting with slight variations only, from proportionality assessments to the implementation of the rule of distinction in ambiguous cases (i.e., in the case of dual-use

objects and civilians directly participating in hostilities). The approach taken by IHL to assumptions, an inevitable element to warfare and even more so in the context of AI use, will be used to illustrate this point. Instead, it is argued that much of the reasoning used to justify any prohibition specifically against anti-personnel targeting systems are of moral and philosophical nature rather than on legal grounds.

The second section will examine closely the different actors involved across the different stages of these technologies' development and the relevant IHL implications. General discussions surrounding the governance of AI for military targeting often refer to 'developers' and their responsibility. It is hoped that this section will not only demonstrate that what is often categorized as 'developers' is not as monolithic as it seems to be; it will also provide a more detailed overview as to what legal considerations are relevant for those involved in the development of these technologies. Admittedly, non-AI weaponry and other technologies used in military targeting usually involve as well a myriad of actors throughout their supply chain, i.e., in their development, fabrication, testing and evaluation. However, a close look at the relevant legal considerations for the crowd in the pre-deployment stages of AI for military targeting is important. First, there is growing scrutiny, by States, over the actors involved in these technologies' supply chain and the desire to understand the relevant legal implications for these actors; hence this examination would be timely in terms of interest and potential impact. Second, the complexity of these technologies' supply chain, some aspects of which can touch upon the civilian realm, will require further reflections on the ways in which the actors involved should (or may even be required to) consider IHL compliance.

The third and final section of this chapter will reflect on existing legal review mechanisms. The discussions will include a reflection on whether Article 36 legal reviews, as conducted for non-AI means and methods of warfare and as documented in the public sphere, are fit for purpose. This section will then examine the need to factor in machine learning's ever-evolving nature,

especially when it comes to reinforcement learning models, and the adequacy of Article 36 legal reviews. Finally, this section will also look into f AI technologies that may not necessarily fall within the scope of 'new means and methods of warfare' but would still play a role in decision-making behind military targeting and thus – and whether they would still require a form of legal review process.

Akin to the previous chapter, for each of the aforementioned aspects, this thesis will not only provide an overview of the key technical considerations, why and how they are relevant to military targeting. Each section will provide a discussion on select legal considerations in IHL. This thesis does not aim to provide a comprehensive and exhaustive list of all legal considerations that are relevant; instead, it will focus on a select few that are most relevant, in the eyes of the author, to each of the aforementioned aspects. The rationale behind each choice of legal focus will be outlined in greater detail in each of the relevant sections later in this chapter.

1.2. Definitions and stages in the design and development of an AI programme

There is no universal, standardized process for the design and development of AI technologies, whether it is due to varying standards and practices, or simply due to the inherent differences between each type of AI application. Nevertheless, there are a number of key steps in these pre-deployment stages that will have a high impact on the technology's performance and, conversely, limitations. Taking into account the possible variations and nuances each of them may present across applications, the following steps can be considered as of critical with high levels of impact across the technology's lifecycle, and will be relevant for the rest of the chapter.

1. **Stage 1: Definition of the problem**

The first key stage relates to the definition of the problem for the programme to solve. AI programmes are, ultimately, designed and developed to solve a specific problem – for example, helping with efficient and safe navigation (i.e., get user from point A to point B); or helping with decision-making through the analysis large volumes of data, identification of unique patterns and insights unobtainable otherwise (i.e., help the user make better decision-making). Defining the problem will help programmers to make choices with regards to the key parameters that they would need to write in their code. These parameters are the key elements and settings that configurate the algorithm and the way it runs, which will then influence algorithm performance.[385]

One example of a problem could be the need to predict the effects of a natural disaster. The ability to anticipate the likely effects and consequences of a natural disaster will then enable policymakers, humanitarian relief organizations and national authorities (e.g., armed forces) to devise and prepare plans accordingly. As such, the use of agent-based modelling could be of particular use to solve such issues, as attested by the number of research undertaken in this field.[386] An agent-based model would, indeed, consist of simulating the actions and interactions of specific agents in a system, both between each other, and with their environment. This would enable the observation of patterns in these interactions, both at macro-level (i.e., system-wide level) and the micro-level (i.e., individual-/agent-level).[387] For such models, the parameters would not only consist of defining how the environment in which the agents will interact would look like; they also consist of defining who the agents are, the known relationship between agents, the weight each agent carries individually and against other agents, etc. These

[385] Nguyen Dang and Patrick de Causmaecker 'Analysis of algorithm components and parameter: Some case studies' (Learning and Intelligent Optimization: 12th International Conference (LION 12) Kalamata, Greece, 10-15 June 2018) <https://core.ac.uk/download/pdf/196582687.pdf> accessed 16 September 2022

[386] Lu Zhuo, Dawei Han 'Agent-based modelling and flood risk managemen: A compendious literature review' (2020) 591 Journal of Hydrology <https://www.sciencedirect.com/science/article/abs/pii/S0022169420310611> accessed 18 September 2024

[387] Volker Grimm, Steven F. Railsback *Individual-based Modeling and Ecology* (Princeton University Press 2005) 10

parameters can be changed as the programme goes through training and testing; which would then enable better performance against the success metrics the programmers would have pre-defined as well.

2. Stage 2: Algorithm writing

The second key stage in the design and development of an AI programme relates to the writing of the algorithm. Once the problem has been identified, as well as the desired sets of output/outcome from the programme, developers will need to write the algorithm, the 'recipe' that will help solve the problem at hand. The developer can either write the code from scratch, less of a common practice today; use a model (e.g., Microsoft Copilot) to generate either entire or parts of algorithms – or, in other words, AI-generated algorithms, a growing practice particularly with the advent of generative AI; or build on pre-existing code that has been pre-written already, some of which could be found in open source on online platforms (e.g., Github), where source codes are shared to enable collaborative software development by programmers across the world – this practice being colloquially referred to as social coding.[388] This chapter will discussed in more detail, in a later section, the legal implications of such social coding that (in)directly feed into the development of AI-enabled software for military targeting.

3. Stage 3: Programme training

Closely linked to the latter stage is the training of the programme, and its subsequent calibrations to consolidate its performance. There are different learning methods depending on the type of programme that is being developed – e.g., supervised learning, unsupervised learning, reinforcement learning, which have been described in detail earlier in this thesis (see

[388] Laura Dabbish and others, 'Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository' (CSCW '12, Seattle, Washington, 11-15 February 2012) <http://www.jsntsay.com/publications/dabbish-cscw2012.pdf> accessed 15 September 2022

chapter 1). How the choice is made, on which learning method to use, depends on a number of factors. This can include: the problem at hand and what would be the most appropriate learning method,; and/or the data available for training and/or use, to be considered against the problem the developers wish to solve with the algorithm; and/or the greater agenda and objectives of the programmers (e.g., companies such as DeepMind may choose to develop reinforcement learning programmes mainly because of its greater strategy/focus on reinforcement learning methods, such as its Flamingo visual language model).[389]

The choice of learning method most appropriate for the problem at hand will also pre-determine the degree of human involvement desirable, or even needed, in the programme's design and development. Problems for which a supervised learning programme may work best, would then entail a need for humans to label the input data (i.e., data that feeds into the programme) as well as the output (i.e., the results of the programme).[390] For example, in order to develop a computer vision software designed to identify the different types of vehicles used by the adversary: programmers would need to train the programme on sample data of what the adversary's vehicles look like and their different types, each with their own label. For problems where reinforcement learning methods may work best (e.g., natural language processing), the programmers would have a different role than in the previous example, and instead focus on establishing when/where the programme should be rewarded or, conversely, exposed to punishment signals.[391]

4. Stage 4: Testing and evaluation

Testing the programme is another key stage in the development of an algorithm. This stage is complementary to the training stage, as results from the tests may lead to the re-calibration of

[389] Jean-Baptise Alayrac and others, 'Tackling multiple tasks with a single visual language model' (*Google DeepMind*, 28 April 2022) <https://www.deepmind.com/blog/tackling-multiple-tasks-with-a-single-visual-language-model> accessed 15 September 2022
[390] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective* (The MIT Press, 2012) 2
[391] Ibid

the programme's parameters, and for the programme to go through further training. Testing is often coupled with the evaluation, verification and validation processes; especially by the US Department of Defense (DoD), which tends to group them together as TEVV. In fact, the DoD has its own TEVV 'methods, processes, infrastructure and workforce – in order to help decision-makers and operators understand and manage the risks of developing, producing, operating and sustaining AI-enabled systems.'[392] This risk assessment would entail going through processes of testing the system's capabilities and performance against pre-defined success metrics and in different settings. For example, models and programmes for simulations would need to be tested to verify that they accurately represent 'the developer's conceptual description and specifications', 'the real world from the perspective of the intended uses', and that they are 'acceptable to use for a specific purpose.'[393] Whether or not these processes are overall fit for purpose with regards to AI technologies is a separate issue many scholars, and even government bodies such as the US Army have been attempting to address for decades.[394] For the purpose of this thesis, this chapter will specifically look at TEVV against legal considerations in a later section; and the following chapter will pay closer attention to the legal implications surrounding elements of uncertainties in TEVV stages.

5. Stage 5: Decision to deploy

The final key step in the pre-deployment stages consists of the decision to deploy the system; in other words, validating the system as ready to be taken 'out of the lab' for real-life applications. Depending on the results from them programme's tests and evaluations, whether it is to test their technical or legal reliability, it would need to be formally vetted so as to be deployed (e.g., in the battlefield for weapons systems).

[392] Flournoy, Haines, Chefitz (n 221)
[393] Ibid
[394] Ibid

While this section of the chapter seeks to be purely descriptive, the way the key stages are presented should not be interpreted as these processes being linear, nor uniform. For example, the development of AI-enabled technologies can require back-and-forth's between the training and testing stages as developers seek to consolidate the programme's performance and reliability. There may be multiple TEVV cycles for the same programme, depending on the goals that developers seek to achieve through each cycle.[395] Depending on the programme, its properties and the level of complexity, the weight/importance of each stage may differ. In addition, there may be instances where the system would need to go 'back in the lab', especially when it is found to not be as reliable anymore compared to when it was vetted as ready for deployment. This is particularly important given that AI-enabled programmes are ever-evolving by nature: their behaviours and performance are non-linear and vary in time – for example due to the data and applications they have been exposed to.[396] In other words, a programme that has been validated as lawful for certain uses may not be such anymore after some time. In the light of this issue, there is a growing body of research dedicated to developing software and mechanisms that enable the auditing of these AI programmes against legal, ethical and technological risks.[397] Such tools, in addition to regular monitoring and assessments, are critical to ensure that the AI-enabled programmes in question, and their applications and use, remain within the boundaries of the law at all times.[398]

[395] Stuart Young, Presentation for the NDIA National Test & Evaluation Conference, 'Autonomy Test & Evaluation Verification & Validation (ATEVV) Challenge Area' (3 March 2016) <https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2016/Test/Young.pdf> accessed 22 September 2022

[396] Office of the Director, Operational Test and Evaluation (US) 'FY 2021 Annual Report' (2022) <https://www.dote.osd.mil/Portals/97/pub/reports/FY2021/other/2021_DOTEAnnualReport.pdf?ver=YVOVPcF7Z5drzl8IGPSqJw%3d%3d> accessed 22 September 2022

[397] Adriano Koshiyama, Emre Kazim, Philip Treleaven 'Algorithm Auditing: Managing the Legal, Ethical, and Technological Risks of Artificial Intelligence, Machine Learning, and Associated Algorithms' (2022) Computer 55(4) <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9755237> accessed 22 September 2022

[398] The US, for example, have a dedicated office (The Office of the Director, Operational Test and Evaluation) that releases, on an annual basis, a report of operational tests and evaluation activities carried out with regards to the performance of existing systems. See: 'DOT&E Annual Reports' (*The Office of the Director, Operational Test and Evaluation*) <https://www.dote.osd.mil/annualreport/> accessed 22 September 2022

2. Machine Learning for Anti-Personnel Targeting: Legal Permissibility

2.1. Background

The term 'military AI' can refer to a host of applications, from the enhancement of logistical and organisational capabilities, to assisting with the navigation of vehicles and artillery (e.g., missile technologies), especially in cluttered environments where remote-controlled navigation may not always be possible. In addition, AI-enabled capabilities can also be integrated to assist, either directly or indirectly, with lethal decision-making (e.g., for target identification and/or proportionality assessments). Yet, from the very beginning, the international community primarily scrutinized AI applications whose effects are of lethal nature.

In 2013, the then-Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns, published a report to the HRC on lethal autonomous robotics that 'once activated, can select and engage targets without further human intervention.'[399] This was the first time AI-enabled technologies capable of engaging lethal force are brought to the attention of the international community at the UN-level. Due to resistance from the diplomatic community in discussing this 'sensitive' topic within the framework of the HRC, the issue of 'emerging technologies in the area of lethal autonomous weapons systems' was brought into the framework of the CCW when its high contracting parties decided that the Chairperson will convene the following year (in 2014) an informal meeting of experts on this topic.[400] From the embryonic stages of these high-level discussions on military AI, the lethality aspect of these technologies has been given a particular emphasis. Even to this day, the sole multilateral forum formally mandated to deliberate on AI in the military domain corresponds to the CCW's

[399] 'Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns' A/HRC/23/47 (n 41)
[400] 'Final Report' CCW/MSP/2013/10 (n 43)

discussions are centred on 'emerging technologies in the area of lethal autonomous weapons systems' – again, giving emphasis on lethality. [401]

Also in 2014, the Stop Killer Robots campaign, now formed of a coalition of over 180 member non-governmental organisations, was publicly launched.[402] The latter advocates for ensuring 'meaningful human control over the use of force' by prohibiting 'all system that use sensors to target humans' and establishing 'additional rules so that other autonomous weapon systems will be used with meaningful human control in practice'.[403] Even before the public launch of the Stop Killer Robots campaign, Human Rights Watch released a report in November 2012 entitled 'Losing Humanity: The Case against Killer Robots'.[404]

Regardless of one's position and opinion on the debate, it is clear that issues surrounding lethality, and by extension humanity, still lie to this day at the forefront of many discussions in this space. From the beginning, they have sparked growing interest on the development, deployment and use of AI in the military domain, both in diplomatic fora but also in academic literature. For a number of actors, particularly civil society organizations and States advocating for a prohibition on lethal autonomous weapons systems, lethality constitutes the red line not to cross on grounds of humanity and ethics and as such, forms the basis for their position.[405] This is particularly the position of the ICRC which states, among others, that the 'use of

[401] Although other UN bodies have formal and informal discussions and fora surrounding AI, they tend to be more general in scope and focused on non-military applications (e.g., the OECD, the ITU, etc.) Ad-hoc discussions have taken place, e.g., within the Human Rights Council, however they remain relatively informal. The CCW remains, to date, the sole multilateral forum with a formal mandate to deliberate on lethal autonomous weapons systems – a very specific subset of military applications of AI; although some NGOs and countries are informally pushing for the discussions to be taken out of the CCW, out of frustration due to the lack of concrete progress. As of November 2024, a UNGA First Committee Resolution on AI in the military domain has recently been adopted by the General Assembly, the content which, however, remains relatively factual for now with the main substantial action point being the solicitation of views by States and the multistakeholder community on this issue, to be submitted by mid-April 2025 and based on which the UN Secretary-General will formulate a report.

[402] 'About Us' (*Campaign to Stop Killer Robots*) <https://www.stopkillerrobots.org/about-us/> accessed 2 September 2022

[403] 'Our policy position' (*Campaign to Stop Killer Robots*) <https://www.stopkillerrobots.org/our-policies/> accessed 2 September 2022

[404] Human Rights Watch (n 75)

[405] This is particularly the position of the ICRC, which states, among others, that the 'use of autonomous weapon systems to target human beings should be ruled out.'

autonomous weapon systems to target human beings should be ruled out.'[406] The seemingly higher stakes raised by anti-personnel targeting, and more specifically the common argument that the development of systems specifically designed to cause lethality should be prohibited, then raises the question of whether such claims are justifiable under international humanitarian law. In other words, the question of legal permissibility of machine learning systems for anti-personnel targeting arises.

There is an increasingly rich (and critical) set of academic literature and research in the realm of ethics surrounding the development, deployment and use of machine learning whose subjects are humans. These particularly revolve around the concept of fairness, and a growing amount of important research is specifically dedicated to the impact of machine learning programmes on under-represented groups, e.g., people with disabilities, ethnic minorities, and those from the 'Global South'.[407] Research in the legal field has started to pick up the pace, particularly in international human rights law, where there is a growing set of pioneering literature.[408] Yet, there is not much clarity as to whether there are specific international

[406] 'ICRC position on autonomous weapon systems' (*ICRC*, 12 May 2021) <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems> accessed 26 September 2024

[407] See select FAccT '21's papers, including Joon Sung Park and others 'Designingg an Online Infrastructure for Collecting AI Data From People with Disabilities' (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445870> accessed 2 September 2022; Joshua L Martin 'Spoken Corporate Data, Automatic Speech Recognition, and Bias Against African American Language: The case of Habitual 'Be'' (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445893> accessed 2 September 2022; Won Ik Cho and others 'Towards Cross-Lingual Generalization of Translation Gender Bias' (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445907> accessed 2 September 2022 and Sambasivan and others 'Re-imagining Algorithmic Fairness in India and Beyond' (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445896> accessed 2 September 2022

[408] For example: Lorna McGregor, Daragh Murray and Vivian Ng 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68(2) International & Comparative Law Quarterly <https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/international-human-rights-law-as-a-framework-for-algorithmic-accountability/1D6D0A456B36BA7512A6AFF17F16E9B6> accessed 3 September 2022; Eleni Kosta 'Algorithmic state surveillance: Challenging the notion of agency in human rights' (2022) 16(1) Regulation & Governance <https://onlinelibrary.wiley.com/doi/full/10.1111/rego.12331> accessed 3 September 2022; also key institutions such as the Alan Turing Institute have a programme of work dedicated to human rights and AI: 'AI for Human Rights' (*The Alan Turing Institute*) <https://www.turing.ac.uk/ai-human-rights> accessed 3 September 2022

humanitarian law implications. In the context of our discussions, it is unclear whether international humanitarian law could be used as grounds to support arguments for a prohibition on the development of systems whose subjects are humans and, more specifically, for lethal targeting. To this end, possible use-cases of such lethal systems will be explored, before unpacking the relevant technical considerations in their development. These discussions will then lead towards a legal analysis as to whether IHL can indeed be used to substantiate arguments for a prohibition on the development of these systems – which this section generally argues against, with the exception of two specific considerations.

2.2. Potential applications in military targeting

Anti-personnel targeting is to be understood, for the present section, as military operations whose objective is to neutralize, i.e., produce lethal effects on one or more than one individual target. A number of AI applications are being developed, or have even been reportedly used particularly in recent and on-going conflicts, to directly enable or at least, to partially support the conduct of such operations. This section will take stock of some of these applications, albeit in a non-exhaustive manner due to space constraints. Specifically, three categories of applications will be discussed:

1. **Intelligence, surveillance and reconnaissance**:

The first group of AI applications for anti-personnel targeting pertains to ISR. Whether it is in the kinetic or virtual realm, there is widespread recognition that artificial intelligence offers tremendous opportunities, for the armed forces, to collect and process data at unprecedented scale and speed.[409] This capability will, naturally, feed into intelligence collection for militaries, including the identification of potential targets.

[409] Afina (n 2)

The US NSA's Skynet is one of such programmes, which has been discussed in more detail earlier in this thesis.[410] Another example corresponds to DARPA's URSA programme, which we also discussed earlier in this thesis.[411] The latter is specifically designed to collect and process vast amounts of data from various sources in the battlefield, thus in theory supporting the commander with advanced ISR capabilities in environments as complex as densely populated areas where risks of civilian casualties are at the highest, and target identification is most difficult.

Beyond the kinetic realm, AI has also been used by armed forces to conduct ISR activities in the digital space. Armed forces have for example used these technologies to establish patterns and trends in social media posts in times of civil unrest, and identify individuals predicted to incite or trigger violence.[412] These technologies are also being used to scan public posts on social media as part of counter-terrorism efforts, with natural language processing systems capable of processing text data in various languages.[413]

Most recently, in the context of the current armed conflict in Gaza, the IDF's use of AI technologies for intelligence processing and to assist with target identification has drawn significant media coverage. Specifically, the Lavender system is used for intelligence collection to support anti-personnel targeting operations, drawing from reportedly 'hundreds and thousands' of datapoints including the belonging to WhatsApp groups with a known militant.[414] In response, the IDF has issued a press release, online, presenting their targeting processes in the context of using systems akin to Lavender. Specifically, the IDF argues that

[410] Chapter 1, Section 1.2.
[411] Chapter 2, Section 2.2.3.
[412] Insights shared by a participant to a UNIDIR roundtable the author took part in, in September 2024, discussing different use-cases where AI has been deployed in security and defence. The meeting was held under the Chatham House rule and this information may thus not be attributed.
[413] Insights shared by a participant to the regional consultations co-organized by the Netherlands and the Republic of Korea and hosted by Türkiye in Istanbul, specifically covering States within South East Europe, the Middle East, the South Caucasus and Central Asia in May 2024. The meeting was held under the Chatham House rule and this information may thus not be attributed.
[414] Abraham (n 207)

these systems are 'merely tools that help intelligence analysts cross-reference' and constitute one source, among others, that feed into target identification – a process done separately to target engagement, a decision allegedly made by another person.[415]

2. Decision support systems:

The second group of AI applications for anti-personnel targeting corresponds to decision support systems. The latter can be 'characterized as computerized tools that are designed to assist humans at different levels in the chain of command to complete decision-making tasks.'[416] These tasks can range from the choice of means and methods of warfare to employ in a specific situation, to that of target engagement in the battlefield (e.g., moment of attack). This is particularly the case for AI-enabled battle management software, enabling the centralisation of data, forces and operations across all domains of combat (e.g., for NATO – air, space, maritime, land and cyber). Leveraging yet again AI's ability to collect and process data at scale and at speed from across domains, these systems can then suggest courses of action to enable mission success.[417]

One specific example is that of Palantir's AI Platform for Defence (AIP for Defense). The system does not only collect, process and analyse data at scale and at speed – reportedly live, providing the user (i.e., the commander) with an overview of the battlefield in real-time with enhanced situational awareness. The system is also integrated with a large language model that enables the user to 'interact' with the system by submitting prompts, but also for the system to 'interact' with the user. Based on the demonstration video of AIP for Defense, it can, for

[415] 'The IDF's Use of Data Technologies in Intelligence Processing' (n 270)
[416] ICRC and Geneva Academy 'Expert Consultation Report on AI and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts' (2024) ICRC <https://www.geneva-academy.ch/joomlatools-files/docman-files/Artificial%20Intelligence%20And%20Related%20Technologies%20In%20Military%20Decision-Making.pdf> accessed 27 September 2024
[417] Grooms (n 215)

example, notify the user of 'anomalous military activity detected', which the user can then respond to by asking for more details.[418] The system will subsequently provide the user with satellite imagery and an analysis of the data and patterns it has detected, followed by a series of back-and-forth's between the system and the user that may, at times, request for different options to support their decision-making.[419] While the demonstration video shows the targeting of military equipment (i.e., anti-materiel targeting), such systems could in principle be used for anti-personnel targeting too. Alex Karp, Palantir's Chief Executive Officer, has claimed that Palantir's software is 'responsible for most of the targeting in Ukraine' – thus not necessarily ruling out the possibility of anti-personnel targeting.[420]

3. Autonomy in weapons systems:

The third group of relevant applications relates to autonomy in weapons systems deployed for anti-personnel targeting. These applications would thus include, among others, the 'lethal autonomous weapons systems' being debated, for over a decade now, within the UN and specifically in the context of the CCW.[421] Broadly speaking, however, this autonomy can correspond to different levels, from basic weapon functions (e.g., navigation, calculating which munition to use) to more advanced functions (e.g., swarming capabilities, loitering, predictive maintenance). Autonomous swarm drones have particularly garnered increasing attention in recent years, notably due to recent advances in swarming in the civilian realm which, thus, spark concerns on their weaponization.[422] In swarming, a system would enable multiple

---

[418] Palantir, *Palantir AIP | Defense and Military* (YouTube, 25 April 2023) <https://www.youtube.com/watch?v=XEM5qz__HOU> accessed 27 September 2024

[419] Ibid

[420] Bergengruen (n 206)

[421] 'Lethal Autonomous Weapon Systems (LAWS)' (*United Nations Office for Disarmament Affairs*) <https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/> accessed 17 November 2022

[422] See for example Ingvild Bode, Anna Nadibaidze '25 Autonomous Drones' in James Patton Rogers (ed) *De Gruyter Handbook of Drone Warfare* (Vol. 4, De Gruyter, 2024) <https://www.degruyter.com/document/doi/10.1515/9783110742039-025/pdf?licenseType=restricted> accessed 27 September 2024

systems (e.g., drones) to 'communicate' with one another, synchronize and act in a coordinated manner to establish specific formation and for effective navigation.[423] Beyond drones, swarming capabilities could, in principle, also be applied in non-aerial environments (e.g., at sea).[424] Another way in which AI could enable weapons systems for anti-personnel targeting is through computer vision capabilities, which may assist with target identification and selection, or even the decision to engage with the target with minimum to no intervention by a human operator. This for example relates to target detection through imagery obtained on CCTV embedded in weapons systems.[425]

2.3. Technical Issues and Considerations

Building on the previous survey of AI applications for anti-personnel targeting, this sub-section will consider technical issues and considerations related to the development of these technologies. In fact, by considering whether there are technical issues and considerations that may be specific, to a certain extent, to the development of these technologies, reflections can then be made on their legal implications. Besides clarifying points of the law, this will ultimately inform our assessment as to whether or not there are specific legal grounds on the development of AI technologies that would warrant a blanket prohibition on lethal applications.

[423] Jawad N Yasin and others 'Navigation of Autonomous Swarm of Drones Using Translational Coordinates' (Advances in Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness, PAAMS Collection 18th International Conference, L'Aquila, Italy, October 7–9, 2020) <https://link.springer.com/chapter/10.1007/978-3-030-49778-1_28> accessed 17 November 2022; see also Fawaz Alsolami and others 'Development of Self-Synchronized Drones' Network Using Cluster-Based Swarm Intelligence Approach' (2021) 9 IEEE Access <https://ieeexplore.ieee.org/abstract/document/9373310> accessed 17 November 2022

[424] HD Hyundai Heavy Industries, in partnership with Palantir, presented during the 2nd edition of the REAIM Summit in Seoul, Republic of Korea, 9-10 September 2024, their Tenebris uncrewed surface vessel concept that would have swarming capabilities at sea. For a report on Tenebris, see: Aaron-Matthew Lariosa, 'HD HHI And Palantir Reveal Tenebris Unmanned Surface Vessel' (*Naval News*, 16 May 2024) <https://www.navalnews.com/naval-news/2024/05/hd-hhi-and-palantir-reveal-tenebris-unmanned-surface-vessel/> accessed 27 September 2024

[425] Jose L. Salazar González and others 'Real-time gun detection in CCTV: An open problem' (2020) 132 Neural Networks <https://www.sciencedirect.com/science/article/abs/pii/S0893608020303361> accessed 17 November 2022

The first considerations relate to the identification of the problem to be solved; in other words, establishing what one wants the algorithm to do and solve. Knowing what the problem is, and the related task can be relatively straightforward. This could, for example, be 'the positive identification of targetable fighters in an urban setting through computer vision'. What is much less straightforward would be to ascertain how to address the task and what would be achievable to meet this end. This complication is notably due to the assumptions and biases that will heavily shape the solution designed to solve the problem at hand.

For example, the identification of targetable fighters in urban settings through computer vision will imply the need for developers to code specific parameters as to what is assumed to positively indicate that someone is indeed a targetable fighter under the law of armed conflict. These assumptions, set by the developers, can for example correspond to straightforward indicators such as the open bearing of arms and uniforms. As discussed in the previous chapter, specifically in the previous chapter, proxy indicators can also be used to assess targetability and may, in fact, at times be necessary in the absence of direct indicators.[426] These computer vision programmes can for example increase the likelihood of an individual's targetability on the basis of their sex (i.e., male) of 'military age', echoing the profiling made by Bosnian Serb forces of those 'who posed a potential military threat' in the context of the killings in Srebrenica.[427] An added layer of verification must be embedded in the programming of such systems to make sure that the fighter identified as targetable is not protected, e.g., if they are *hors de combat*, e.g., as they aresurrendering. This could, yet again, imply the need to use proxy data in addition to direct indicators, such as behavioural patterns indicating an absence of resistance, environmental context (e.g., defencelessness in non-combat environments), as well

[426] Chapter 2, section 2.3

[427] *Prosecutor v Radislav Krstic* (Judgment) ICTY IT-98-33-T (2 August 2001) <https://www.icty.org/x/cases/krstic/tjug/en/krs-tj010802e.pdf> accessed 17 November 2022

as the eventual tactical disadvantage in which the fighter is in (e.g., no realistic means of escape or resistance).

In this context, beyond using proxy data, there is also the question of how much weight do these data carry in the final assessment – a decision that can either be set by the developer, or set by the programme (particularly if it operated on a reinforced learning basis), or a combination of the two. In any event, even if the weighing of proxy indicators is done by the programme, it may still be subject to re-calibration based on testing and evaluation metrics that the human testers would have set on what they assume to be benchmarks for success. As such, it is clear that assumptions made by humans will heavily influence the development of AI systems for military targeting – unsurprisingly so, given that assumptions are an inherent part of military planning and decision-making especially in combat where uncertainty and other variables may come in, which will be further discussed later.[428]

Facial recognition programmes are also known as being subject to biases and assumptions. For example, there is a wealth of research conducted on the use of facial recognition technologies by law enforcement agencies, and how the programmes are fed with inherently biased data reflecting deeply rooted societal issues, with a particular emphasis on racism. The Met Police in London has even disclosed issues with their facial recognition technology programmes in identifying women as part of its trials period; and its performance rate was even worse when it comes to people of colour.[429] As discussed in the previous chapter, in the context of armed conflicts, such biases may also be an issue for facial recognition technologies tasked with the positive identification of potential targets.[430]

[428] See Chapter 4, section 2
[429] Metropolitan Police (UK), National Physical Laboratory 'Metropolitan Police Service Live Facial Recognition Trials' (2020), <https://www.met.police.uk/SysSiteAssets/media/downloads/central/services/accessing-information/facial-recognition/met-evaluation-report.pdf> accessed 17 November 2022
[430] Chapter 2, section 2.2.3

The question of human assumptions and biases shaping the programme's parameters boil down to one key issue: the need to quantify social constructs and metrics, some of which may fundamentally not be quantifiable, at all, in an accurate way.[431] For example, as discussed in the previous chapter, there have been a number of qualitative studies undertaken by social scientists attempting to identify 'predispositions' of members of certain NSAGs based on the environment surrounding them, including socio-cultural, political and psychological contexts.[432] As enlightening as these studies can be, not only are these findings inherently complex and of qualitative nature. Even in their qualitative state, they do not necessarily align with IHL targeting criteria; and there is only so much that can be translated into quantitative metrics that could then be coded into an algorithm. These parameters are indeed highly context-dependent, and a lot of the key agents and considerations identified in these studies would be interconnected and interdependent in various ways. In the light of these technical considerations deeply relevant to AI applications designed to assist with anti-personnel targeting, it is therefore important that we conduct an assessment on the subsequent legal implications.

2.4. The Permissibility of Machine Learning Development for Anti-Personnel Targeting

The present section is the culmination of this first substantive segment of Chapter 3. Building on the discussions above, this section will ultimately examine the extent to which IHL rules can be incorporated in the development stages of AI for anti-personnel targeting, specifically focusing on the technical characteristics and implications of using machine learning for such applications. Following a deep-dive on international humanitarian law's approach to the issue of assumptions, an integral component to machine learning, the argument will be that the

[431] Abigail Z. Jacobs, Hanna Wallach 'Measurement and Fairness' (FaccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada, 3-10 March 2021) <https://dl.acm.org/doi/abs/10.1145/3442188.3445901> accessed 11 November 2022
[432] See Florez-Morris (n 327) and Gill & Horgan (n 328)

development of AI-enabled anti-personnel targeting is not necessarily of unlawful nature in general, with a number of boundaries and considerations that will be discussed (i.e., the customary prohibition for weapons of a nature to cause superfluous injury or unnecessary suffering, and specific dispositions with regards to the protection of persons *hors de combat*).

2.4.1. Assumptions and International Humanitarian Law Considerations

The previous section discussed the decisive role of human assumptions and subsequent biases in the development of AI technologies for anti-personnel targeting. Specifically, these assumptions are central to the quantification of qualitative social constructs. This sub-section seeks to unpack two legal considerations relevant to these discussions.

1. **Rule of distinction**

The first relates to the rule of distinction, which makes a clear differentiation between who would be lawfully targetable and, conversely, who would be non-targetable. AI-enabled, or at least AI-assisted distinction assessments means the programme's inherent assumptions and biases will bear a certain level of influence. The extent to which these assumptions and biases can cause issues with regards to compliance with the rule of distinction will depend on a number of factors, namely:

- The assumptions and biases in question: Depending on what these assumptions and biases are, they may not necessarily be an issue and, if anything, could even foster compliance. For instance, if a computer vision-based target identification system is programmed to assume that anyone lying flat on the ground is *hors de combat* and thus cannot be targeted, this protective assumption is actually more aligned to the rule of distinction's protective stance in case of doubt regarding an individual or an object's

status.[433] On the flipside of the coin, programming said system to assume that individuals of a certain age and sex increase their chances to be considered as targetable fighters may be more problematic. In the same vein, on biases, whether they may be problematic or not will also depend on their definition. The general and neutral definition of a bias, i.e., as a derivation from a standard, may not necessarily and in and of itself be problematic. Meanwhile, the legal implications of biases prone to be problematic (i.e., reflective of unjustified inclination) will be more nuanced. This issue is relevant from the development stages of AI technologies (e.g., through the overfitting of systems into categories deemed as 'too rough') and may have subsequent legal implications.[434] On the rule of distinction specifically, overfitting may prevent the system from capturing the nuances brought by certain variables that would affect distinction assessments such as, for example, combatants that may be considered as 'sick and wounded'.

- The type of technology and its inherent purpose: Assumptions and biases may raise more, or less, compliance issues depending on the technology's application and its intended purpose. For example, the assumptions calibrated into an ISR capability that is being developed to identify potential targets for the planning stages is likely going to carry weight on the commander's decision-making at a further stage in the kill-chain; however, this weight may be less than that of ATR systems embedded into weapons systems designed for dynamic targeting, where assumptions and biases are more likely

[433] Melzer (n 336)
[434] Prratek Ramchandani 'Random Forests and the Bias-Variance Tradeoff' (*Medium*, 10 October 2018) <https://towardsdatascience.com/random-forests-and-the-bias-variance-tradeoff-3b77fee339b4?gi=6cdd6a5535b8> accessed 28 September 2024, as cited in Ingvild Bode, Ishmael Bhila 'The problem of algorithmic bias in AI-based military decision support systems' (*Humanitarian Law & Policy*, 3 September 2024) <https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based-military-decision-support-systems/> accessed 28 September 2024

going to have repercussions on compliance due to the technology being 'closer' to target engagement.

- Pre-deployment considerations and practices: Variations in the legal implications of assumptions and biases will also be influenced by the considerations and practices adopted in the technology's pre-deployment stages. This includes the extent to which, for example, reflections have been made and measures have been adopted to ensure harmful biases will not be reflected in data labelling. Another example is the extent to which testing and evaluation benchmarks take into account possible compliance issues stemming from assumptions and biases. Generally, benchmarks and testing and evaluation metrics are, in themselves, are also subject to assumptions and biases. Developers of search engine tools, such as Google, have their definition of what constitutes 'high quality' results; these will inherently depend on the assumptions of those involved in developing and maintaining the tools in question on what would indeed constitute results of 'high quality.[435] Similarly, in the context of military AI development, what is considered as an effective and reliable system will depend on assumptions and biases: What a lawyer would consider as effective and reliable in a system may not be the same as that of an ethicist or someone coming from an operational perspective. What is certain is that these communities are working towards addressing their respective issues, worries and concerns from early on in these technologies' lifecycle. The more and the earlier good practices are adopted to foster compliance, the more compliance issues will be pre-empted or at least, known and anticipated.

[435] Danny Sullivan 'An overview of our rater guidelines for Search' (*Google The* Keyword, 19 October 2021) <https://blog.google/products/search/overview-our-rater-guidelines-search/> accessed 17 November 2022

- Use: A number of considerations surrounding to the use of the system will also be relevant with regards to addressing the compliance issues raised by assumptions and biases. This includes, for example, the intended level of human judgment and oversight conferred and that have subsequently been encoded into the system's functioning and interface. For, if the commander has been informed of the assumptions and biases that have formed part of a decision support system's development, e.g., through the programme's system card. In case the commander sees possibly unlawful or at least contentious outputs as a result of these assumptions and biases, their ability to either challenge them or at least, request other courses of action will have a direct impact on the outcome's legality. Another consideration is the number of sources used to feed into the system's final output. If, for instance, a target identification system relies on a combination of computer vision and speech analysis from intercepted calls, at the very least the assumptions and biases forming part of the final output will not come from one single source – although this may also complicate further traceability, which may have implications on the ability to investigate in case of violations, as discussed later in Chapter 4.[436]

Hence, and in the light of the highly context-dependent nature of such assessments, it would be difficult, if not impossible to substantiate claims that assumptions and biases would inherently make the development of military AI unlawful. It can even be argued that most, if not all algorithms are inherently based on a set of assumptions. In other words, given that AI systems are fundamentally developed to solve a problem, their development will depend on what is assumed to be a solution to said problem. Feed recommendations on online platforms operate on what is assumed to be what the user is looking for, whether it is based on what is

[436] Chapter 4, section 3.4

assumed to be their preferences or, conversely, the type of content they are less likely to engage with based on their activities.[437]

Similarly, target identification systems operate on what is assumed to make an individual or object a target or not. With regards to IHL compliance, targetability assessments will depend on a number of questions that, while relatively objective, will have answers that will inevitably be influenced by assumptions such as:

- *Are they a targetable fighter?* This question would constitute the first layer of distinction assessments that generally differentiates civilians from combatants. Indicators of someone's status as a targetable fighter, or not, will heavily rest on a series of assumptions such as, for example, that combatants would be clearly identifiable through their uniform.[438] Yet this may not always be the case, particularly in non-international armed conflicts taking place in densely populated areas, where identity concealment is of tactical interest, or simply that identifying targetable fighters (e.g., members of a non-State armed group) would be more challenging and not as straightforward.[439] For these situations, assumptions will play an even greater role.
- *If they are a targetable fighter, are they hors de combat?* This question will be necessary to answer, providing the rule of distinction extends beyond the differentiation between combatants and civilians. Combatants are not always targetable – particularly when they are wounded, sick, or surrendering, i.e., *hors de combat*. Again, what would indicate someone's status as *hors de combat* would be based on a series of assumptions

[437] For TikTok's algorithm and its set of user assumptions, for example, see: Daniel Klug and others 'Trick and Please. A Mixed-Method Study on User Assumptions About the TikTok Algorithm' (WebSci '21: Proceedings of the 13th ACM Web Science Conference 2021) <https://dl.acm.org/doi/abs/10.1145/3447535.3462512> accessed 17 November 2022

[438] Cordula Droege 'Get off my cloud: cyber warfare, international humanitarian law, and the protection of civilians' (2013) 94(886) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/get-off-my-cloud-cyber-warfare-international-humanitarian-law-and-the-protection-of-civilians/72114EF6E71757FAB2B37E7DE918B2BB> accessed 17 November 2022

[439] Ibid

– e.g., minimal movement, position, as well as body language indicating surrender. These assumptions should also be evaluated against operational considerations, e.g., what assumptions can be made based on the situation and what would be feasible to evaluate these assumptions? For example, what assumptions can be made regarding the targetability of a fighter based on ISR data collected via aerial drones, as opposed to the information that can be collected from troops on the ground? What are the measures that would be feasible to undertake to verify the veracity of these assumptions including as to whether or not they are *hors de combat*?

- *If they are not a targetable fighter, are they specifically protected?* This question will seek to determine whether or not the person in question, if not a combatant, is specifically protected e.g., because they are medical or religious personnel. Medical personnel may be more easily identifiable through the wearing of certain symbols (e.g., red cross, red crescent or red diamond – although this must also take into account local contexts where, in certain countries, the red crescent would be more prominent than the red cross, for example). Religious personnel may not always be as straightforward, and what is assumed to indicate an individual's status as such may be biased and not reflective of local realities.
- *If they are not a combatant, are they civilians directly participating in hostilities?* If the identified person is neither a combatant, nor specifically protected, the rule of distinction would also encompass assessments as to whether or not a civilian is directly participating in hostilities. Yet again, this would imply relying on a series of assumptions, e.g., considering certain actions that would make them directly participating to hostilities or, conversely, considering certain actions as not amounting to directly participating to hostilities. An added layer of complexity is further added

depending on the context in which such assessments are made, and the impact they may have on the targetability of those civilians directly participating in hostilities.

In addition to the assumptions that, will inherently form and shape the answer to those questions through the model's output, the commander's own assumptions will form part of their risks assessments and, ultimately, whole decision-making surrounding the targeting process. In fact, one could even argue that assumptions are an inevitable aspect of warfare notably due to the dynamic, messy and uncertain nature of conflict – an aspect that will be discussed in further detail in Chapter 4.[440] In short, assumptions lie at the heart of conflict, and so it is the case for IHL.

This makes it difficult to substantiate the hypothesis that AI programmes holding assumptions and biases would be inherently incompatible with the rule of distinction. Instead, an approach that could be more productive would rest on the premise that developers need to undertake the appropriate measures to:

- Identify the assumptions and biases encoded into the system under development,
- Minimise risks of harm and of non-compliance, which includes the eventual need to decide not to deploy a particular technology if risks of non-compliance are too high; and
- Ensure a level of transparency about the assumptions, biases and other limitations that bear an impact on the programme to the end-users or at least, document known assumptions and biases e.g., through system cards.

The latter point is particularly important, given that end-users are unlikely to have been involved in the development process of the programme, and/or may not have the technical literacy necessary to fully appreciate the way it has been shaped and its eventual boundaries.

[440] Chapter 4, section 2

Transparency on the assumptions surrounding the programme will not only inform procurement, testing and evaluation processes to determine whether or not the programme is purchasable, at all, and if so whether it is deployable and under what circumstances. This will then also enable the end-users to fully appreciate the breadth of factors that need to be taken into account while they conduct their own decision-making based on the programme's results, and the subsequent risks assessment.

2. Rule of precautions

Second, as established earlier in the thesis, the legal requirement to undertake precautionary measures to minimise risks of harm to civilians and/or damage to civilian objects also extends to the design and development stages of an AI-enabled programme. In addition to the precautionary measures needed to minimise such harms through appropriate training and testing processes, another key consideration relates to the danger of over-estimating the programme's capabilities. How can this risk be minimised? This aligns perfectly with the point argued in the previous paragraph on the need for transparency with regards to the programme's capabilities and boundaries – including the assumptions it is conditioned to. In fact, such transparency will enable end-users to fully appreciate the performance of the programme and its limitations. Informed decision-making does not only consist of the amount of information the commander has: it is also about them knowing what they do not know. A lack of awareness on these limitations would increase risks of over-estimating the programme's capabilities and thus, the commander's decision-making will be flawed and potentially lead to violations of the law.

Coming back to the examples raised in the previous paragraph: if developers do not disclose, to the end-users, that their facial recognition programme's performance is poorer for women and people of colour. The end-users might not take the necessary precautionary measures to exclude women and people of colour from being scanned by the machine. This in turn could

lead to the misidentification of individuals from these groups of people – and subsequently expose them to risks of harm. Risks of harm will always be present, especially in conflict where many uncertainties lie. Some risks, however, could be minimised by taking the appropriate steps, hence the rule of precautions is there. A full disclosure on a programme's limitations to the end-users (and identifying these limitations) is one of those steps and would constitute part of the precautionary measures to foster downstream compliance post-deployment of the system.

2.4.2. Anti-Personnel and Anti-Materiel Targeting: Source of Inherently Similar Yet Nuanced Legal Issues

The previous sub-section provided the reader with a deep-dive on how IHL approaches the question of assumptions and biases; its main conclusion being that these two do not unnecessarily equate to the unlawfulness of a system's development. In parallel, another observation emerges: Whether or not these systems are anti-personnel or anti-materiel, assumptions will play an equally influencing role and will have implications on both applications. The question of whether an individual is a targetable fighter – and the assumptions surrounding what makes a fighter as targetable, resonates with the question of whether an object constitutes a military objective (i.e., whether by its nature, location, purpose or use, the object makes an effective contribution to military action, and whether its partial or total destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage).[441]

Generally speaking, the legal considerations surrounding anti-personnel and anti-material targeting are, actually inherently similar despite a few nuances, as attested by the issue of assumptions. Proportionality assessments are framed by the same rule for both cases, i.e.,

[441] 'Rule 8: Definition of Military Objectives' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule8> accessed 29 September 2024

whether the incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof expected to be caused by the launch of an attack be excessive in relation to the concrete and direct military advantage anticipated.[442] The same applies for compliance with the rule of precautions in attack. With regards to the rule of distinction, some nuances are applied when it comes to the status of individuals and objects of ambiguous status, i.e., civilians directly participating in hostilities and dual-use objects respectively. These nuances primarily boil down to the inherently different nature of individuals and objects; both approaches, however, are inherently similar with limited temporality and specific conditions to be met in order for the target to be lawful.

In the same vein, prohibitions are also very much alike. Whether the system under development is for anti-personnel or anti-material targeting, if, through testing and evaluation, its intended use is found to be inherently incompatible with the applicable dispositions set by international humanitarian law, its deployment would be unlawful. For example, if a computer vision-based target identification system under testing is found to exhibit racial biases and is then unable to distinguish between targetable fighters and civilians due to the subjects' skin colour, its deployment would unlikely be lawful in the first place. Similarly, if another or even the same computer vision programme-based target identification system struggles to recognize medical workers as such because they use a red crescent symbol instead of a red cross, it is unlikely to be deployable especially in contexts where the red crescent is more prevalent. Ultimately, the rule that the choice of means and methods of warfare in armed conflict is not unlimited is all-encompassing, applying both to anti-personnel and anti-materiel targeting.[443]

Aside from the slight nuances, there are two additional considerations to this observation: the prohibition of weapons systems of a nature to cause superfluous injury and unnecessary

[442] API art 51(5)(b), see also 'Rule 14: Proportionality in Attack' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule14> accessed 29 September 2023
[443] API, art 35(1)

suffering; and the specific dispositions with regards to the treatment of persons *hors de combat* and particularly with regards to the customary prohibition of violence to life. A statutory prohibition on the development of AI systems for anti-personnel targeting contravening any of these two customary rules would be legally justified, due to their inherent link to persons and not objects.

First, the prohibition of weapons, projectiles and material and methods of warfare of a nature to cause unnecessary suffering and superfluous injury is enshrined in Additional Protocol I.[444] Generally accepted as of customary nature, this prohibition has been ascertained as a cardinal principle by the International Court of Justice in its Advisory Opinion on the *Legality of the Threat or Use of Nuclear Weapons*.[445] A couple of international treaties have also been drafted and adopted on the basis of this prohibition, namely the 1980 Convention on Certain Conventional Weapons and the 1997 Mine Ban Treaty. This prohibition is centred around the suffering and injury caused by the weapons, projectiles, material and methods of warfare in question, both inherently experienced by humans only. In this sense, an international treaty crystalizing this prohibition on systems for anti-personnel targeting that, by nature, would cause unnecessary suffering and superfluous injury, would be justifiable; although such text will need to be specific on what would concretely constitute unnecessary suffering and superfluous injury in the context of AI development, deployment and use.

Second, there are specific dispositions with regards to the treatment of protected persons, e.g., combatants and more generally targetable fighters found to be *hors de combat*. In fact, while the rule of distinction practically allows the killing of targetable fighters in armed conflict, they may not be the subject of direct attack if they are found to be *hors de combat*, i.e., wounded or sick. This status comes with a host of specific provisions as to how they should be treated,

[444] API, art 35(2)
[445] *Legality of the Threat or Use of Nuclear Weapons* (n 69) [78]

including the customary prohibition of violence to their life – a prohibition generally applicable to all protected persons.[446] In fact, Common Article 3 of the 1949 Geneva Conventions prohibits 'violence to life and person, in particular murder of all kinds' of both civilians and those *hors de combat*. Furthermore, 'wilful killing' of protected person is listed as a grave breach in all four 1949 Geneva Conventions; both 1977 Additional Protocols I and II reinforce this prohibition.[447] As such, and echoing the previous paragraph, an international treaty crystalizing this prohibition on systems specifically designed to inflict violence to the life of civilians and those *hors de combat*, would be justifiable. Whether or not States are looking into developing such systems in the first place with this very specific purpose is another question.

3. The (Indirect) Role of 'Developers' and Relevant Implications

This section will be dedicated to examining closely the different actors involved in the development of these technologies. Specifically, as discussions on military AI governance increasingly pay attention to 'developers' and in particular their accountability and responsibility, this section will seek to examine the relevant legal considerations for their implication and role in the development of AI for military targeting. To this end, this section will first demonstrate that what is often referred to as 'developers' is not a monolithic bloc but rather, a diverse group of individuals involved in the development of AI systems with diverse roles and responsibilities. Subsequently, these 'developers' will have different legal considerations relevant to their respective work, duties and activities, which this section will seek to identify and discuss.

This discussion is a critical one to have, not only because of the increased scrutiny, by the international community over the 'developer's' accountability and responsibility in the context

[446] 'Rule 89: Violence to Life' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule89> accessed 29 September 2024

[447] API, art 75(2)(a) and Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the protection of victims of non-international armed conflicts (Protocol II) (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 609 (APII), art 4(2)(a), respectively

of AI development for military targeting. It is also important to acknowledge the complexity of these technologies' supply chain, with many aspects that can touch upon the civilian realm (e.g., collection and processing of 'civilian' data for military applications – although what would be considered as 'civilian' data is another question). Similar arguments could, to a certain extent, be raised for non-AI capabilities, e.g., the extent to which chip manufacturers would know whether the chips will be used for laptops used by military personnel. However, as yet again policy deliberations in this space are paying a particular attention to the 'developers', it is hoped that this section will provide some food-for-thought on the relevant legal considerations to these discussions.

3.1. Who is the 'Developer'?

While much attention is given to 'the developer', their responsibility, and accountability, there is no clear definition as to who 'the developer' refers to. In fact, depending on the context, what is often referred to as 'the developer' in international discussions can either correspond to a range of professional groups with different titles and conducting different duties as part of the model's development. In fact, the development of complex AI systems generally requires collaborative work between a large number of developers. There is no single developer that would be responsible for the development of AI systems from A to Z: A range of AI developers can be involved in the development of a programme at various stages and under different capacities.

While there is no standardized set of titles associated with specific duties (e.g., 'software engineer', 'data scientist', 'AI research scientist'), what is often referred to as 'developers' in international discussions tend to refer to those working in either of the following:

- **Data**: The training and testing of AI systems will usually require the conduct of a number of data-related tasks by so-called 'developers'. These include, but are not

limited to: data collection (i.e., the sourcing and collection of relevant data such as text, images or audio); data labelling (i.e., for supervised models); and data processing.

- **Coding**: 'Developers' can also refer to those involved in the algorithm's coding which, in itself, can involve a number of tasks. Those often referred to as 'AI architect', for example, are usually in charge of conceptualising the bigger picture with regards to the programme's general purpose and define its subsequent development pipeline. AI engineers are also involved in the development process of developing AI solutions to the pre-identified problem in question.
- **Training**: Processes related to the training of the system and subsequent modification of the code for optimization form part of the tasks completed by those often referred to as 'developers'.
- **Testing and evaluation**: Finally, 'developers' can also refer to those specialists specifically working on the testing and evaluation of systems against a set of benchmarks, which can take many forms (e.g., on the basis of legal compliance, alignment with ethics, and/or operational effectiveness).

While this list is non-exhaustive, it shows how diverse the role of 'developers' can be across all stages of AI development; indicating how 'crowded' the process can be.

3.2. Role of Developers

Developers play a key role in shaping the performance of AI-enabled technologies, as they are the ones with the technical expertise required to drive the process behind their development 'in the lab' (i.e., prior to deployment) – from the conceptualisation/design phase to the coding, training, testing, evaluation, verification and validation stages. Of course, these developers would be working together with other stakeholders in identifying the programme's overarching goal (i.e., the problem to solve and how), as well as evaluation benchmarks and success metrics with regards to the programme's performance. These stakeholders, including representatives

from the procuring entity (e.g., armed forces) and the to-be users of the system, while they could influence the overall strategic purpose of the programme and its functionalities, they may not have the technical literacy required to be at the core of the programme's development; hence a discussion on the developers behind the programmes is necessary.

These developers could be 'in-house', as in the Army and the wider Ministry in charge of defence and security issues would have the financial, human and technological resources required to conduct research in the field of AI and eventually develop their own programmes. For example, the United States established in 2018 the JAIC, embedded within the US Armed Forces, and which was eventually absorbed, along with the Defense Digital Service and the Office of Advancing Analytics into the CDAO. One of the JAIC's primary objectives consists of delivering, to the US, AI-enabled capabilities addressing key missions; which implies the DoD's development of their own AI technologies in-house. In addition, the US DoD has also got DARPA, its specialised agency generally known for undertaking research and development to advance existing capabilities (e.g., AI to enhance air combat capabilities, battlefield medicine, etc.), as well as research in more 'frontier', exploratory areas (e.g., hypersonics, quantum computing, neurotechnology, etc.)[448] DARPA has got extensive resources, in-house to the DoD, ranging from human resources and technical literacy, to the necessary financial resources, infrastructure and equipment to conduct their work (e.g., with large compute power).

When not directly developing the programmes, such in-house technical expertise could also be used to coordinate and enable public-private partnerships with external organisations and contractors in the development of AI capabilities. For example, DARPA has a wide range of partners from industry as part of its Air Combat Evolution programme.[449] The latter seeks to 'automate air-to-air combat and build human trust in AI'; with partners including Heron

[448] 'Our Research' (*DARPA*) <https://www.darpa.mil/our-research> accessed 29 September 2022
[449] Ryan Hefron 'Air Combat Evolution (ACE)' (*DARPA*) <https://www.darpa.mil/program/air-combat-evolution> accessed 29 September 2022

Systems, Lockheed Martin, PhysicsAI etc. taking part in a number of activities, including 'dogfight' trials simulating air combat pitting the programme of one company against another.[450] In the UK, the Ministry of Defence (MoD) has for example partnerships with the Alan Turing Institute, tapping into its in-house analysts and data scientists, along with the intelligence community, including the Government Communication Headquarters (GCHQ) and the Security Service (MI5) to implement projects enhancing the UK's AI capabilities.[451] Aside from a handful of States, most States would actually be heavily dependent on such public-private partnerships to procure AI-enabled military capabilities due to limited in-house technological, financial and/or human resources.[452]

In addition, even developers who are not part a governmental agency, nor a partner organisation, could somewhat be involved in the development of AI-enabled technologies for military targeting. This is particularly because of the general tendency, by the AI community, to upload, and use, source codes on open-source platforms (e.g., Github) with the aim to contribute to the development of greater AI research while also seeking feedback from other programmers part of the same platform/community. The benefits of open-source software have been recognised in the military sphere, both in terms of accessing these codes to develop, strengthen and deploy certain capabilities and in terms of gaining feedback on unclassified code.[453] Similarly, software libraries available in open-source that are not necessarily developed for military purposes could, however, be used as such.

This was the case, for example, of the now defunct Project Maven, which involved Google providing application programme interfaces (APIs) from TensorFlow to develop computer

[450] 'AlphaDogFight Trials Go Virtual for Final Event' (*DARPA*) <https://www.darpa.mil/news-events/2020-08-07> accessed 29 September 2022
[451] 'Defence and security' (*The Alan Turing Institute*) <https://www.turing.ac.uk/research/research-programmes/defence-and-security> accessed 29 September 2022
[452] Afina (n 2)
[453] Defense Digital Service 'Code.mil: An Open Source Initiative at the Pentagon' (*Medium*, 13 March 2017) <https://medium.com/@DefenseDigitalService/code-mil-an-open-source-initiative-at-the-pentagon-5ae4986b79bc#.i78how76u> accessed 29 September 2022

vision capabilities, which in turn would have supported military targeting.[454] Even though, in the case of Project Maven, Google had a direct contract with the DoD (although details surrounding their agreement and the role each party plays in this partnership remains unclear), such use of TensorFlow attests to the growing interest, by the military, to use open-source AI platforms to strengthen their AI capabilities.[455] Reliance on external AI platforms by the military means that developers that were/are involved with the platforms in question will in turn be involved, even indirectly, with the development of military AI.

Again, these points show how 'crowded' the development stages of an AI-enabled system can be, even those in the military realm where such endeavours are generally shrouded with secrecy. In fact, there may be AI systems that the military would end up using, but where it would be impossible to re-trace all individuals involved in the greater supply chain and development processes of the system in question (e.g., because of the use of open-source source codes); when generally such processes in the military are strictly limited and those involved in such processes are thoroughly vetted and subject to security clearance.

3.3. Legal discussion

Due to the number of individuals involved in the development processes of AI systems, questions arise surrounding the legal implications of such a 'crowd' in these upstream stages of the technology's lifecycle. This section will specifically look at the following questions: the common assumption that this crowd results in an 'accountability gap'; issues surrounding the use of civilian 'codes' found in open-source by the military; and the implications of involving such a crowd with regards to risks of bias and cognitive limitations.

---

[454] Kate Conger 'Google Is Helping the Pentagon Build AI for Drones' (*Gizmodo*, 6 March 2018) <https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533> accessed 29 September 2022

[455] Marijn Hoijtink, Anneroos Planqué-van-Hardeveld 'Machine Learning and the Platformization of the Military: A Study of Google's Machine Learning Platform TensorFlow' (2022) 16(2) International Political Sociology <https://academic.oup.com/ips/article/16/2/olab036/6562417> accessed 30 September 2022

3.3.1. Legal accountability

There is an assumption shared by many that because the development of AI involves such a crowd, it exacerbates the risks of a supposed accountability gap.[456] In the light of the multi-layered and complex supply chain prior to the deployment of AI technologies, questions indeed arise with regards to the developers' liability and at what level. Worries pertaining to accountability gaps are generally a recurrent theme in discussions surrounding military AI, whether it is because of the number of individuals involved in the development of these technologies, or because of the 'human out of the loop' context where certain decisions are automated.[457] In fact, it is one of the key driving arguments for advocacy groups seeking to advance the prohibition of autonomous weapons in international discussions since their embryonic stages.[458]

Questions subsequently arise on who should be held accountable in the chain of command (e.g., the commander, the developers, etc.); how to factor in issues surrounding technical literacy and expectations with regards to the machine's reliability; state responsibility for internationally wrongful acts; the conduct of investigations in the light of IHL violations; etc.[459] As a result, there is a growing, and rich, literature seeking to shed light on these diverse issues

[456] This point has been raised a number of times during research roundtables and workshops on this topic, all under the Chatham House rule. For a discussion on developers' liability, see: Afonso Seixas-Nunes, SJ 'Scapegoats! Assessing the Liability of Programmers and Designers for Autonomous Weapons Systems' in Jan Maarten Schragen (ed) *Responsible Use of AI in Military Systems* (CRC Press, 2024) <https://library.oapen.org/handle/20.500.12657/89843> accessed 8 October 2024

[457] Horowitz (n 78)

[458] 'Mind the Gap: The Lack of Accountability for Killer Robots' (*Human Rights Watch*, 9 April 2015) <https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots> accessed 14 October 2022

[459] See for example Robin Geiß 'Autonomous Weapons Systems: Risk Management and State Responsibility' (3rd CCW meeting of experts on lethal automous weapons systems (LAWS), Geneva, 11-15 April 2016) <https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/Geiss-CCW-Website.pdf> accessed 14 October 2022; Daniele Amoroso & Guglieelmo Tamburrini 'Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues' (2020) 1 Current Robotics Reports <https://link.springer.com/article/10.1007/s43154-020-00024-3> accessed 14 October 2022; Giovanni Sartor and Andrea Omicini 'The autonomy of technological systems and responsibilities for their use' in Nehal Bhuta, Susanne Beck, Robin Geiß (eds) *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press 2016) <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/books/autonomous-weapons-systems/autonomy-of-technological-systems-and-responsibilities-for-their-use/9C47BF93BCF884E3D4735C95509790BD> accessed 14 October 2022

surrounding human responsibility over the use of military AI.[460] In fact, the question of algorithmic accountability is so pertinent across all possible applications of AI that it stretches beyond the military sphere: in the US, for example, a bill had been formally introduced as early as in 2019 on algorithmic accountability for automated decisions in the realm of consumer protection and commerce.[461]

A lot of these discussions, however, focus on accountability for the implications of AI technologies downstream, i.e., post-deployment. There is increasing attention given to the need for anticipatory thinking in terms of human accountability with regards to algorithmic decision-making, particularly under the term of 'responsible AI' by design. One of the possible approaches to this question is framing algorithmic accountability within the framework of human rights: whether it is from a legal perspective or more from a general, philosophical and ethical standpoint, the argument would be to include and streamline human rights considerations from the design and development stages to shed light on accountability questions.[462]

In the context of military AI and IHL, the sheer number of individuals and entities involved in the development of these technologies should not pave the way towards claims of an 'accountability gap' resulting from confusion as to who should be held accountable, what for, and how far can a programme be retraced to. In fact, the accountability question could be approached from two perspectives: 'public' state responsibility; and 'private' individual responsibility for violations of IHL, including those amounting to international crimes.

---

[460] Marta Bo, Laura Bruun and Vincent Boulanin 'Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS' (2022) SIPRI <https://www.sipri.org/sites/default/files/2022-10/2210_aws_human_responsibility.pdf> accessed 14 October 2022

[461] Algorithmic Accountability Act of 2019, H.R. 2231, House – Energy and Commerce, 116th Congress (2019) <https://www.congress.gov/bill/116th-congress/house-bill/2231> accessed 14 October 2022

[462] For example, see McGregor, Murray and Ng (n 408) and Vinodkumar Prabhakaran and others 'A Human Rights-Based Approach to Responsible AI' (2022) ArXiv <https://arxiv.org/abs/2210.02667> accessed 14 October 2022

State responsibility

The question of state responsibility with regards to military AI is a question that has attracted some attention already. Whether it is in the context of the CCW's discussions or in academic writing, these analyses, however, remain limited in number and most generally follow the tendency of focusing legal analyses on violations 'post-deployment' in the battlefield.[463] Furthermore, these analyses tend to cover state responsibility under public international law, instead of focusing exclusively on state responsibility over IHL violations in the context of military AI. This point deserves further reflections: In fact, it is commonly accepted as customary law that states can be held accountable for violations of the law, both in the context of international and non-international armed conflicts.[464] This means that all IHL considerations discussed in the present thesis can lead states to being liable for their eventual violation in the context of designing, developing and testing AI systems for military targeting – despite the eventual confusion caused by the 'crowd' involved in these processes.

The International Law Commission (ILC)'s Articles on State Responsibility, and more specifically through its Article 4 , provides for the consideration of the conduct of 'any State organ' as 'an act of that State under international law, whether the organ exercises legislative, executive, judicial or any other functions, whatever position it holds in the organization of the State, and whatever its character as an organ of the central Government or of a territorial unit of the State.' Furthermore, Article 5 also provides for the conduct of 'a person or entity which is not an organ of the State under article 4 but which is empowered by the law of that State to

[463] See for example Geiß (n 459); Daniel N Hammond 'Autonomous Weapons and the Problem of State Accountability' (2014-2015) 15 Chicago Journal of International Law <https://heinonline.org/HOL/LandingPage?handle=hein.journals/cjil15&div=26&id=&page=> accessed 20 October 2022; although more recently there has been a very helpful analysis on state responsibility over the development of military AI (i.e., pre-deployment), see Berenice Boutin 'State responsibility in relation to military applications of artificial intelligence' (2023) 36(1) Leiden Journal of International Law <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/leiden-journal-of-international-law/article/state-responsibility-in-relation-to-military-applications-of-artificial-intelligence/1B0454611EA1F11A8B03A5D2D052C2BE> accessed 22 November 2024

[464] 'Rule 149: Responsibility for violations of International Humanitarian Law' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule149> accessed 20 October 2022

exercise elements of the governmental authority', also considering their act as an act of the State under international law 'provided the person or entity is acting in that capacity in the particular instance.' These provisions have been echoed by the ICRC which establishes, in its study on customary law and building on the ILC's work, a list of entities, seemingly non-exhaustive, whose violations of IHL could hold a state responsible. The list includes the state's own organs, including its armed forces, but also those individuals and groups and entities 'empowered by the state to exercise elements of governmental authority'; or 'under its direction or control, by the state'; or whose violations are acknowledged and adopted as own.[465] End-users of AI technologies for military targeting will ultimately fall within one of these categories (i.e., the armed forces); and the State will be responsible for eventual violations of IHL any of these entities would commit as a result of deploying and using the technology in question. It is therefore in the state's best interests to oversee and ensure that those technologies are as compliant 'by design' as possible, so end-users (and ultimately the state) can rest on the assumption and reassurance that these technologies are indeed safe and reliable for their intended uses.

Beyond end-users, and as this thesis also seeks to narrow its focus to pre-deployment, one of the organs, groups and entities whose actions could entail State responsibility corresponds to the research laboratories and centers where the procurement, development, testing and evaluation of these technologies occur. In the US, this would be DARPA for example, while in the UK this would be the Defence Science and Technology Laboratory (Dstl) and, in France, the research unit of the *Agence ministérielle pour l'intelligence artificielle de la défense*.[466] Whether it is through their in-house developers team to those in charge of initiating, managing

[465] Ibid

[466] 'Sébastien Lecornu lance la stratégie ministérielle sur l'intelligence artificielle' (*Ministère des Armées*, 8 March 2024) <https://www.defense.gouv.fr/actualites/sebastien-lecornu-lance-strategie-ministerielle-lintelligence-artificielle> accessed 18 November 2024

external relationships, or those evaluating compliance in the procurement of new capabilities, their actions could indeed hold the State responsible. This could be, for example, the case of testing and evaluation processes that would fail to demonstrate the implementation of the duty to respect and ensure respect. Not only could their actions for in-house development of AI capabilities constitute grounds for State responsibility; their actions related to external entities also qualify. For example, in the purchase of AI-enabled surveillance drones from a private company, these agencies will need to stress-test and evaluate these systems' compliance with IHL as part of its procurement process. Not only is this important due the increased volume of public-private partnerships in the space of military AI, with for example OpenAI acknowledging its collaboration with DARPA.[467] It is also important to ascertain State responsibility over the purchase, by these agencies and/or its other organs such as the armed forces, of off-the-shelves capabilities. States will typically have little to no influence over the development of these technologies as they will be off-the-shelves, sometimes the only type of capabilities smaller States can afford due to their lower costs.[468] Thus, to ensure compliance, it would be important for States to ensure that they allocate the resources necessary to at least test and evaluate these off-the-shelves capabilities while procuring these systems and before their deployment, even though they may not have been able to intervene earlier in the technology's lifecycle.

Individual responsibility

With regards to individual responsibility, much of the discourse presented by advocacy groups, but also and more generally proponents of a prohibition on autonomous weapons systems, raise concerns with regards to a supposed 'accountability gap'. Many of these arguments are based

[467] 'OpenAI's approach to AI and national security' (*OpenAI*, 24 October 2024) <https://openai.com/global-affairs/openais-approach-to-ai-and-national-security/> accessed 18 November 2024

[468] This insight draws from the five regional consultations the author has attended on Responsible AI in the Military Domain, which the Netherlands and the Republic of Korea held in the first half of 2024 in Singapore, Istanbul, online, Nairobi and Santiago. While these consultations were held under the Chatham House rule, key findings may be found at: Afina (n 2)

on the premise that the unpredictable nature of AI systems will, inherently, stem into such gap, supposedly absolving individuals (e.g., the commander) in being held accountable and responsible.[469] Yet, as an emerging but growing number of scholars are pointing out, the issue is much more nuanced and if anything, virtually under all circumstances can individual accountability and responsibility be upheld with regards to the development of AI for military targeting. If anything, arguments justifying the existence of an accountability gap primarily fall under moral and philosophical premises, if not driven by political motives. The earliest accounts presented by Human Rights Watch, advocating for a ban on so-called 'killer robots', specifically revolve around the supposed accountability gap, featuring prominently in their arguments and even more so in the title of their 2015 report.[470]

Some legal scholars have pointed out the issue of impunity stemming from the use of autonomous weapons systems that, subsequently, result in an accountability gap.[471] Yet, impunity and, more generally, the question of enforcement of international law is not an issue that is exclusive to autonomous weapons. In fact, international law enforcement corresponds to a much wider issue that is met across disciplines and is of increased evidence, from human rights and the law of international organizations to international criminal law, transitional justice, as well as corporate accountability – especially in the context of the growing decolonization of legal scholarship.[472]

---

[469] Thompson Chengeta 'Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law' (2016) 45(1) Denver Journal of International Law and Policy <https://heinonline.org/HOL/P?h=hein.journals/denilp45&i=9> accessed 11 October 2024

[470] 'Mind the Gap: The Lack of Accountability for Killer Robots' (n 458)

[471] Thompson Chengeta 'Autonomous Weapon Systems: Accountability Gaps and Racial Oppression' in Christopher Sabatini (ed) *Reclaiming Human Rights in a Changing World Order* (Brookings Institution Press, 2022) <https://www.chathamhouse.org/2022/10/reclaiming-human-rights-changing-world-order/9-autonomous-weapon-systems-accountability> accessed 11 October 2024

[472] See, respectively, Joyce De Coninck 'The EU's Human Rights Responsibility Gap – Exhuming Human Rights Impunity of International Organizations' (2024) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4706514> accessed 11 October 2024; Michelle Burgis-Kasthala, Barrie Sander 'Conteporary International Criminal Law After Critique: Towards Decolonial and Abolitionist (Dis-)Engagement in an Era of Anti-Impunity' (2024) 22(1) Journal of International Criminal Justice, <https://academic.oup.com/jicj/article/22/1/127/7695263> accessed 11 October 2024; Farah Mihlar 'Global Models, Victim Disconnect and Demands for International Intervention: The Dilemma of Decoloniality and

Perhaps those systems do exacerbate issues with regards to impunity and enforcement. In fact, it has been argued that military AI technologies are actually symptoms, highlighting the absence of an 'international accountability mechanism for most unintended civilian harms in armed conflict.'[473] Yet again, these issues are more of a systemic nature instead of being unique to the development, deployment and use of AI technologies for military targeting. If the issue is to ensure the individual responsibility of anyone post-deployment, the responsibility of military commander is explicitly enshrined both in statutory and customary IHL.[474] Ultimately, the commander's decision to deploy and use an AI system for military targeting operations will also fall under their responsibility. This will entail, among others, the conduct of risk assessments with regards to the technology's functionalities, known limitations, operational needs and the situation in the battlefield; and there are many measures that can be adopted from the pre-deployment stages of these technologies to support such risks assessments (and thus, foster, compliance) – including through the extensive documentation of the technology's testing and identified flaws, as well as the robust training of users-to-be.[475] Post-deployment, the accountability and responsibility of commanders will then be framed by international criminal law. Further discussions on the exact and many nuances specific to the commander's responsibility from the use of these technologies, especially with regards to international criminal law, would fall outside of this thesis' scope. Considerations include, for example, the

Transitional Justice in Sri Lanka' (2024) 16(3) Journal of Human Rights Practice, <https://academic.oup.com/jhrp/advance-article-abstract/doi/10.1093/jhuman/huae024/7748714> accessed 11 October 2024; and Flávia do Amaral Vieira, 'The Struggles for Corporate Accountability in the UN: A Global South Perspective' in Thamil Venthan Ananthavinayagan, Amritha Viswanath Shenoy (eds) *The Wretched of the Global South: Critical Approaches to International Human Rights Law* (Springer, 2024) <https://link.springer.com/chapter/10.1007/978-981-99-9275-1_12> accessed 12 October 2024

[473] Rebecca Crootof 'AI and the Actual IHL Accountability Gap' (2022) CIGI <https://www.cigionline.org/articles/ai-and-the-actual-ihl-accountability-gap/> accessed 12 October 2024

[474] API, art 87 and 'Rule 153: Command Responsibility for Failure to Prevent, Repress or Report War Crimes' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule153> accessed 12 October 2024

[475] On documentation, see for example Florian Königstorfer, Stefan Thalmann 'AI Documentation: A path to accountability' (2022) 11 Journal of Responsible Technology <https://www.sciencedirect.com/science/article/pii/S2666659622000208> accessed 12 October 2024

intricacies of establishing the cognitive element of *mens rea* in the context of human-machine teaming, as well as evidentiary difficulties in proving casualty.[476]

Additional questions arise for instances where civilians are indirectly involved in the development of AI-enabled technologies for military targeting; for example, because they uploaded a source code on an open access platform (e.g., Github) and it was subsequently used for developing such technologies. Another example would be large language models (e.g., GPT-3 by OpenAI) used for non-intended uses (e.g., to support disinformation campaigns such as a deepfake video ordering an attack). Although this last example falls beyond the scope of military targeting per se, the question of AI technologies used for non-intended uses is generally of interest and relevance today. Can developers be held liable for eventual violations of the law – e.g., from the technology developed based on their source code? One possible approach to this issue is again through international criminal law, which somewhat grapples with this issue for many decades now and extending beyond the realm of artificial intelligence. Specifically, the term of the 'Problem of Many Hands' has been coined in 1980 to describe 'the general difficulties experienced in criminal law when the assumption of a lone, solitary figure making decisions is broken.'[477] This issue is further exacerbated due to the dual-use nature of many, if not most AI technologies and the possibility of civilian AI being re-purposed for military applications. Yet again, this issue is not novel and in fact, ultimately, accountability and responsibility for battlefield use will rest with the commander who will, in their risks assessment, have taken into account the dual-use nature of the AI technologies in use (or at least, in consideration for use). Unless *mens rea* can be established, in the first place, by

[476] See Marta Bo 'Autonomous Weapons and the Responsibility Gap in light of the *Mens Rea* of the War Crime of Attacking Civilians in the ICC Statute' (2021) 19(2) Journal of International Criminal Justice <https://academic.oup.com/jicj/article-abstract/19/2/275/6181757> accessed 12 October 2024; and Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies* (Springer, 2024) 269

[477] Dennis F. Thompson 'Moral Responsibility of Public Officials: The Problem of Many Hands' (1980) 74(4) American Political Science Review <https://doi.org/10.2307/1954312> accessed 12 October 2024, as cited in Kwik (n 476) 270

developers to build and train systems for military targeting, it would be difficult from the outset to establish their accountability and responsibility. Expanding beyond legal discussions, the dual-use nature of AI technologies and their potential re-purposing for military purposes – whether by the armed forces or even non-State armed group – is an issue many States are grappling with and as such, they seek the right balance in their policy approach to ensuring continuous accessibility of open-source AI on the one hand, and mitigating proliferation and misuse risks on the other hand.[478]

3.3.2. Biases and Cognitive Limitations: Due Diligence as a Solution to Accountability and Responsibility?

Decision-making processes can be extremely complex; not only because of the nature of the matter to be decided on, but also because of other factors influencing, directly or indirectly, these processes. Human judgment is particularly affected by 'perceptions, beliefs, culture, religion and education', and more generally complex limitations surrounding behavioural decision-making.[479] The influence of these factors could be particularly amplified by biological factors in high-stress situations, where those at the heart of decision-making would be exposed to certain hormones due to the surrounding environment (e.g., adrenaline due to high-stress). While those on the one end of the spectrum (i.e., arguing for a prohibition on military AI) advocate for a ban due to the lack of humanity from the development and use of such technologies; others would argue that this human element can constitute more of a liability – and could paradoxically lead to inhumane outcomes in armed conflict. In fact, from the dawn

[478] Afina, Grand-Clément (n 260)

[479] See in the context of nuclear decision-making Beyza Unal and others 'Uncertainty and complexity in nuclear decision-making' (2022) Chatham House <https://www.chathamhouse.org/2022/03/uncertainty-and-complexity-nuclear-decision-making/> accessed 28 October 2022; behavioural insights are also used to analyse and better understand decision-making processes in other disciplines, including economics, public health and other areas of policy design and policy-making. See for example: Colin R Kuehnhanss 'The challenges of behavioural insights for effective policy design' (2019) 38(1) Policy and Society <https://academic.oup.com/policyandsociety/article-abstract/38/1/14/6403979> accessed 28 October 2022; Lara Carminati 'Behavioural Economics and Human Decision Making: Instances from the Health Care System' (2020) 124(6) Health Policy <https://pubmed.ncbi.nlm.nih.gov/32386789/> accessed 28 October 2022

of discussions surrounding autonomous weapons systems, arguments were set forth somewhat in favour of these systems due to the issues surrounding emotions and how they can lead to 'horrendous suffering', such as in the case of Rwanda, the Balkans, Darfur and Afghanistan.[480] Cognitive, and physical limitations have indeed been argued as risk factors for excess and exacerbating risks of human losses in armed conflict.[481]

Regardless of where one stands on this debate, one thing is certain: 'automated' decision-making processes are never truly deprived from the influence of biases and cognitive limitations, i.e., these 'human' factors proponents of military AI argue these technologies lack of. In fact, those individuals involved in the design, development, testing, evaluation, verification and validation of the technology in question will inevitably bear some influence on its performance through their own cognitive limitations and biases. This influence can go even further than the developers in themselves. As discussed in the previous chapter, historical data that reflects certain, established societal biases used as training and testing datasets will inevitably affect the way predictive algorithms will work.[482] For example, predictive policing algorithms are particularly, and often, put in evidence as an example of algorithms reflecting systematically and institutionally biased data with regards to policing practices in society (i.e., by reflecting deeply-rooted issues in policing such as institutional racism, classicism, etc.)[483] It has been established by a number of research that these biases can lead to harmful outcomes, in particular towards minority and vulnerable groups which, as discussed in the previous chapter, can have subsequent implications on IHL compliance such as the rule of distinction.[484]

---

[480] Schmitt & Thurnher (n 155)

[481] Robin Geiß 'The International-Law Dimension of Autonomous Weapons Systems' (2015) Friedrich Ebert Stiftung <https://library.fes.de/pdf-files/id/ipa/11673.pdf> accessed 28 October 2022

[482] One example is the inheritance of historical racial biases reflected in datasets – as discussed in Section 2.2.3, Chapter 2

[483] See for example Adriane Chapman and others 'A Data-driven analysis of the interplay between Criminological theory and predictive policing algorithms' (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3533071> accessed 28 October 2022

[484] Section 2.2.3, Chapter 2

As such, the biases and cognitive limitations of those involved, directly and indirectly, in the development of military AI technologies will bear influence over their performance. One way of approaching and addressing this issue is through the concept of due diligence. More specifically and in this context, due diligence is to be approached as the 'extent to which states must prevent, halt and redress different types of harm or injury' or in other words, framing the behaviour and conduct of states with diligence.[485] As such, and in a nutshell, due diligence is not in and of itself a standalone rule or obligation in international law but rather, it is a standard that may be used to assess States' compliance with various primary obligations.

There are a couple of main reasons why due diligence, as an approach to addressing biases and cognitive limitations, could work. First, it would be impossible to expect algorithms to be completely deprived of biases and all kind of influence from developers and all behind the data and the programme. This is particularly the case for bias in data, which may reflect deeper, underlying societal issues and human biases reflected in these datasets, such as institutional racism, classicism, and/or gender bias – as discussed in the previous chapter.[486] If anything, it has been argued that in order to rid systems of harmful biases, there is research looking at amplifying other biases that, instead, would quash these harmful biases.[487] As such, a legal obligation to remove all biases from the programme would practically be unrealistic. However, what would be realistic and, in fact, useful in fostering compliance in the development of these technologies, would be for States to establish risk reduction measures through the lens of due

[485] Antonio Coco and Talita Dias ''Handle with care': due diligence obligations in the employment of AI technologies' in Robin Geiß and Henning Lahmann (eds) *Research Handbook on Warfare and Artificial Intelligence* (2024, Edward Elgar Publishing) <https://www.elgaronline.com/edcollchap/book/9781800377400/book-part-9781800377400-19.xml> accessed 12 October 2024

[486] Risks of bias stemming from well-established biases embedded into data is generally well established and recognized; see for example Department for Science, Innovation & Technology (UK), Centre for Data Ethics and Innovation 'Review into bias in algorithmic decision-making' (2020) <https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making> accessed 7 November 2022

[487] Jungsoo Lee and others, 'Revisiting the Importance of Amplifying Bias for Debiasing' (Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington DC, USA, 7-14 February 2023) <https://ojs.aaai.org/index.php/AAAI/article/view/26748> accessed 12 October 2024

diligence. Specifically, by establishing a specific standard of care through the lens of due diligence, States can help translate the legal requirements set by international law into concrete recommendations and good practices for the development, testing and evaluation of military AI and, ultimately, help prevent harm.

Second, approaching the issue of biases and cognitive limitations through the lens of due diligence helps addressing the need to establish and maintain States' responsibility and accountability in the development of military AI. Specifically, biases and cognitive limitations are two inherently socio-technical issues that, ultimately, will require the upstream intervention of developers; however, concerns are often raised by the fact that these developers, often being from the private sector, are not subjects of international law, thus creating a supposed accountability gap. Due diligence can provide a useful framework to transpose international law obligations to corporations and ground their practices in relation to primary obligations. Specifically, due diligence provides for States' responsibility if they had 'actual or constructive knowledge (that is, it either knew or should have known) that an act contrary to the rights of another state emanated from its territory'.[488]

The implications are two-fold. First, States will need to ensure that industries within their own territory are not developing military AI technologies that would, by nature or through their use, violate international law. Similarly, States with the in-house capacity to develop their own military AI capabilities must ensure that the technologies under development will not, by nature or through their use, violate international law. Second, States procuring military AI capabilities from foreign industries, meant for use by its own governmental bodies (e.g., its own armed forces), will also need to ensure the appropriate protective measures are in place to foster compliance; this assessment should generally be taken throughout the procurement process of

[488] Coco and Dias (n 485)

these technologies, from the requirements set in the call for tenders to their testing and evaluation, including through the conduct of legal reviews.

In international humanitarian law, the essence of due diligence, i.e., the 'extent to which states must prevent, halt and redress different types of harm or injury', can be found in a number of considerations, including: the duty to respect and ensure respect, the rule of precautions, and limitations to means and methods of warfare. On the duty to respect and ensure respect of IHL, the latter is crystalised in common article 1 to the 1949 Geneva Conventions. This provision not only sets obligations for States in the use of AI technologies in military targeting; the duty to ensure respect also means that States must undertake the necessary precautions to uphold compliance from the pre-deployment stages of AI technologies for military targeting.

Furthermore, due diligence somewhat echoes the core of the rule of precautions. By mandating parties to armed conflicts to proactively adopt and implement measures to take constant care during the conduct of the military operations, and minimizing risks of harm, the rule of precautions essentially rests on anticipatory risk management and reduction. While the rule of precautions is generally more focused on the conduct of military operations, we have also established that this rule can also be applied to the pre-deployment stages of AI technologies for military targeting.[489] Specifically to mitigate compliance risks from biases and cognitive limitations, precautionary measures could for example be adopted by mandating pre-deployment assessments of the presence and impact of such biases, e.g., through audits and red-teaming. These assessments would, ideally, need to be undertaken prior to each operational environment, especially as biases and their impact can vary from one theatre to another depending on the varying geographical, demographic, and architectural realities, to cite a few. Finally, with regards to limitations on means and methods of warfare, their choice is not

[489] See discussions on temporality in Chapter 2, Section 2.1.2.1

unlimited. Not only should this choice minimise risks of civilian loss and damage by virtue of the rule of precautions.[490] They must also not be of a nature to cause superfluous injury or unnecessary suffering.[491] As such, States must ensure, from the procurement and pre-deployment stages of military AI, that these capabilities would not, by nature, cause such injury or suffering; and the contexts in which they are meant for deployment are such that they are there are no alternative choices that would be available to minimize risks of civilian loss and damage.

In complement to international humanitarian law, international human rights law can provide a useful framework too for Specifically, in the field of business and human rights, the Guiding Principles on Business and Human Rights were unanimously endorsed in the context of the HRC, which requires corporations to exercise 'human rights due diligence process to identify, prevent, mitigate and account for how they address impacts on human rights.'[492] This due diligence process notably includes the need to identify and assess 'actual or potential adverse human rights impacts' that may be caused, or contributed to by the enterprise's own activities or products; 'integrating findings from impact assessments across relevant company processes and taking appropriate action' accordingly; 'tracking the effectiveness of measures and processes to address adverse human rights impacts in order to know if they are working'; as well as 'communicating on how impacts are being addressed and showing stakeholders – in particular affected stakeholders – that there are adequate policies and processes in place.'[493] Such approach would merit further exploration in the context of AI for military targeting. There

[490] API, art 57(2)(a)(ii)(a)
[491] 'Rule 70: Weapons of a Nature to Cause Superfluous Injury or Unnecessary Suffering' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule70> accessed 14 October 2024
[492] 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework' (2011) OHCHR <https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf> accessed 7 November 2022
[493] 'Corporate human rights due diligence – identifying and leveraging emerging practices' (*OHCHR*) <https://www.ohchr.org/en/special-procedures/wg-business/corporate-human-rights-due-diligence-identifying-and-leveraging-emerging-practices> accessed 7 November 2022

is, however, no standardised process on actual and potential impact assessment; nor there is an explicit mandate for parties to the armed conflict to undertake the appropriate measures to monitor the effectiveness of steps taken to minimise those risks – if not through the duty to respect and ensure respect under IHL. There is no obligation neither, under IHL, for parties developing and deploying military AI to publicise the policies and processes in place they are undertaking to address risks and impacts. In a way, for practical reasons, these limitations are quite understandable. States and their military contractors cannot be expected to fully disclose details surrounding their AI technologies due to secrecy and in order to maintain a strategic advantage over adversaries – especially when it comes to risks, limitations and shortfalls from their own technologies. This secrecy also explains why, in the military sphere, there is generally a lack of universal standards on processes surrounding the innovation, research and development of technologies – even for 'established' processes such as Article 36 legal reviews, there is no standardised way of conducting these reviews as they remain within the discretion of states, most probably due to confidentiality issues.

It has been argued that due diligence cannot be applied in relation to negative obligations, i.e., when these obligations are negative commands (e.g., to not target civilians) because states would implement that obligation only by reaching the specific negative result demanded by that negative obligation (e.g., there are no civilians killed as a result of direct targeting).[494] This approach, however, does not factor in the positive actions a state can take in order to achieve the negative result demanded by the negative obligation: for example, by specifically choosing high-precision weapons for targeting missions in densely populated areas to not cause any civilian casualties and/or damage to civilian objects (or, at the very least, minimise risks of these results). The ICJ has established that 'a state cannot be under an obligation to succeed,

[494] Marco Longobardo 'The Relevance of the Concept of Due Diligence for International Humanitarian Law' (2019) 37 Wisconsin International Law Journal <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3570423> accessed 7 November 2022

whatever the circumstances, in preventing' a specific event; however they do have an obligation to 'employ all means reasonably available to them, so as to [reach the desired result] so far as possible'.[495] This ruling was in relation to the commission of genocide; or rather, the customary, negative command to not commit genocide and undertake all necessary measures in preventing, as much as possible, the commission of genocide. Here, there is an obligation for states to undertake active steps and measures to implement a negative command, the implementation of which shall be reflected through a negative result (i.e., no genocide committed). It is however important to note here that the negative obligation to refrain from genocide is distinct from the positive obligation to prevent it, thus demonstrating that both positive and negative obligations can, and should, be applied in concert.

The same reasoning could thus be applied to IHL, with a strong caveat: in principle due diligence can indeed be applied to all IHL obligations, as even negative commands do imply active measures and steps behind in order to facilitate the implementation of these obligations. It is however important to stress that such argument should, by no means, undermine strict negative obligations under IHL such as the duty to not take hostages, or the duty to not directly attack civilians by subjecting them to an additional standard of due diligence. Risks would indeed arise that the notion of due diligence would be instrumentalized to justify non-respect with these strict negative obligations and subsequently result in 'compliance-washing' practices.

As such, due diligence could work as one of the solutions to addressing issues from the biases and cognitive limitations of programmers and those directly, and indirectly, behind the technology in question. Due diligence could indeed step in to mitigate and reduce risks of non-compliance with IHL as a result of these biases and cognitive limitations. Echoing the ICJ's

[495] *Case Concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia & Herzegovina v Serbia & Montenegro*) (ICJ Judgment) 2007 <https://www.icj-cij.org/public/files/case-related/91/091-20070226-JUD-01-00-EN.pdf> accessed 7 November 2022 [430]; as cited in Longobardo (n 494)

judgment quoted earlier, this is not about succeeding the prevention of harmful biases, but it is rather about doing everything reasonably possible and available to reduce risks from these biases so far as possible. There is a growing and rich literature on approaches and techniques available to address issues stemming from algorithmic biases – for example, anticipatory bias correction methods, bias bounties, and other best practices such as fitting a model to data (instead of exclusively letting the model be shaped by the data) – from which lessons can be learned in the context of military AI.[496] These measures undertaken 'upstream' in the technology's lifecycle could indeed be considered as part of implementing due diligence measures to minimise risks of IHL violations – all possible without compromising the state and industry's secrecy surrounding the technology they are developing if the appropriate measures are taken (e.g., bias bounties for non-classified codes and databases).

Far from being the perfect solution, the present discussions on due diligence do not imply that the mere undertaking of such measures is sufficient to ensure compliance. Due diligence should rather be seen as a framework that can contribute minimizing risks of non-compliance; however, it does not mean that they are, in and of themselves, sufficient alone for a system to be found in compliance. Due diligence can, for example, help reduce the risks of a system being prone to making mistakes on race or gender; yet, it would not necessarily be enough to ensure such system can be used in compliance with international law, although the risks may have been reduced in practice provided due diligence is indeed conducted in good faith and not for the purpose of compliance-washing.

[496] See for example: Abdulaziz A Almuzaini 'ABCinML: Anticipatory Bias Correction in Machine Learning Applications' (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3533211> accessed 7 November 2022; Ira Globus-Harris, Michael Kearns, Aaron Roth 'An Algorithmic Framework for Bias Bounties' (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3533172>; Chapman (n 483)

4. Legal reviews and keeping up with ever-learning machines

4.1. Article 36 reviews: a background

As discussed in the introduction of this chapter, AI-enabled systems go through thorough processes of evaluation, verification and validation before they receive the 'stamp of approval' for deployment. In addition to technical checks, these systems would/should also go through legal review processes, colloquially referred to as Article 36 reviews. The name comes from the fact that the obligation to undertake these legal review processes has been enshrined in Article 36 of the 1977 Additional Protocol I to the 1949 Geneva Conventions (API). Beyond its status as a statutory obligation for State parties to API, whether it qualifies as of customary nature is subject to much debate and remains, to this day, not universally accepted.[497]

This obligation to undertake legal reviews 'in the study, development, acquisition or adoption of a new weapon, means or method of warfare' is aligned with the customary rule that the choice of means and methods of warfare is not unlimited.[498] In fact, these Article 36 reviews have two dimensions.

The first dimension corresponds to evaluating, and establishing whether the weapon, means or method of warfare under scrutiny would be unlawful by nature (i.e., any use of it would constitute violations of IHL). This is particularly with regards to two aspects, both of which have been argued as customary by nature. First, they must not be 'of a nature to cause superfluous injury or unnecessary suffering'.[499] This framing has been particularly adopted in defining the scope of the CCW – its full title being the 'Convention on Prohibitions or

[497] See Natalia Jevglevskaja 'Weapons Review Obligation under Customary International Law' (2018) 94 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1724&context=ils> accessed 22 September 2022, and Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (CUP 2017) rule 110, para 2

[498] API, art 36; on the customary limits surrounding the choice of means and methods of warfare, see: 'Rule 17: Choice of Means and Methods of Warfare' (*ICRC IHL Databases)* <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule17> accessed 23 September 2022

[499] 'Rule 70: Weapons of a Nature to Cause Superfluous Injury or Unnecessary Suffering' (n 491)

Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects'.[500] Second, these weapons, means and methods of warfare must not be, by nature, indiscriminate (i.e., unable to be directed at military objectives, and/or whose effects cannot be contained within the boundaries of the law).[501]

The second dimension corresponds to ascertaining the boundaries within which the use of these weapons, means or methods of warfare would be lawful (and, conversely, the 'line not to cross'). In fact, the ICRC has established that the weapon or means of warfare under scrutiny must not be assessed in isolation from the method, i.e., the ways in which it is intended, and expected to be used. The legal review is, in this sense, holistic, as it prescribes an assessment of the weapon as an object (i.e., its legality by nature), but also the greater context, tactics and strategies surrounding the deployment and use of the weapon in question.

In the greater scheme of things, both dimensions have translated, over decades, in two approaches in regulating, limiting, or even prohibiting certain types of weapons. First, the international community has negotiated and adopted a number of international treaties forming 'blanket bans' (disarmament treaties) over certain specific types of weapons that have been found as unlawful by nature. For example, the chemical weapons convention, the treaty on the prohibition of nuclear weapons, and the biological weapons convention, all effectively prohibit the development, production, acquisition, transfer, stockpiling and use of specific types of weapons of mass destruction (WMD). Beyond WMDs, treaties such as the Mine Ban Treaty also exist to ban conventional weapons that are found to be unlawful by nature. The second approach consists of treaties where certain types of weapons may not be unlawful *per se*, but are (at least partly) limited in their use and the way they are used (arms control treaties). Its

[500] 'The Convention on Certain Conventional Weapons' (*United Nations Office for Disarmament Affairs*) <https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/> accessed 23 September 2022

[501] 'Rule 71: Weapons That Are by Nature Indiscriminate' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule71> accessed 23 September 2022

most notable example is the CCW, which consists of an 'umbrella treaty' that both limits the deployment and use of certain types of specific weapons based on each of its additional protocols, and prohibits certain weapons such as its First Additional Protocol, which prohibits the use of small fragments non-detectable by x-ray, as well as its Fourth Additional Protocol, which constitutes a blanket ban on blinding lasers. The latter two would rather be disarmament protocols (i.e., the first category presented above) despite being embedded within an arms control treaty.

These treaties were negotiated, and came to be to complement the legal reviews to be undertaken under Article 36. In fact, for example, the CCW was negotiated because there was a sense that further clarity was needed because API did not go far enough to answer some of the concerns relating to some particular conventional weapons.[502] This does not mean, however, that because a weapon, means or method of warfare is not covered by a specific treaty, that they remain in a lawless 'grey zone': this would be where Article 36 legal reviews step in. Although there is no universal standard on how these Article 36 reviews ought to be conducted, they maintain a minimum threshold, requiring a systematic legal review over new weapons, means and methods of warfare. Whether they are fit for the purpose of AI-enabled systems for military targeting is another question the following section will discuss.

4.2. Evaluation of machine learning programmes: are Article 36 legal reviews fit for purpose?

There are two main questions that need to be taken into account when reflecting on Article 36 legal reviews vis-à-vis AI-enabled systems. First, it is not certain whether all AI-enabled systems to assist with military targeting would fall within the scope of Article 36 reviews (i.e.,

[502] Justin McClelland 'The review of weapons in accordance with Article 36 of Additional Protocol I' (2003) 85(850) <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/review-of-weapons-in-accordance-with-article-36-of-additional-protocol-i/61B38D9967B45A0EDE039E13D2024EAA> accessed 23 September 2022

as a weapon, mean or method of warfare', as articulated in API). Second, it is argued here that Article 36 legal reviews must not be a 'one-off' event done in the process of getting the stamp of approval for a system's deployment. Because of AI's ever-learning and ever-evolving nature, they must be subjected to an iterative process, subject to legal reviews on a regular basis to ensure that it remains clear under which circumstances the deployment and use of these systems would comply with the law.

4.2.1. A question of scope: the applicability of Article 36 legal reviews on AI-enabled systems for military targeting

Article 36 legal reviews feature prominently in international deliberations surrounding lethal autonomous weapons systems within the framework of the CCW. Among the earliest suggestions made by some delegates to the discussions, some included the exchange of information of best practices relating to the implementation of the review process, the teams involved in the review, and oversight with regards to the reliability of review data.[503] On the basis that these reviews are indeed of customary nature in international law, this would indeed be a useful approach as all lethal autonomous weapons systems, which fall within the scope of the CCW discussions, would need to be subject to such reviews.

Beyond the 'lethal autonomous weapon systems' discourse

Yet, not all AI-enabled systems designed to assist with military targeting are necessarily lethal autonomous weapons systems. The technology discussed within the framework of the CCW is indeed important and disruptive; but these deliberations are limited in their mandate. They are limited to weapons systems only (i.e., those deployed in the battlefield), and among these weapons systems, exclusively those of lethal nature (i.e., designed to produce lethality as an

[503] GGE of the CCW on LAWS 'Weapons Review Mechanisms: Submitted by the Netherlands and Switzerland' (2017) CCW/GGE.1/2017/WP.5 <https://undocs.org/ccw/gge.1/2017/WP.5> accessed 30 September 2022

effect). Yet, it is not because an AI system is not a weapon system that produces no lethal effect that it is any less problematic.

AI systems hold so much potential as enablers, that they are expected to be developed and eventually deployed for a wide range of applications and across the entire targeting cycle. Obviously, one of the key applications of AI would be to enable robotics and other 'physical' systems in the battlefield, whether it is with regards to target selection capabilities in weapons systems; enabling the deployment of swarm (armed) drones; or assisting with the navigation of surveillance robots in cluttered environments (e.g., the deployment of drones at large in the sea for the detection of enemy submarines)[504]. Another key application of AI for military targeting would be to enable intelligence collection and forming that precedes, and underpins, targeting operations. As outlined in the previous chapter, AI constitutes a major enabler to extract insights from data, for example by identifying suspected members of non-state armed groups through GSM metadata. Finally, another key application of AI for the military relates to enhancing the army's digital systems – whether it relates to increasing the efficiency of logistics, to enabling highly sophisticated and complex, targeted cyber operations.[505] In fact, with the growing digitisation of conflict, it would be unreasonable to exclude military targeting operations in the digital realm in our discussions: NATO has, for example, recognised since 2016 cyber as one of its domains of operation, putting cyber operations on equal footing as those on land, air and sea.[506] Parties to an armed conflict could indeed conduct military targeting operations that may not necessarily be of kinetic, 'tangible' nature (e.g., through the

[504] Natasha Bajema 'Will AI Steal Submarines' Stealth?' (*IEEE Spectrum*, 16 July 2022) <https://spectrum.ieee.org/nuclear-submarine> accessed 6 October 2022

[505] Wyatt Hoffman 'AI and the Future of Cyber Competition' (2021) CSET <https://cset.georgetown.edu/wp-content/uploads/CSET-AI-and-the-Future-of-Cyber-Competition-4.pdf> accessed 6 October 2022

[506] 'NATO Recognises Cyberspace as a 'Domain of Operations' at Warsaw Summit' (*CCDCOE*) <https://ccdcoe.org/incyder-articles/nato-recognises-cyberspace-as-a-domain-of-operations-at-warsaw-summit/> accessed 6 October 2022

launch of missiles), but rather through surgical, highly-targeted cyber operations against the adversary's strategic assets.[507]

Applicability of Article 36 to non-weapons systems

These possible applications of AI for military targeting not only show that these systems are not limited to weapons systems, they also do not need to be aiming to produce lethal effects on the target. It is however argued here that these systems may still need to be legal review processes, and not necessarily under the framework of Article 36, despite not necessarily being weapons systems. The present section will indeed discuss the scope of 'means and methods of warfare' subject to Article 36 reviews in the light of AI-enabled technologies.

There is no universal definition of what constitutes 'a new weapon, means and methods of warfare' subject to Article 36 reviews. While the way these reviews are conducted remains at the discretion of each state, so does the determination of what technology/system falls within the scope of this legal obligation. AI in and of itself not being a weapon, nor a means and methods of warfare, it is questionable whether all AI systems would indeed need to be subject to Article 36 legal reviews. In a way, such ambiguity is not new: There was a time when states had to grapple with fire as an element that would enable weapons, means and methods of warfare.[508] While fire can indeed be destructive, it is not because a certain technology produces fire that it is necessarily a weapon, means or method of warfare (e.g., everyday household items using fire such as fireplaces, kitchen hobs, etc.) States eventually cleared out the ambiguity by adopting the additional protocol to the CCW dedicated to incendiary weapons; where the focus is on what the weapon or munition in question was designed for (i.e., 'designed to set fire to

[507] Beyza Unal 'Cybersecurity of NATO's Space-based Strategic Assets' (2019) Chatham House <https://www.chathamhouse.org/2019/07/cybersecurity-natos-space-based-strategic-assets> accessed 6 October 2022

[508] William Hays Parks 'Means and Methods of Warfare' 38 George Washington International Law Review <https://heinonline.org/HOL/P?h=hein.journals/gwilr38&i=521> accessed 6 October 2022

objects or to cause burn injury to persons through the action of flame, heat, or combination thereof, produced by a chemical reaction of a substance delivered on the target').[509]

Such approach was helpful in clarifying the law surrounding the use of fire as an enabler for weapons and munitions; and the AI field could indeed draw some inspiration from this approach by focusing on the aim/what the technology was designed for, and the intended effects and consequences of their use. This approach, however, would not be sufficient, in the sense that as we have seen earlier, not all AI systems are designed to be directly engaging with the target (e.g., battle management software, AI programmes designed to collect and process 'bulk' datasets for intelligence collection and creation, AI-enabled drone deployed for ISR). In fact, the approach adopted for incendiary weapons would have led to the conclusion that these AI software/programmes would not constitute weapons: the CCW Third Additional Protocol uses verbs that imply a direct causation link between the weapon and the target (i.e., 'set fire to objects or to cause burn injury'); something autonomous weapons systems may do, but not big data programmes for intelligence collection, for example. In addition, the ICRC defines 'means of warfare' as a term encompassing 'weapons, weapons systems or platforms employed for the purposes of attack in an armed conflict', generally referring 'to the physical means that belligerents use to inflict damage on their enemies during combat'.[510] The intangible nature of software would therefore exclude it from 'means of warfare'. Finally, 'methods of warfare' have been argued as generally understood 'to mean the way in which weapons are used'.[511] While this paves the way for appropriate legal reviews that do not only look at the object in question in itself, but also the context in which it is expected and intended to be used; it does

[509] CCW Protocol III art 1
[510] 'Means of warfare' (*ICRC*) <https://casebook.icrc.org/glossary/means-warfare> accessed 6 October 2022
[511] Isabelle Daoust, Robin Coupland, Rikke Ishoey 'New wars, new weapons? The obligation of States to assess the legality of means and methods of warfare' (2002) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/new-wars-new-weapons-the-obligation-of-states-to-assess-the-legality-of-means-and-methods-of-warfare/89B6C45677FB9F7833B9EBF5B1A4B895> accessed 6 October 2022

not give room for non-weapon AI programmes to be subject to Article 36 reviews. These reasonings and current interpretations of Article 36's scope would lead us to the conclusion that not all AI technologies designed to assist in military targeting would be subject to Article 36 legal reviews, even if their deployment and use would (in)directly lead to engaging the targets in question.

However, it does not mean that these programmes should not be subject to any form of legal scrutiny/examination. In fact, the author would argue here that these programmes falling below the threshold of 'weapons' would still need to go through legal reviews by virtue of the rule of precautions. As established much earlier in the thesis, the duty to take precautionary measures extends to those 'who plan or decide upon an attack'.[512] This implies a certain assumption the commander and decision-makers rest on: that the AI system to be used is reliable and that it will operate as intended. The commander and decision-makers may not be those with the required technical expertise to understand, in breadth and depth, the inner workings of the AI-enabled system in question. They may either rely on a user-friendly interface the developers will have developed and assumed to be tailored to the commanders' needs, and/or on the assessment of analysts who have the technical expertise required to operate the system but will need to summarise and present their findings, assumed, again, to be reliable for mission success.

In the end, in both cases, the commander will rest on the assumption that the system is reliable. In order to be reliable, both from an operational/technical perspective but also legal perspective, the system must have gone through a number of tests, evaluations, reviews and vetting processes before being validated as fit for deployment for the intended uses. In this sense, one of the reviews in question would be legal reviews through which the AI system has been

[512] API, art 57

validated as 1) not unlawful by nature/in and of itself; and 2) its intended uses would indeed comply with the applicable laws; regardless whether the system is a weapon – as long as it contributes and will actively feed into the decision-making processes.

The first element echoes the customary prohibitions for means and methods of warfare to not be of a 'nature to cause superfluous injury or unnecessary suffering', as codified in Article 35(2) of API. As such, developers will need to work closely with lawyers in the development stages of the systems in question to ensure that this obligation is met and well implemented. The second element serves as an umbrella requirement to ensure that the use of the system in question does not lead to the violation of other rules pertaining to the conduct of hostilities – most notably, but not exclusively, the rule of distinction, as alluded in the previous paragraph, as well as the rule of proportionality and (circling back to) the rule of precautions during the conduct of the attack, i.e., constant care to spare civilians, and take all feasible precautions to avoid, and to minimise, incidental loss of life and damage. In this sense, legal assessors will not only need to ascertain under which conditions, and for which uses the AI system would be lawful to operate. They will also need to identify and establish, as required, the necessary safeguards measure to ensure that the operation of the AI system in question will remain within the boundaries of the law. This includes, for example, working with the developers to embed fail-safe programmes in situations where the system's confidence rate in target identification falls below a certain percentage; and that the interface enables the commander to see the programme's rate of confidence and thus, including when it is falling (i.e., transparency requirement).

In addition, the conduct of legal review of systems falling below the threshold of 'weapons' is arguably also part of the duty to respect and ensure respect for IHL. An unarmed surveillance drone, for example, may not necessarily be subject to Article 36 legal reviews per se, however the duty of the purchasing State to respect and ensure respect for IHL has a number of

implications. First, they need to ensure that the development, deployment and use of such surveillance drones would not result in the violation of IHL either by its deployment or use. Second, the duty to 'ensure' respect is such that the procuring State must verify that the processes surrounding the drone's development, deployment and use would minimize risks of IHL violation.

4.2.2. The need for an iterative review process

Legal reviews, despite not necessarily 'Article 36 reviews' per se, therefore remain necessary to ensure legal compliance in the development and deployment of AI systems, even if they fall outside of the scope of weapons systems. While the assumption is that these reviews are conducted prior to the system's deployment (i.e., as part of the TEVV processes), it is important that they are undertaken on a regular basis throughout the system's lifecycle, and not just as 'one-off' process required for the stamp of approval needed for its deployment.

Those developing and deploying AI systems for military targeting will indeed need to ensure that their systems remain lawful by nature, and that it remains clear which applications and under which circumstances would their use be lawful. It is not because a system was deemed as lawful at the time it was first deployed, that it would remain as such – especially because of the ever-evolving nature of machine learning systems. This is particularly the case for reinforcement learning programmes, which learn 'on the spot' as the system is used for real-life applications; thus, their performance will evolve through time depending on their experience after each deployment.

The United States, for example, have acknowledged the need to re-visit the ways testing and evaluations are undertaken for AI systems, given risks of 'non-linear, time-varying, and

emergent behaviors'.[513] Challenges arise from flexible behaviours and evolving performance from these systems, especially when they are deployed in untested environments – and this not only affects the overall confidence of human operators in the system, but also the overall reliability of the system, including whether it is still reliable for legal compliance. It is therefore important that these systems are subject to regular scrutiny with regards of their compliance to the law, especially as the military will grow increasingly dependent on AI. In fact, in light of this growing dependency, this further reinforces the case that AI systems have to be subject to legal reviews to ensure that the necessary precautions are taken to comply with the law – regularly – even if the system falls below the threshold of 'weapons, means and methods of warfare'.

In response to this ever-evolving nature of machine learning, it would therefore be important that these systems are scrutinised on a regular basis. It is however difficult to ascertain for all systems a 'fixed' calendar period where they would benefit from this systematic review (e.g., once a month, once a quarter, etc.) This need for legal review is constant; but the pace at which the AI system's performance will evolve depends on a number of factors, including the learning method (e.g., reinforcement learning, versus supervised learning, etc.), what it has been designed for, the complexity of the programming, etc. – if it will evolve at all. The risk of fluctuating behaviour and performance however remains. Whether armed forces are even keen on procuring, developing and deploying systems with self-learning abilities for applications with stakes as high as military targeting is another question. The United States, for example, have outlined measures for such systems in their 2023 Political Declaration on Responsible Military Use of AI and Autonomy, which has been endorsed by 57 States as of November 2024. Specifically, for 'self-learning or continuously updating military AI capabilities, States

[513] Office of the Director, Operational Test and Evaluation (US) (n 396)

should ensure that critical safety features have not been degraded, through processes such as monitoring.'[514]

As such, one approach could be to develop and implement an auditing systems and processes, including the use of software that will constantly monitor the programme's compliance with the law, even post-deployment.[515] These constant audits will be particularly important for those systems that, by nature, are likely going to change in performance (e.g., reinforcement learning). These solutions will indeed be key in ensuring consistent compliance with the law, in addition to serving as a trust-building mechanism. In fact, these audits would constitute one of the key mechanisms that would help flag, to the end-users and those involved in its maintenance, that the machine's performance may not meet the legality threshold anymore; or at least the degree of risks and uncertainties stemming from its change in performance have passed a certain threshold that would require the machine to undergo a new round of thorough and in-depth reviews (and subsequent TEVV processes again before going back for deployment). Embedding such auditing and periodic review processes from the development stages of the programme is therefore essential to ensure that legal compliance remains at the heart of AI development from the beginning – and not as an afterthought.

[514] 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy' (*U.S. Department of State*, 9 November 2023) <https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/> accessed 18 November 2024

[515] Inioluwa Deborah Raji and others 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing' (FAT* '20, Barcelona, Spain, January 27–30, 2020) <https://dl.acm.org/doi/pdf/10.1145/3351095.3372873> accessed 10 October 2022; see also for an example of software assisting with algorithmic auditing: Koshiyama, Kazim, Treleaven (n 397)

Chapter IV – The Question of Accuracies and (Un)Certainty – Knowns versus Unknowns

1 Introduction

The promise for greater precision and accuracy, in addition to increased speed and the potential for scale, constitute key drivers for high levels of interest and subsequent investment in military AI. Whether it is to enhance the complexity of decision support systems or to enable 'surgical strikes', such potential is, at least, in theory, appealing both from an operational standpoint and that of legal compliance.[516] The appeal that comes with AI technologies resonates well with long-lasting issues military forces have tried to grapple with since the pre-AI era, even over centuries, to ultimately generate and sustain power.[517] These can include, but are not limited to, the desire for information, speed, and scale. Uncrewed vehicle systems were (and still are) seen as key in providing commanders with the critical information required at the strategic level.[518] During both the First and Second World War, codebreaking was a key element of warfighting in order to ensure information superiority – including the tremendous work undertaken by Alan Turing, among others, in Bletchley Park to decipher German communications and support intelligence efforts. As far as history is concerned, records of innovation in intelligence collection methods trace back to the age of the Roman Empire and

[516] The appeal for 'surgical strikes', or precision targeting, has been increasing notably since the 1991 Gulf War – see: J. Marshall Beier 'Short circuit: retracing the political for the age of 'autonomous' weapons' (2017) 6(1) Critical Military Studies <https://www.tandfonline.com/doi/full/10.1080/23337486.2017.1384978?needAccess=true> accessed 18 October 2023

[517] Michael C. Horowitz and Shira Pindyck define military innovations as 'changes in the conduct of warfare designed to increase the ability of a military community to generate power.' For a more in-depth discussions of both authors on what military innovation consists of, see: Michael C. Horowitz and Shira Pindyck 'What is a military innovation and why it matters' (2022) 46(1) Journal of Strategic Studies <https://www.tandfonline.com/doi/full/10.1080/01402390.2022.2038572> accessed 16 October 2024

[518] The role of uncrewed systems for information gathering (and, even, lethality) has particularly been brought to the fore in the light of the on-going conflict in Ukraine, see most recently: K. Bondar, 'Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare' (2025) Center for Strategic International Studies, <https://www.csis.org/analysis/ukraines-future-vision-and-current-capabilities-waging-ai-enabled-autonomous-warfare> accessed 3 June 2025. The criticality of uncrewed systems has, in addition, been recognized for over twenty years now: Brian Burridge 'Post-Modern Warfighting with Unmanned Vehicle Systems: Esoteric Chimera or Essential Capability?' (2005) 150(5) The RUSI Journal <https://www.tandfonline.com/doi/abs/10.1080/03071840509431879> accessed 16 October 2024

Ancient Egypt.[519] The desire for speed is well reflected in investments for research and development to enhance missile technologies; while the pursuit for scale was evident in the Second World War over the use of tanks and military aircraft to counter firepower.[520]

AI innovation holds promises in these three domains.[521] As discussed throughout this thesis, AI capabilities bring the potential to increase situational awareness, the scale of operations and information being processed, as well as the speed at which decisions are made. This is even more important in the light of technological convergences, e.g., between cyber and AI, thus requiring greater degrees of autonomy and the ability to respond at speed.[522] The sophistication of offensive capabilities, both kinetic and cyber, is such that the need for rapid responses that surpass human cognitive abilities will only grow. This has been a key driver in the deployment of automated systems in place already, such as the Patriot missile system or Israel's Iron Dome.[523] AI is seen as a key enabler in accelerating their sophistication.

In addition to enhancing capabilities in the battlefield, AI innovation also holds promises to address the uncertainties brought by warfare. By processing vast amounts of data at massive scales and at unprecedented speed, users are provided with an avenue to navigate the dynamic and messy nature of war and, subsequently, ensure supremacy over the adversary. Investment in this area is high, with data analytics software offerings mushrooming in the security and

[519] For a fascinating historical account on intelligence, see: Christopher Andrew *The Secret World: A History of Intelligence* (Penguin Books, 2018)

[520] William J. Perry 'Military technology: an historical perspective' (2004) 26(2-3) Technology in Society <https://www.sciencedirect.com/science/article/abs/pii/S0160791X04000363> accessed 16 October 2024

[521] Afina (n 2)

[522] The intersection between AI and cyber operations is of increased interest: for example, the subject notably featured in a number of discussions as part of the 2023 Internet Governance Forum. For a discussion, see: Nektaria Kaloudi, Jingyue Li 'The AI-Based Cyber Threat Landscape: A Survey' (2020) 53(1) ACM Computing Surveys <https://dl.acm.org/doi/abs/10.1145/3372823> accessed 18 October 2023. On missile technology, one key concern in security and defence discussions pertains to the development and testing of hypersonic missiles, against which many states seek to develop AI-enhanced missile defence capabilities. See for example: Richard H. Speier 'Hypersonic Missiles: A New Proliferation Challenge' (*The RAND Blog*, 29 March 2018) <https://www.rand.org/blog/2018/03/hypersonic-missiles-a-new-proliferation-challenge.html> accessed 18 October 2023

[523] Joel Block 'A laws of war review of contemporary land-based missile defence system 'iron dome'', (2017) 45(2) Scientia Militaria: South African Journal of Military Studies <https://www.ajol.info/index.php/smsajms/article/view/164585> accessed 24 November 2022

defence sector.[524] Beyond the sheer scale and speed, it has also been argued that AI could aid critical decision-making processes in situations where psychological biases and other human cognitive limitations could constitute a liability to mission success.[525] This is particularly argued in the light of past incidents, for example the Operation Provide Comfort (1994) where the US attacked their own Black Hawk helicopters as a result of 'a series of avoidable errors and failure of safeguards in place at the time of the shootdown'.[526]

These potential benefits, however, rest on the assumption that the technology is working as intended, i.e., that its output is predictable; it performs at a high rate of accuracy; and that it is reliable for mission success. Yet, there is no universal norm, nor standard, to evaluate and benchmark the accuracy and efficiency of military AI– whether it is in terms of technical reliability or with regards to legal compliance. Furthermore, there is a lot of misconception as to the nature of AI solutions and the subsequent, misleading expectation that these technologies must be at one hundred percent accurate in order to be reliable. It is important, from the outset, to consider the outputs of AI-enabled programmes as, essentially, approximations or predictions based on what it has been taught – and not as oracles. In other words, the functioning and outputs of AI programmes will, inherently, have a certain degree of inaccuracy. This leads to the question of how to deal with this issue of inherent inaccuracy and subsequent uncertainties stemming from this: The present section hopes to explore whether international humanitarian law could potentially provide answers or at least, possible approaches to address

[524] One of the most striking examples being Palantir with its AI Platform for Defense: 'AIP for Defense' (n 259)
[525] Missy Cummings 'Lethal Autonomous Weapons: Meaningful human control or meaningful human certification?' (*Technology and Society*, 23 December 2019) <https://technologyandsociety.org/lethal-autonomous-weapons-meaningful-human-control-or-meaningful-human-certification/> accessed 23 November 2024; see also Keith Dear and Magdalena Pacholska, 'Machine Inhumanity? The Misguided Approach to Human Control in Advanced Weapon Systems' (forthcoming)
[526] General Accounting Office (US) (1997) 'Operation Provide Comfort: Review of U.S. Air Force Investigation of Black Hawk Fratricide Incident' (1997) GAO/OSI-98-4 <https://www.gao.gov/assets/osi-98-4.pdf> accessed 25 November 2022

the issue of accuracy and to what extent are AI technologies expected to exhibit such trait under the law.

For example, the leaked documents revealing the existence of the NSA's Skynet show that the statistical algorithms used were able to find the couriers with a false alarm rate of 0.18%, assuming that it was programmed to allow missing half (50%) of them.[527] These seemingly low rates of inaccuracy must be caveated by the fact that they are the result of testing in highly-controlled and certain environments, which may not necessarily reflect reality on the ground as we have seen with the case of the Al Jazeera journalist wrongfully identified as a target by Skynet due to the nature of his activities.[528] In fact, evaluating accuracy rates in testing environments on the one hand, and in the battlefield on the other hand, can prove to be difficult for a number of reasons. From discrepancies between both environments to the lack of established testing and evaluation benchmarks, a lot of uncertainties arise. As such, paradoxically, the supposed solutions to addressing warfare's uncertainties bring, themselves, their own set of uncertainties. In the absence of universally accepted benchmarks to evaluate the accuracy of military AI, this chapter will aim to examine the ways through which international humanitarian law can be used as a reference for such assessments.

1.1 Probabilistic AI Technologies: Inherent Inaccuracies and Uncertainties

From its genesis in the 1950s, one of the main objectives as to why AI is being developed and pursued in the first place is to simulating 'every aspect of learning or any other feature of intelligence' through their precise description.[529] Through time, the overarching objective and purpose behind AI research and development seems to have evolved and grown in nuance.

[527] 'SKYNET: Courier Detection via Machine Learning' (*The Intercept*, 8 May 2015) <https://theintercept.com/document/skynet-courier> accessed 25 November 2022
[528] D. Parvaz 'Journalists allege threat of drone execution by US' (*Al Jazeera*, 2 April 2017) <https://www.aljazeera.com/news/2017/4/2/journalists-allege-threat-of-drone-execution-by-us> accessed 25 November 2022
[529] Stuart Russell, *Human Compatible: AI and the Problem of Control* (Allen Lane, 2019) 4

Beyond simply emulating intelligence, some even went so far as pursuing super-intelligence exceeding that of human cognition and intelligence which, in fact, constituted the main drive behind AI research in its first few decades before the gradual shift towards 'narrow' AI, designed to solve and be very efficient at very specific tasks yet would be 'dumb' for tasks other than the pre-designated ones.[530]

As technological progress is proceeding apace, a sub-set of studies in the field of artificial intelligence has emerged: that of testing and evaluation, and the development of benchmarks and success metrics to assess a system's performance. Perhaps the most famous evaluation to date may actually be the oldest, which even pre-dates the Dartmouth College workshop: the Turing Test, proposed by Alan Turing in 1950. The latter consists of evaluating AI systems' intelligence against that of humans on the basis of questioning and human judgment.[531] In medicine, the need for 'rigorous evaluation of the quality and impact of AI was recognised in [as early as] the 1980s' and with early work dating from back to the 1970s.[532] Today, the field of testing and evaluation has grown so much that not only have a number of approaches and testing methods emerged; the range of evaluation and success metrics has also grown.

With regards to the purely technical evaluation of the system, for example, a number of approaches have emerged. These include, but are not limited to:

- Cross-validation, which consists of assessing the ability of models to 'generalize' by subjecting it to testing datasets that are different to those used for its training.[533] The aim is to prevent models from being over-fitted, i.e., 'too' trained on a very specific and

[530] The shift from artificial general intelligence to narrow AI has been discussed in Chapter 1, Section 2.1.1
[531] Feng Liu, Ying Liu, Yong Shi 'Three IQs of AI systems and their testing methods' (2020) 2020(13) The Journal of Engineering <https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/joe.2019.1135> accessed 17 October 2024
[532] Farah Magrabi and others 'Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications' (2019) 28(01) Yearbook of Medical Informatics, <https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0039-1677903> accessed 17 October 2024
[533] Daniel Berrar 'Cross-validation' in Shoba Ranganathan and others (eds) *Encyclopedia of bioninformatics and computational biology* (1st edition, vol. 1, 2019) <https://oro.open.ac.uk/96776/> accessed 23 November 2024

narrow dataset that would make it unable to function effectively with data other than what it had been exposed to.

- Stress-testing, which consists of assessing the system's performance under 'stress execution conditions', i.e., by simulating edge-cases to test the system's behaviour under such abnormal and extreme circumstances.[534] The aim is to test and measure the extent to which the model can be exposed to 'stress' and at what point would its performance start to deflate.
- Adversarial testing, which consists of purposefully submitting inputs designed to mislead the model and tamper its calculations to undermine its performance.[535] The aim is to assess the system's resilience against deliberate attempts at tampering its calculations and, eventually and as appropriate, for the developers to devise counter-measure solutions. For example, this would consist of testing the efficiency of anti-spam filters in email inboxes by exposing it deliberately to emails that would tamper with its ability to filter out undesirable messages, which would subsequently enable harmful messages to 'beat' the system and land in the users' main inboxes without ending up in the junk folder.

Building on these methods, a number of metrics have subsequently emerged to measure the performance of systems when subject to those tests. Beyond technical considerations, these evaluation and success metrics generally tend to capture a number of considerations drawing from multiple disciplines. These include, but are not limited to:

[534] Mahshid Gelali Moghadam 'Machine learning-assisted performance testing' (Proceedings of the 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), New York, 2019) <https://doi.org/10.1145/3338906.3342484> accessed 17 October 2024

[535] Xianmin Wang and others 'The security of machine learning in an adversarial setting: A survey' (2019) 130 Journal of Parallel and Distributed Computing <https://www.sciencedirect.com/science/article/pii/S0743731518309183> accessed 17 October 2024

- Accuracy scores, which correspond to metrics that measure how correct the model is.[536] Accuracy scores are typically used in classification models such as, for example, the determination as to whether or not an object is targetable or non-targetable. What is considered as 'targetable' or 'non-targetable' here would extend beyond the technical realm and, would rather, at minimum, draw on international humanitarian law and, possibly, ethics and other aspects (e.g., military manuals, existing rules of engagement, etc.) The task of determining what is 'accurate' and what is 'non-accurate' must be undertaken carefully, especially with regards to the implications of false positives versus false negatives. In the context of models designed to identify objects as 'targets', false positives for example (i.e., inaccurately classifying a protected object as a target) would, legally, be more consequential than false negatives (i.e., inaccurately not identifying a lawful target as a military objective), although from an operational perspective the latter may be more consequential in terms of tactical and strategic advantage.
- Confidence scores, which correspond to metrics that measure how confident the model is in its predictions, i.e., the chances of its output being correct.[537] Unlike accuracy scores, confidence scores do not operate on the basis of 'correctness' but rather, the probabilities corresponding to each of the system's outputs. The use of these metrics would be particularly important for military targeting, especially for decision support systems commanders may rely on in time-sensitive operations.[538] Confidence scores in such contexts would, for example, correspond to 'Object X is 80% likely to be a lawful

[536] Yunfeng Zhang, Q. Vera Liao, Rachel K. E. Bellamy (2020), 'Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making' (FAT* '20: Proceedings of the Conference on Fairness, Accountability, and Transparency 2020) <https://dl.acm.org/doi/abs/10.1145/3351095.3372852> accessed 19 October 2023
[537] Ibid
[538] Arthur Holland Michel, 'The Black Box, Unlocked: Predictability and Understandability in Military AI' (2020) UNIDIR, <https://unidir.org/files/2020-09/BlackBoxUnlocked.pdf> accessed 18 October 2024

target', or 'Person Y is 76% likely to be a targetable fighter'. Here, a couple of considerations must be raised. First, confidence scores are not necessarily correlated with correctness. The system can be 80% confident that Object X is a lawful target (i.e., a military objective), yet the reality is such that this assessment is more complex and nuanced (e.g., Object X in question is a civilian shelter at night, and a command-and-control center during the day). What the system's 'beliefs' are that form the basis for its confidence (or non-confidence) is often uncertain and further exacerbated by the black box problem, thus making it difficult to estimate whether the system is under- or over-estimating its capabilities.[539] Second, in the military domain, a system's ability to estimate and communicate uncertainty is of particular importance due to the high stakes at hand.[540] This is particularly important for classification models meant to determine the targetability, or not, of an object or person. Effective communication of the system's uncertainty to the commander could, for instance, push the latter to undertake extra assessment with regards to the target and thus, foster compliance with the rule of distinction.

In all the proposed approaches to measuring an AI system's success, one clear and consistent pattern has emerged: all assume that AI has an inherent degree of uncertainty. In other words, while these systems' performance and subsequent outputs will be heavily influenced by their development and testing, yet, there will always be some level of uncertainty as to their outputs which can only be captured and eventually measured after deployment.[541] In addition to its inherent, probabilistic nature, many factors contribute to this uncertainty surrounding the programme's outputs and its final accuracy rate, including the limitations around the success

[539] Ibid
[540] Richard Tomsett and others 'Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI' (2020) 1(4) Patterns <https://www.cell.com/patterns/fulltext/S2666-3899(20)30060-X> accessed 18 October 2024
[541] Zhang Liao, Bellamy (n 536)

metrics and risks of overestimating the technology's performance, the differences between the testing environment and realities in the battlefield, as well as the inherent messy and variable nature of war.[542] The chances that the system's output will indeed be correct and meet set expectations can be high – and heightened, but it is assumed that it will never be perfect: chances are, by nature, probabilistic and not synonymous to absolute certainty.

Generally, 'uncertainty' is defined as a 'situation in which something is not known, or something that is not known or certain.'[543] Uncertainty has always been a prevalent theme present across disciplines, each of which have attempted to define and measure it against a range of different contexts and applications – and the subsequent implications of not knowing in these situations. Uncertainty has, for one, been studied in 'abstract', theoretical contexts: for example, in the context of economics, uncertainty has been examined against variation, as well as its magnitude and the correlation of uncertainties with other elements (e.g., real activity).[544] In the field of statistical studies, understanding the different types of uncertainties and their respective implications is of great importance too. In statistical analysis, uncertainty would for example need to be factored in against their subsequent trade-offs – whether this uncertainty is, by nature, structural, technical, or related to risk.[545] In international relations theory, the force of uncertainty has also been a central, recurrent theme of study, in addition to the different meanings it may hold: for example, in certain contexts, uncertainty can be understood as induced fear, ignorance, confusion, or indeterminacy.[546] The question of uncertainty and its implications have also been examined against more 'concrete' situations. For example, in the

[542] Afina (n 85)
[543] 'Uncertainty' (*Cambridge Dictionary*) <https://dictionary.cambridge.org/dictionary/english/uncertainty> accessed 2 February 2023
[544] Kyle Kurado, Sydney C. Ludvigson, Serena Ng (2015), 'Measuring Uncertainty' (2015) 105(3) American Economic Review <https://www.aeaweb.org/articles?id=10.1257/aer.20131193> accessed 8 December 2022
[545] James S. Hodges (1987) 'Uncertainty, Policy Analysis and Statistics' (1987) 2(3) Statistical Science <https://projecteuclid.org/journals/statistical-science/volume-2/issue-3/Uncertainty-Policy-Analysis-and-Statistics/10.1214/ss/1177013224.full> accessed 2 February 2023
[546] Brian C. Rathburn 'Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory' (2007) 51(3) International Studies Quarterly <https://academic.oup.com/isq/article/51/3/533/1795465> accessed 2 February 2023

field of medical practice, uncertainty has been studied against public health responses to crises – such as the COVID-19 pandemic, whereby its management was rapidly and deeply shrouded with uncertainty stemming from a number of factors, ranging from the lack of a sophisticated understanding of the virus in itself to the appropriate response required to the disease's rapid proliferation at a global scale.[547] In the light of its high stakes – quite literally, of life and death – it was critical that healthcare professionals providing care, researchers in the medical field, as well as those responsible for the public health response to adopt, would adequately address the issue of uncertainty and factor it in as an inevitable part of their decision-making.

In all these examples, a number of recurring patterns emerge. First, measures of uncertainty are of particular importance in risk assessments. Regardless of the field and the situation in question – whether it is in the context of economics, statistics, public health or, in our case, in the context of military targeting, a clearer understanding of uncertainty as a concept, but also the forms that it can take and the sources and factors contributing to uncertainty – they all will help decision-makers in the choice and direction they adopt in the light of trade-offs and risks they are willing to accept as a result of 'not knowing'. In fact, a number of multidisciplinary studies have been specifically dedicated to dissecting decision-making under uncertainty both in theory and application. This was done, for example, from a computational perspective, but drawing from a number of disciplines, including philosophy, behavioural game theory and statistics.[548] In the context of military targeting, a clear and well-defined approach to uncertainty should, in principle, enable better decision-making, as it will then factor in the subsequent risks against what the commander is ready to take (e.g., of inaccuracy and thus, of

[547] Jonathan Koffman and others 'Uncertainty and COVID-19: how are we to respond?' (2020) 113(6) Journal of the Royal Society of Medicine <https://journals.sagepub.com/doi/pdf/10.1177/0141076820930665> accessed 2 February 2023

[548] Mykel J. Kochenderfer, *Decision Making Under Uncertainty: Theory and Application* (MIT Lincoln Laboratory Series (2015)

causing disproportionate harm to civilians and civilian objects; mission failure or any other unpredicted outcome; etc.)

The second pattern observed pertains to the complexity surrounding the situations in which uncertainty arises. Complex decision-making environments are often referred to as complex 'systems'. This is in the light of the number of factors and agents that not only interact with this complex ecosystem in a 'messy', non-linear way – but are also mutually interacting and somewhat interdependent in certain cases.[549] A clash between complexity and uncertainty makes decision-making in certain situations extremely difficult, the possible outcomes of which are can even lead to 'catastrophic outcomes' such as financial instability, existential risks such as climate change, nuclear waste management, or even lethality.[550]

Against the high stakes and the pressure that comes with uncertainty, decision-makers are met with the need to come up with tools that will aid overcome their own cognitive limitations. Going even beyond just compensating for the human decision-maker's limitations, these tools can even go further and not only enable, but ultimately enhance decision-making processes. Computational methods have been particularly put in the spotlight given the major advances and investment in the field over the past few years – such as 'big data' to create intelligence and, subsequently, enable better decision-making.[551]

In the same vein, the development of military AI could be thought as a means to decrease the uncertainty brought in warfare. Paradoxically, these inherently probabilistic technologies are being used to address uncertainties. This paradox is not unique to the military domain and is,

[549] S Lloyd 'Measures of complexity: a nonexhaustive list' (2001) 21(4) IEEE Control Systems Magazine <https://ieeexplore.ieee.org/document/939938> accessed 2 February 2023, as cited in Unal and others (n 479)
[550] Daniel A Farber 'Uncertainty' (2011) 99 Georgetown Law Journal <https://heinonline.org/HOL/LandingPage?handle=hein.journals/glj99&div=30&id=&page=> accessed 8 December 2022
[551] Peter Bossaerts, Carsten Murawski 'Computational Complexity and Human Decision-Making' (2017) 21(12) Trends in Cognitive Sciences <https://www.sciencedirect.com/science/article/abs/pii/S1364661317301936> accessed 2 February 2023

arguably, present across sectors and is thus of general nature. From the medical field to disaster planning, whether it is to reduce uncertainty with regards to a patient's likelihood to have cancer or predict migration flows in the wake of a natural disaster, AI is being developed and used to address uncertainties, despite these technologies' inherent degrees of uncertainties due to their probabilistic nature. As such, it is important that AI technologies are not seen as producing outputs with absolute and prophetic accuracy. What is unique to the military domain, however, is the very nature of warfare and its complex legal, socio-political, and environmental implications and ramifications that may be deeply transformative across borders and generations. In the light of these high stakes, there are unique incentives, and risks, that surround the development and deployment of AI technologies in the military domain. Addressing any form of uncertainty from the latter will thus be critical, including through technological means – e.g., the use of AI-enabled decision support systems to support proportionality assessments with enhanced data analytics. Yet, while these technologies could in principle reduce these uncertainties, a new set of uncertainty can emerge or at least, exacerbate and complexify existing uncertainties, specifically in relation to their reliability and accuracy. For example, these uncertainties pertain to a classification system's outputs designed to assist with target identification. This chapter will seek to dissect the possible legal implications of these specific uncertainties.

1.2 Human Involvement in AI Development: A Decisive Factor to Uncertainties and Inaccuracies

Since the early stages of multilateral discussions and deliberations in the sphere of lethal autonomous weapons systems at the UN, within the framework of the CCW, there has been a significant emphasis, by many, on the role of humans in the development, deployment and use

of military AI and the need to maintain a degree of oversight, control, and/or judgment.[552] Yet, these narratives tend to omit the extent to which this human element is a decisive factor to uncertainties and inaccuracies in the system's performance, and how its influence can go either way, i.e., either in maintaining a high degree of uncertainty and thus, decreasing accuracy; or conversely, in lowering degrees of uncertainty and thus, increasing the chances for accuracy. In other words, amidst the fervent calls for human control (or whichever term is being used), it is unclear whether it would actually solve issues related to uncertainty and inaccuracies.

Amidst the constellation of appellations and terms, from 'human oversight' to 'human involvement,' the concept of 'meaningful human control' over AI-enabled decision-making systems has been relatively popular both among policymakers, experts/scholars and advocacy groups.[553] This concept has been interpreted in different ways – and some are even using different yet adjacent terms, with the United States for example using the term 'appropriate levels of human judgment over the use of force' in CCW discussions and in their national policy documents.[554] The notion that meaningful human control, or at least human intervention, is necessary for lethal decision-making, has been increasingly gaining traction, notably from advocacy groups and aligned states seeking to obtain a mandate for the negotiation of an

[552] A range of terms have emerged in discourses surrounding the human element, including 'human control', 'meaningful human control', 'human oversight', 'human judgment', etc. The UK generally uses the framing 'context-appropriate human involvement' in its statements related to autonomous weapons systems. See for example: United Kingdom, *Input to UN Secretary-General's Report on Lethal Autonomous Weapons Systems (LAWS)* (2024) <https://docs-library.unoda.org/General_Assembly_First_Committee_-Seventy-Ninth_session_(2024)/78-241-UK-EN.pdf> accessed 18 October 2024

[553] This notion of 'meaningful human control' has notably been popular in the multilateral discussions surrounding lethal autonomous weapons systems at the UN, within the framework of the CCW; although there are a number of states and/or organisations using synonyms/alternative terms to refer to the involvement of human operators at different levels and different degrees. For discussions on this issue, see: Michael C. Horowitz and Paul Scharre 'Meaningful Human Control in Weapon Systems: A Primer' (2015) CNAS <https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf> accessed 22 April 2022; Merel Ekelhof 'Autonomous weapons: Operationalizing meaningful human control' (*Humanitarian Law & Policy*, 15 August 2018) <https://blogs.icrc.org/law-and-policy/2018/08/15/autonomous-weapons-operationalizing-meaningful-human-control/> accessed 22 April 2022.

[554] Congressional Research Service (US), 'Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems' (2024) <https://crsreports.congress.gov/product/pdf/IF/IF11150#:~:text=Lethal%20autonomous%20weapon%20systems%20(LAWS,human%20control%20of%20the%20system> accessed 23 November 2024

international treaty in the realm of lethal autonomous weapons systems. For example, in the 2018 iteration of the CCW Group of Governmental Experts, 'further consideration of the human element in the use of lethal force' formed part of the formal agenda of deliberations, specifically under agenda item 6.b. From an ethical perspective, it has been argued that such retention of human oversight (or whatever form/term may be used) is necessary with regards to the human dignity of targets.[555] This argument for 'human dignity' has been further echoed in statements delivered in the context of formal CCW deliberations, for example by Human Rights Watch and the Holy See.[556]

The ICRC has been vocal at advocating for, and 'recommending', the adoption of new, legally binding rules (i.e., an international treaty) on autonomous weapons systems to ensure that 'sufficient human control and judgment is retained in the use of force'.[557] Furthermore, in partnership with the Stockholm International Peace Research Institute (SIPRI), the ICRC has also published an in-depth analysis of the type and degree of human control required to mitigate risks emanating from autonomous weapons systems, specifically with regards to: controls on the weapon's parameters; controls on the environment of use; and controls in the form of human supervision.[558] This argument rests on the premise that IHL rules (and the law more

[555] Daniele Amoroso & Guglielmo Tamburrini (2021) 'Toward a Normative Model of Meaningful Human Control over Weapons Systems' (2021) 35(2) Ethics & International Affairs <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/ethics-and-international-affairs/article/toward-a-normative-model-of-meaningful-human-control-over-weapons-systems/A3FD9EC4CBD6EA77439211537B94A444> accessed 23 November 2024

[556] See, respectively: Mary Wareham, Statement, 'Statement on Lethal Autonomous Weapons Systems to the CCW Annual Meeting' (Geneva, 16 November 2022) <https://www.hrw.org/news/2022/11/16/statement-lethal-autonomous-weapons-systems-ccw-annual-meeting-0> accessed 6 January 2023 and Permanent Mission of the Holy See to the United Nations and Other International Organizations in Geneva at the 2021 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS) of the Convention on Certain Conventional Weapons, Statement, 'Item 5(b) Characterization of the systems under consideration' (Geneva, 4 August 2021) <https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2021/gge/statements/4Aug_HolySee.pdf> accessed 6 January 2023

[557] ICRC, Statement delivered at the CCW GGE on LAWS, 'Autonomous weapons: The ICRC recommends adopting new rules' (Geneva, 3-13 August 2021) <https://www.icrc.org/en/document/autonomous-weapons-icrc-recommends-new-rules> accessed 19 December 2022

[558] Vincent Boulanin and others 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control' (2020) SIPRI <https://www.sipri.org/publications/2020/policy-reports/limits-autonomy-weapon-systems-identifying-practical-elements-human-control> accessed 6 January 2023

generally) are 'ultimately implemented by human subjects who are responsible for complying with these rules in carrying out attacks […]', and that therefore 'some degree of human control over the functioning of an autonomous weapon system, translating the intention of the user into the operation of the weapon system, will always be necessary to ensure compliance with IHL'.[559] In other words, this position and the centrality of humans in the development, deployment and use of AI technologies for military targeting somewhat presents human involvement as a *sine qua non* condition for legal compliance.

However, if AI innovation is to be driven by legal compliance and more specifically by IHL, upstream governance intervention from these technologies' development should not be limited to the ability to identify and pinpoint a specific individual as accountable and responsible for implementing the law, i.e., the user supposedly entrusted with 'control' over the system. This is not to discount its importance: it is in fact critical that, in cases of incidents and potential violations through the use of AI in military targeting operations, accountability and responsibility are upheld. However, this thesis is ultimately about fostering compliance by design and, more generally, responsible AI. This extends beyond maintaining accountability, which should be the very minimum. It is about identifying key legal considerations for the development of military AI, not only to justify their use further down the line in the battlefield but also, if possible, to help increase compliance rate through the use of these technologies. This motivation is driven by the very purpose of international humanitarian law – to protect lives, not justify and legitimize destruction and 'large-scale devastation'.[560]

[559] Neil Davison 'A legal perspective: Autonomous weapon systems under international humanitarian law' (2018) UNODA Occasional Papers No. 30 <https://www.globalgovernancewatch.org/library/doclib/20180208_ICRCAutonomousWeaponSystems.pdf> accessed 6 January 2023

[560] Cordula Droege 'War and what we make of the law' (*Humanitarian Law & Policy*, 18 July 2024) <https://blogs.icrc.org/law-and-policy/2024/07/18/war-and-what-we-make-of-the-law/> accessed 18 October 2024

Protecting lives is more than ensuring accountability: it is not because a weapon system is supposedly under 'human control' that it would necessarily protect civilians from indiscriminate attacks or disproportionate damage, whether due to malice from the user, or due to inaccuracies in the system. As such, the idea that human control is essential for legal compliance may to a certain extent be true, but it is not enough. In fact, it may even be argued that human operators would constitute the 'weak link' under certain circumstances, such as those operations that take place at machine-speed.[561] In such situations, the human element adds to the uncertainty factor and thus may even heighten inaccuracy risks.

One of such instances where the human operator would constitute a liability relates to issues regarding automation bias. The latter specifically refers to the tendency, for humans, to favour the suggestions and outputs generated by an autonomous system at the expense of 'vigilant information seeking and processing'.[562] In other words, automation bias corresponds to the tendency of not questioning and verifying the veracity of a system's outputs due to over-confidence in its performance and accuracy. A solid number of studies have looked at the issue of automation bias, its implications and the risks that ensue in all sectors, from healthcare provision to the administrative procedures.[563] In the military domain, there has been research on automation bias and its impact for example in time-critical decision support systems.[564]

---

[561] Hendrik Huelss 'Transcending the fog of war? US military 'AI', vision, and the emergent post-scopic regime' (2024) European Journal of International Security <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/european-journal-of-international-security/article/transcending-the-fog-of-war-us-military-ai-vision-and-the-emergent-postscopic-regime/35BCDEE8E28B076BCD597AFDC8976824> accessed 18 October 2024; see also H. W. Meerveld and others, 'The irresponsibility of not using AI in the military' (2023) 25 Ethics and Information Technology <https://link.springer.com/article/10.1007/s10676-023-09683-0> accessed 18 October 2024

[562] Mor Vered and others 'The effects of explanations on automation bias' (2023) 322 Artificial Intelligence <https://www.sciencedirect.com/science/article/pii/S000437022300098X> accessed 19 November 2024

[563] See, respectively, Rohan Khera, Melissa A. Simon, Joseph S. Ross 'Automation Bias and Assistive AI: Risk of Harm From AI-Driven Clinical Decision Support' (2023) 330(23) JAMA <https://jamanetwork.com/journals/jama/fullarticle/2812931> accessed 19 November 2024 and Peter Parycek, Verena Schmid and Anna-Sophie Novak 'Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions' (2024) 15 Journal of the Knowledge Economy <https://link.springer.com/article/10.1007/s13132-023-01433-3> accessed 19 November 2024

[564] See for example Missy L. Cummings, 'Automation Bias in Intelligent Time Critical Decision Support Systems' in Don Harris and Wen-Chin Li (eds) *Decision Making in Aviation* (1st edn, Routledge, 2015)

Here too, issues are raised for example with regards to lowered human performance due to automation bias. Further studies have tried to add nuance to the examination of automation bias, by for example looking at the level at which such bias is more likely to occur in the use of such systems.[565] This would be the case if, for example, a system is known to be capable of operating at 80% certainty rate at most; yet the commander will follow its recommendations thinking, and without questioning and verifying, that these systems are at 100% certainty rate. Automation bias issues, however, are not exclusively a problem relevant to use. Human biases such as automation bias can also affect the pre-deployment processes of a system, e.g., leveraging the perceived objectivity of the system to cover or even justify biased performance, which would then have downstream repercussions.[566]

Nevertheless, although human control does not constitute in and of itself a silver bullet to the issues raised by algorithmic uncertainties and risks of inaccuracies, credit remains nevertheless due to the role of the human element as a contributing remedy. In fact, it is assumed in the present thesis, as described in the introduction, that There are nevertheless nuances to highlight here. First, on the assumption that because humans obviously remain the sole subjects of the law, subsequently a degree of human control by the commander (i.e., the end-user) must be maintained over the use of autonomous weapons systems as they retain legal responsibility through their intent executed by the machine.[567]

This assumption, however, rests on the premise that commanders/end-users are the sole 'human influence' over an algorithm's decision, and that only they carry legal responsibility over the

<https://www.taylorfrancis.com/chapters/edit/10.4324/9781315095080-17/automation-bias-intelligent-time-critical-decision-support-systems-cummings> accessed 19 November 2024

[565] Michael C Horowitz, Lauren Kahn 'Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts' (2024) 68(2) International Studies Quarterly <https://academic.oup.com/isq/article-abstract/68/2/sqae020/7638566> accessed 19 November 2024

[566] Reva Schwartz and others 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence' (2022) NIST Special Publication 1270 <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf> accessed 19 November 2024

[567] Boulanin and others (n 558)

system's outcomes (thus, excluding other actors involved in the technology's lifecycle, including developers and those involved in the testing and evaluation of the programme). As established in the previous chapter, AI-enabled technologies will inherently involve a crowd throughout all stages of their development, and what can be considered as 'developers' also involve a wide range of actors, from software engineers to data scientists, who will inevitably also gain input from others – and most notably their client (e.g., a state's armed forces); thus end-users cannot be the only human influence over the system's performance.

Beyond the diversity in actors, humans will provide input throughout the many stages a system goes through in its development, which include, but are not limited to:[568]

1. Conceptualisation/definition of the system's purpose: The mere identification of the problem in itself will require the involvement and judgment of human developers and their clients. For example, in the context of military targeting, the problem could be 'the identification of targetable fighters in the battlefield'. How 'targetable fighters' will be defined will inevitably reflect some of the assumptions of developers and clients, e.g., what parameters and characteristics would they assume such 'targetable fighters' have that would amount to their positive identification.

2. Choice of learning method: This stage will bear influence on subsequent steps, notably code-writing and testing methods. As covered earlier in this thesis, machine learning methods have a number of learning approaches, all different depending on what developers and potentially their clients are trying to achieve. This could be, for example,

[568] While the list is non-exhaustive, the author has tried to regroup and summarise the key stages where humans will be involved in the development of an AI-enabled programme based not only on literature review, but also and mostly based on discussions the author had with experts and practitioners in the field in various contexts – for example as part of roundtables held under the Chatham House rule, informal engagements, etc, For an example of literature on this, see: David Lehr & Paul Ohm 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 UC Davis Law Review <https://heinonline.org/HOL/Page?handle=hein.journals/davlr51&collection=journals&id=667&startid=&endid=732> (accessed 28 April 2022).

supervised learning methods (e.g., classifying data to certain pre-determined categories based on pre-labelled training data); unsupervised learning methods (e.g., for the algorithm to cluster data in a novel way); or reinforcement learning (e.g., the programme will seek to come up and learn new strategies and approaches to achieve a certain pre-defined reward). The choice will inevitably define, as well, the degree of human involvement in the learning process and the way their intervention is needed: for instance, supervised learning programmes will require humans to label the training data; whereas unsupervised learning programmes will not have humans involved in the labelling of the training data, but more on overseeing the relevance and accuracy of clusters and categories the programme comes up with, for example.

3. Testing, evaluation, verification and validation: Such involvement will include the choice and design of, for example, the testing environment; the testing parameters; success metrics; number of test cycles; as well as the choice and definition on the way forward (i.e., how to take into account feedback from testing – for example change the way the programme is trained, or add further training data, or even discontinue the programme's development); and eventually the final validation of the programme.

4. Deployment: Beyond the decision to deploy (when and how), another important aspect of deployment is the design and development of the interface the end-user will be exposed to while manoeuvring the deployed system. Some of the key considerations include: the choice of functionalities the end-users will have access to (or not); the parameters the end-users can and cannot change; how the developers would be involved post-deployment; training of end-users to use the machine; etc.

In the light of this rich ecosystem of actors involved in the development of an AI-enabled system and the processes surrounding it, the previous chapter covered, in particular, the many

layers of legal considerations relevant to developers. As such, designing human control by end-users in and of itself is not sufficient to upholding legal compliance and accountability.

Second, this argument also rests on the premise that the commander exercising human control over the system has the level of technical literacy needed to exercise influence over its calculations and results; and that the system's interface allows the exercise of such influence. In reality, it may be that the end-users of the AI-enabled programmes, especially those in the battlefield 'supervising' autonomous weapons systems, are relying on a user-friendly interface actually allowing little manoeuvrability because of their limited level of technical literacy. Arguably, an interface allowing the commander to process in time the system's calculations, exercise control over its decision-making as needed – including possibly aborting (parts of) the mission, may very well be just the level of control needed to foster compliance with the applicable laws. In other words, it may be that the commander does not need advanced technical literacy – just enough for them to ensure that the use of these systems comply with IHL. This, however, also means that developers are responsible for designing a user interface that would allow commanders/end-users to exercise control as required by the law with the appropriate level of manoeuvrability; and ensure that the programme functions as intended and reflecting the commander's intent.[569] In other words, a well-thought interface can, in principle, decrease uncertainties and increase accuracy in military decision-making and targeting; conversely, a poorly designed interface will exacerbate uncertainties and thus, decrease chances of accuracy in the conduct of military targeting operations.

Ultimately, AI-enabled systems and their algorithms do not 'write themselves'; and even algorithms designed to write other codes and/or to optimise themselves are programmed,

[569] In line with our discussions, there is increasing interest in looking at the technological aspect of interfaces from a political sciences and humanities perspective, see for example: Pedro Maia 'The Case for Interfaces in International Relations' (2023) 3(3) International Political Sociology <https://academic.oup.com/isagsq/article/3/3/ksad054/7274820?login=false> accessed 16 November 2023

trained and tested by humans. Of course, there are learning processes that may involve fewer human agents (e.g., in the case of unsupervised learning), with their very added-value lying in this 'absence' (or in reality, minimised level) of human intervention. These models remain, however, subject to human testing, further training, evaluation – and rightly so. In this sense, as human influence over the accuracy levels of the system is inevitable, there is a need to examine how such influence over the system's performance, and subsequent limitations (e.g., risks of human error, biases, etc.) translate against the applicable laws; so as to adopt the adequate measures to foster compliance with the law in the spirit of protecting civilians. Amidst all this, the right balance must be struck between a non-idealised view of 'the human' on the one hand, and the need to steer away from unnecessary and harmful techno-solutionism on the other hand.[570]

1.3 AI and the Black Box Issue

1.3.1 Introduction: What is the black box issue?

In addition to the probabilistic nature of AI, another key feature of most, if not all of these technologies, is their 'black box'. Many machine learning models' functions are so inherently complex that it has become too complicated, or even impossible, for any human to unpack and comprehend fully the calculations. In other words, black box models are those where no human can understand how the programme came up with its results because it is so, deeply complicated that it has become opaque – even to their own developers. Opening the black box has even been likened, to a certain extent, to understanding, in neuroscience, the networks inside the human brain, which current scientific research and knowledge is not even close to understanding either.[571] The black box issue has been widely discussed in many disciplines and

[570] Beyond the field of AI-enabled technologies for military targeting, see the following for a discussion on techno-solutionism and views on 'the human': Stefania Milan 'Techno-solutionism and the standard human in the making of the COVID-19 pandemic' (2020) 7(2) Big Data & Society <https://journals.sagepub.com/doi/full/10.1177/2053951720966781> accessed 13 January 2023

[571] Davide Castelvecchi 'Can we open the black box of AI?' (*Nature*, 5 October 2016) <https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731> accessed 19 January 2023

for various AI applications. For example, one of the most widely researched AI application scrutinised against the black box issue relates to autonomous vehicles.[572]

The black box raises a seeming paradox with regards to the question of uncertainty and accuracy. On the one hand, the black box creates or even exacerbates epistemic uncertainty about the model itself, i.e., what it is doing and why it is making certain decisions.[573] This uncertainty makes it difficult to not only verify, but also potentially predict what the system's outcomes would be. In environments where certainty is key, such as the military domain, such uncertainty can cause reluctance and distrust that the machine would, in fact, operate predictably and accurately. On the other hand, and paradoxically, some would argue that the black box nature of models could actually lead to greater accuracy.[574] A study has been conducted to survey the different methods and approaches aimed at overcoming the opacity brought by the black box.[575] At times, certain of these attempts at making the model less opaque would result in the 'sacrificing' of accuracy for the sake of interpretability. It remains however unclear whether there is any correlation at all between opacity and inaccuracy, as the same study established a number of approaches that may allow for interpretability without sacrificing the model's accuracy.

[572] For example: Fabian Utesch and others 'Towards behaviour based testing to understand the black box of autonomous cars' (2020) 12 European Transport Research Review <https://etrr.springeropen.com/articles/10.1186/s12544-020-00438-2> accessed 4 January 2023; Quanxin Zhang and others 'Towards cross-task universal perturbation against black-box object detectors in autonomous driving' (2020) 180 Computer Networks <https://www.sciencedirect.com/science/article/abs/pii/S138912862030606X> accessed 4 January 2023

[573] Omer Faruk Tuna, Ferhat Ozgur Catak, M. Taner Eskil, 'Uncertainty as a Swiss army knife: new adversarial attack and defense ideas based on epistemic uncertainty' (2023) 9 Complex & Intelligent Systems, <https://link.springer.com/article/10.1007/s40747-022-00701-0> accessed 18 October 2024

[574] Manuel Carabantes 'Black-box artificial intelligence: an epistemological and critical analysis' (2020) 35 AI & Society <https://link.springer.com/article/10.1007/s00146-019-00888-w> accessed 19 January 2023

[575] Riccardo Guidotti and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51(5) ACM Computing Surveys <https://dl.acm.org/doi/10.1145/3236009> accessed 18 October 2024

The black box issue can be particularly relevant and problematic for a number of issues in the development of AI systems more than others. While the following list does not seek to be exhaustive, they may include:

- Testing and evaluation: The black box issue may hinder any effort at understanding the results of testing and evaluation processes and as such, would affect the ability for programmers to improve the performance of systems e.g., by increasing rates of confidence and certainty, addressing inaccuracies and biases, as well as optimizing the performance and quality of the system's output.
- Transparency and oversight: Opacity surrounding a system's functioning from the development stages will limit the information available to programmers, procuring entities, and ultimately end-users. A commander who is unable to have certain details regarding the parameters and performance of a system may find it difficult to evaluate when it would be legally permissible for it to be deployed. The inability to understand, was well, how a system's confidence and certainty rates came to be could also affect the commander's ability to exercise judgment as to whether or not the system's outputs are indeed reliable.
- Duty to investigate: The inability to access the inner workings of a system can stand in the way of States' ability to conduct effective investigations in situations where there may be violations of IHL. This will be discussed in further detail later in this Chapter.[576]

1.3.2 Proposed solutions to 'opening' the black box

[576] Chapter 4, Section 3.3

Efforts to 'open' the black box have primarily revolved around two concepts: transparency, and explainability.[577] The latter approach focusing on building 'explainable' AI has been the subject of research from as early as the 1970s.[578] In essence, explainability seeks to enable humans to understand the system's output (e.g., decisions or predictions); hence through this understanding, a person should be able to 'assess whether this output was based on valid criteria or not.'[579] In the same vein, algorithmic transparency consists of making transparent the factors informing and influencing the system's calculations and subsequent output, such as for example the training data it has been subject to (in line with discussions earlier on 'garbage in, garbage out').[580]

These two concepts have indeed been widely adopted by many disciplines: reaching beyond the field of computer science, explainability and transparency lie at the heart of discussions and deliberations in ethics, policymaking, as well as the legal field. In fact, while this chapter will, later, examine more in detail the IHL implications of 'not knowing' with regards to the development of military AI, the black box issue more generally (i.e., beyond the context of military applications) also lies at the heart of reflections in other branches of the law. For example, in the context of EU law, the General Data Protection Regulation (GDPR) provides

[577] Amina Adadi, Mohammed Berrada 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' (2018) 6 IEEE Access <https://ieeexplore.ieee.org/abstract/document/8466590/> accessed 4 January 2023

[578] See for example A. Carlisle Scott and others 'Explanation Capabilities of Production-Based Consultation Systems' (1977) American Journal of Computational Linguistics <https://aclanthology.org/J77-1006/> accessed 23 February 2023, as cited in Feiyu Xu and others 'Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges' (Natural Language Processing and Chinese Computing 2019) <https://link.springer.com/chapter/10.1007/978-3-030-32236-6_51> accessed 23 February 2023

[579] Leilani H. Gilpin and others 'Explaining Explanations to Society' (Proceedings of the NeurIPS 2018 Workshop on Ethical, Social and Governance Issues in AI, Montréal, Québec, Canada, 2018), <https://arxiv.org/abs/1901.06560> accessed 23 February 2023, as cited in Timo Speith 'A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods' (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3534639> accessed 23 February 2023

[580] Nicholas Diakopoulos & Michael Koliska 'Algorithmic Transparency in the News Media' (2017) 5(7) Digital Journalism <https://www.tandfonline.com/doi/abs/10.1080/21670811.2016.1208053?journalCode=rdij20> accessed 23 February 2023. For the saying 'garbage in garbage out', it is relevant here as, to a certain extent, knowing what the programme has been fed with can provide some form of accountability and explainability with regards to its outputs. See Section 1, Chapter 2

for a right to explanation for those individuals subject to automated decision-making through 'meaningful information about the logic involved' in these processes.[581] The highly-anticipated and now adopted EU AI Act, which many see as a pioneering regulatory tool in the sphere of AI, is expected to put an even greater emphasis on transparency as a response to the black box issue. This is particularly the case for 'high-risk AI systems', which are subject to a higher standard and requirement for transparency than those systems considered as of 'non-high-risk'.[582]

It is however worth noting that these two approaches are far from constituting the perfect remedy to issues surrounding the opacity of models. In fact, a number of limitations have been identified to this approach. For example, the complexity and sheer number of data, as well as the algorithm in itself, in most of today's AI-enabled systems would make it difficult, if not impossible, to design these systems are inherently explainable – or in other words, explainable by design.[583] There is also the general assumption that there is a certain trade-off between the need for transparency on the one hand, and accuracy of the system on the other hand.[584] This circles back to the first point made about increasingly complex programmes constituting most of the systems deployed today, which provide a high rate of accuracy, yet are inherently opaque. Regardless of whether this assumption is indeed well-grounded in scientific evidence or whether it is a myth, in response, plenty of research focuses on developing methods to make

[581] European Union (EU) Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) [2016] OJ L119/1, art 13(2)(f), see also Andrew D Selbst, Julia Powles 'Meaningful information and the right to explanation' (2017) 7(4) International Data Privacy Law <https://academic.oup.com/idpl/article/7/4/233/4762325> accessed 23 February 2022; and Jarek Gryz, Marcin Rojszczak 'Black box algorithms and the rights of individuals: No easy solution to the "explainability" problem' (2021) 10(2) Internet Policy Review <https://www.econstor.eu/handle/10419/235967> accessed 23 February 2023

[582] EU AI Act (n 23)

[583] Marzyeh Ghassemi, Luke Oakden-Rayner, Andrew L Beam 'The false hope of current approaches to explainable artificial intelligence in health care' (2021) 3(11) The Lancet Digital Health <https://www.sciencedirect.com/science/article/pii/S2589750021002089> accessed 23 February 2023

[584] Fatimah Ishowo-Oloko and others 'Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation' (2019) 1 Nature Machine Intelligence <https://www.nature.com/articles/s42256-019-0113-5> accessed 23 February 2023

these systems explainable post-hoc.[585] This approach, however, comes with its own set of criticisms. These include the fact that this post-hoc approach makes explainability more of an afterthought while the focus should be to make algorithms interpretable by design – especially for those with high societal stakes; explanations may not necessarily be aligned and faithful to what the original model computes; and explanations are often not providing enough details and only constitute the tip of the iceberg on the black box model's calculations.[586]

1.3.3 The Black Box in the Military Domain

Building on the previous section, studies seeking to 'open' the black box are on-going; however, the state of affairs is such that, today, completely opening the black box is not feasible. As such, military AI technologies are bound to maintain a degree of opacity. Two main issues subsequently arise, the legal implications of which will be unpacked in further detail later in this chapter.

1. **Uncertainty acceptance**

First, there is the question of how much uncertainty are military forces ready to accept, whether it is with regards to operational reliability or legal compliance. In fact, as accurate as a system can be, the extent to which the procuring forces would accept the uncertainties that come with its black box is uncertain. In other words, the black box raises the question of whether the end justifies the means. It is argued here that accuracy is not enough both with regards to operational reliability, but also in the spirit of fostering legal compliance. More specifically, this is due to the black box of those systems and the obstacles it brings to improve the system's performance.

[585] Daniel Vale, Ali El-Sharif, Muhammed Ali 'Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law' (2022) 2 AI and Ethics <https://link.springer.com/article/10.1007/s43681-022-00142-y> accessed 23 February 2023

[586] Cynthia Rudin 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' (2019) 1 Nature Machine Intelligence <https://www.nature.com/articles/s42256-019-0048-x> accessed 23 February 2023

Before a more detailed discussion on the military domain specifically, it is worth mentioning perhaps the most famous and among the earliest examples of the black box issue: Dean Pomerleau's Humvee military vehicle, equipped with a computer, which he sought to programme to learn to drive autonomously. Trained on images collected through a camera, the computer was designed to process and interpret what it saw as Pomerleau took the vehicle out for a drive through city streets; with the hope that the programme will eventually learn how to steer the vehicle on its own through the streets of Pittsburgh.[587] Through an adaptive neural network, Pomerleau's aim was for the algorithm to learn to adapt to unprogrammed scenarios based on the data it has collected from camera, on how a human would react under similar circumstances; instead of the developers having to patch the programme every time it was encountering situations that were unfamiliar to it.[588] The test drives went well until one incident, when the vehicle 'approached a bridge and suddenly swerved to one side'.[589] The crash was eventually avoided by retaking control over the steering wheel in time. Upon studying the incident in its aftermath, Pomerleau was met with the black box issue: due to the neural network nature of his algorithm, it was difficult to understand how it processed and diffused the information it was fed with. Through extensive testing, Pomerleau eventually managed to 'open' the black box: 'the network had been using grassy roadsides as a guide to the direction of the road, so the appearance of the bridge confused it.'[590]

Pomerleau's incident dates way back to 1991; and since then, not only has the number of opaque models skyrocketed, many commercial applications and scientific research today heavily depend on them. In the banking field, for example, a lot of programmes depend on artificial neural networks to establish someone's loan eligibility. In addition to obvious

[587] Castelvecchi (n 571)
[588] Shubham Agarwal 'How a big blue van from 1986 paved the way for self-driving cars' (*Digital Trends*, 6 February 2022) <https://www.digitaltrends.com/cars/first-self-driving-car-ai-navlab-history/> accessed 20 January 2023
[589] Castelvecchi (n 571)
[590] Ibid

variables that would indicate their eligibility (e.g., income history, type of employment, criminal records etc.), the programme will have found others that are not so obvious in themselves, but are nevertheless highly informative and add to the accuracy of the programme's estimation due to the complex – and subsequently opaque – correlations discovered by its algorithms.[591]

Aside from challenges to legal compliance, which will be covered in further detail later, Pomerleau's case study reveals major challenges with regards to the operational reliability of military AI. One of these challenges pertains to risks assessments in the testing, evaluation, and decision to deploy those systems. In fact, at the procurement stage and before these systems are even deployed, the procuring armed forces will need to conduct a series of testing and evaluation to assess their performance. In addition, risks assessments must be conducted with regards to their intended use, whether it is from an operational or legal standpoint. Assessed risks may include, but would not be limited to, those of malfunction, the system's performance and limitations in specific environments, as well as compliance with the law (Article 36 legal reviews). These risks assessments, while not necessarily unique to AI, are particularly important due to these technologies' probabilistic nature and the inherent uncertainties that come with it. Yet, the development and subsequent testing and evaluation of these technologies generally tend to be done in highly controlled settings, and they may not necessarily perform well in real-life especially in environments as dynamic, messy and unpredictable as warfare.[592] A system tested a thousand times and exhibiting a 100% accuracy rate would not necessarily be as performing in real-life, thus creating uncertainties with regards to the system's reliability

[591] Carabantes (n 574)

[592] The issues raised by discrepancies between controlled training and testing environments on the one hand, and real-world deployment on the other hand, are not necessarily unique to the military domain. However, it is argued that due to war's dynamic, messy and unpredictable nature, associated risks from discrepancies in performance are further exacerbated. For a general account on the challenges arising from these differences, see: Andrei Paleyes, Raoul-Gabriel Urma, Neil D. Lawrence 'Challenges in Deploying Machine Learning: A Survey of Case Studies' (2023) 55(6) ACM Computing Surveys <https://dl.acm.org/doi/10.1145/3533378> accessed 18 October 2024

in combat. As such, it would be important for procuring forces to ascertain ways of measuring this uncertainty by anticipating and calculating the rate of such discrepancies. Procuring forces should then consider the level of uncertainty they would be willing to accept – with such assessments done notably with IHL considerations in mind.[593] However, once deployed, if the system behaves unpredictably and a similar incident to that of Pomerleau occurs (i.e., it is uncertain why the system behaved in such way that it causes the incident), this will have operational repercussions. When incidents occur, the system would generally be suspended from use until the source of the problem has been addressed; however, in the context of black box models, this could potentially prove to be impossible.

Another and closely related challenge pertains to the continued reliability of reinforcement learning models. As discussed in the previous chapter, the performance of reinforcement learning models is likely going to change from the moment they are being tested.[594] Not only is this due to discrepancies between testing and real-life environments, but particularly due to their ever-learning and ever-evolving nature. In the previous chapter, technical solutions to flagging such changes were discussed. However, another question that arises is – once the performance of a reinforcement learning model has been flagged as having steered away too much from the one measured at the evaluation and approval stage, the system will need to be evaluated on its reliability in the light of its updated performance. Yet, there may be challenges as to how this assessment can be conducted if the model is also a black box. In fact, a number of studies suggest that reinforcement learning models are also likely more opaque due to their complexity, thus exacerbating the challenges in ensuring they remain continuously reliable from an operational standpoint.[595]

[593] Acceptability thresholds are discussed in Sections 2.2.1 and 3.2 of this chapter.
[594] Chapter 3, section 4.2.2
[595] George A. Vouros 'Explainable Deep Reinforcement Learning: State of the Art and Challenges' (2023) 55(5) ACM Computing Surveys <https://dl.acm.org/doi/10.1145/3527448> accessed 19 October 2024; see also Claire Glanois and others 'A survey on interpretable reinforcement learning' (2024) 113 Machine Learning

2. **Military decision-making under uncertainty**:

The second main issue relates to the uncertainty the black box issue further adds to military decision-making. As discussed in the previous section, warfare is inherently complex and shrouded in uncertainties, and the black box could potentially exacerbate those. The opacity of black box models will inevitably cause trust issues which, in many (if not most) areas, could be problematic. In the medical field, for example, physicians are generally expected to not only provide diagnoses and prescriptions, they must also be able to provide an explanation as to why they have taken certain courses of actions, and justify their decisions.[596] This transparency requirement is justifiable for a number of reasons – whether it is in terms of morality, ethics, and the patients' right to information, or with regards to accountability in situations where there may have been medical negligence – with the latter potentially resulting in life-altering outcomes for patients, or even be a matter of life and death. Diagnoses informed by black box models may be problematic in this case, as they offer no transparency over the results that may have informed the contentious decision-making in question. In the field of public policy, the increasing deployment and use of AI-enabled technologies for surveillance and, more generally, as part of transitions towards 'smart cities' not only raises questions with regards to the implications of such opacity and its democratic legitimacy; it is also source of concern with regards to over-reliance on these technologies all while it exacerbates and intensifies deeply

<https://link.springer.com/article/10.1007/s10994-024-06543-w> accessed 19 October 2024; and Alexandre Geuillet, Fabien Couthouis, Natalia Díaz-Rodríguez 'Explainability in deep reinforcement learning' (2021) 214 Knowledge-Based Systems <https://www.sciencedirect.com/science/article/pii/S0950705120308145> accessed 19 October 2024

[596] William R Swartout 'XPLAIN: a system for creating and explaining expert consulting programs' (1983) 21(3) Artificial Intelligence <https://www.sciencedirect.com/science/article/abs/pii/S0004370283800149> accessed 20 January 2023

rooted social injustices without possibility of oversight and scrutiny due to the black box issue.[597]

Similarly, in the military domain, black box models can raise issues with regards to their reliability, and the risks of introducing more uncertainties in decision-making. Ultimately, the black box issue is about the development, adoption, integration, deployment and use of a system known to be inherently uncertain, and the extent to which the users, facing the system's opacity, would be expected to have the ability to provide explanations. Additionally, the uncertainty issues brought by the black box are, in the end, non-binary: it is not about whether a decision support system is completely opaque or, conversely, completely transparent to the user. It is rather much more nuanced, particularly in the light of the paradox brought by black boxes with regards to accuracy rates (or at least, assumptions of accuracy) discussed earlier.[598] This raises, yet again, the question of whether this uncertainty is, in and of itself, an issue with regards to legal compliance, which will be discussed in further detail later in this chapter.

2 Uncertainty in War and expectations for accuracy: What does the law say?

2.1 War: An inherently uncertain environment

War, by nature, is inherently complex and shrouded with uncertainty. In the US, the Army War College has attempted to describe environments of armed conflict as characterized by 'volatility, uncertainty, complexity and ambiguity' (VUCA).[599] In fact, war has always been inherently uncertain – and reflections on the unknown in a military context precedes way back the 1980s when the US Army War College coined the aforementioned acronym. In fact, even tracing back to Sun Tzu's work in the 5th Century Before Christ (BC), he referred to war as varying by nature – and that both orthodox and unorthodox attacks are bound to be endlessly

[597] Gavin J D Smith 'The politics of algorithmic governance in the black box city' (2020) Big Data & Society <https://journals.sagepub.com/doi/pdf/10.1177/2053951720933989> accessed 23 February 2023
[598] Section 1.3.1
[599] Benjamin E Baran, Haley M Woznyj 'Managing VUCA: The human dynamics of agility' (2020) 20 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7439966/> accessed 2 February 2023

varying as well.[600] In the 19th century, Clausewitz went even further: in his much-cited book 'Vom Kriege', among the many metaphors used to characterise war, he refers to a 'fog' in the light of war being 'the realm of uncertainty', whereby 'a sensitive and discriminating judgment is called for; a skilled intelligence to scent out the truth.'[601] He goes further, in his reflection, by adding the element of chance as also inherent and 'incessant' to war, making 'everything more uncertain and interferes with the whole course of events.'[602] Clausewitz essentially argues that in the light of these inevitable uncertainties (as they are inherent to war) in the battlefield, commanders and decision-makers must adapt accordingly, knowing that 'all information and assumptions are open to doubt' against the uncertainty that characterizes war.[603]

Beyond Sun Tzu, Clausewitz and the US Army War College's VUCA concept, more recent scholarship has continued this centuries-long reflection on war and uncertainty. From strategic planning to decision-making in the battlefield, military operations will always hold a certain element of uncertainty that stretches far beyond the battlefield. [604] From the intelligence collection and analysis stages, and even post-operation in the context of feedback and reviews, uncertainty will always be present and affecting not only the way decisions are made, but also how information – and more generally, the conflict – is perceived, interpreted and processed.

In fact, some scholars have made the case for uncertainty as a necessary element in war and conflict. Uncertainty and the subsequent fear of (or at least, concern from) not knowing is precisely what pushes decision-makers to err on the side of caution.[605] In other words,

---

[600] Sun Tzu (n 231) 70-17, as cited in Bleddyn E Bowen 'The Application of Force and Strategy in Sun Tzu and Clausewitz' (*E-International Relations*, 16 December 2010) <https://www.e-ir.info/2010/12/16/the-application-of-force-and-strategy-in-sun-tzu-and-clausewitz/> accessed 2 February 2023

[601] Clausewitz, Carl von, Vom Kriege (Book 1, 1832) 101

[602] Ibid

[603] Ibid, 102

[604] Talbot C. Imlay, Monica Duffy Toft (eds) *The Fog of Peace and War Planning: Military and Strategic Planning under Uncertainty* (Routledge, 2006) 4

[605] For example, see: David Whetham, Kenneth Payne 'AI: In Defence of Uncertainty' (*Defence-In-Depth*, 9 December 2019) <https://defenceindepth.co/2019/12/09/ai-in-defence-of-uncertainty/> accessed 19 February 2023

uncertainty can be argued as an enabler for sound, cautious decision-making and even, arguably, at times, is the ultimate barrier against potentially catastrophic outcomes. This was the case, for example, during the Falklands crisis: in 1982, the British Admiral Woodward decided not to engage with a detected aircraft suspected to be a Boeing part of the Argentinian fleet despite all authorisations and legal permissions granted. Uncertainty pushed Woodward to double check and re-confirm the positive identification of the aircraft in question, and eventually led him to command not to fire 20 seconds short of missile launch upon confirmation that the aircraft was, in fact, a Brazilian civilian airliner running from Durban to Rio de Janeiro.[606]

This situation is very similar to the incident involving Lieutenant Colonel Petrov, a Soviet officer who is known to have adverted nuclear war in the following year. On 26 September 1983, Petrov, stationed in Serpukhov-15 at the Oko early-warning satellite system's control console, was met with a sudden, launch warning indicating that a total of five intercontinental ballistic missiles were launched from the United States.[607] Petrov, despite the command console exhibiting the highest level of confidence in the reliability of its assessment, decided to report the notification as a false alarm.[608] Petrov allegedly had a number of considerations that informed and influenced his final decisions – but, in the end, in response to uncertainty, he had to rely on his gut feeling, part of the 'second brain' and the gut-brain axis, currently still poorly understood.[609] As such, his response to uncertainty was, in a way, a black box. This

[606] T Van Baarda 'Chapter 4: Ethics in the Royal Netherlands Navy' in T Van Baarda, D E M Verweij (eds) *Military Ethics, The Dutch Approach – A Practical Guide* (Brill | Nijhoff, 2006) <https://brill.com/display/book/edcoll/9789047411253/Bej.9789004154407.i-395_005.xml> accessed 23 November 2024

[607] G Forden, P Podvig, T A Postol 'False alarm, nuclear danger' (2002) 37(3) IEEE <https://ieeexplore.ieee.org/document/825657> accessed 19 February 2023

[608] Unal and others (n 479)

[609] See: Justin Sonnenburg & Erica Sonnenburg 'Gut Feelings–the "Second Brain" in Our Gastrointestinal Systems [Excerpt]' (*Scientific American*, 1 May 2015) <https://www.scientificamerican.com/article/gut-feelings-the-second-brain-in-our-gastrointestinal-systems-excerpt/> accessed 19 February 2023 and Marilia Carabotti and others 'The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems' (2015) 28(2) Annals of Gastroenterology <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4367209/> accessed 19 February 2023

incident is the perfect, textbook case study often used in the nuclear weapons policy sphere to illustrate not only a concrete example of nuclear near-miss, but also the extent to which major, high-stakes decisions with possibly catastrophic outcomes – including in the realm of nuclear weapons – are shrouded with uncertainty.

Yet, the opposite could also be argued where uncertainty could add stress to human judgment and push for destructive behaviour. If, for example, Admiral Woodward succumbed to pressure from the pre-existing tensions brought by the Falkland crisis, as well as the uncertainties brought by the unknown identity of the aircraft – uncertainty could have led him towards adopting destructive behaviour that would have led to the shooting of the Brazilian civilian airliner and thus, the death of civilians due to the unknown. Similarly, if Lieutenant Colonel Petrov reacted to the uncertainties brought by the situation in a way that would push him towards recommending a pre-emptive strike in response to the threat presented, this could have potentially led the world towards a catastrophic outcome marked by a global nuclear war.

Uncertainty is thus inherent to conflict and is present across all levels of decision-making, including those with the highest level of impact. Regardless of how one perceives uncertainty – as either a liability or an asset/a safeguard measure forcing commanders to err on the side of caution, it essentially boils down not to a question of eliminating uncertainty, but rather acknowledging its inevitably and adapting accordingly, cognisant of the subsequent risks that come with it. It is not a question of 'clearing' the fog, but rather navigate through relative blindness and uncertainties.

The same must therefore be said with regards to the development and deployment of AI: echoing what has been said earlier, AI technologies can in principle help address uncertainties in decision-making; thus, hold the potential to at least, in theory, increase accuracy in the execution of certain actions, such as target engagement. Uncertainties surrounding a cluttered environment can be decreased though AI-enhanced situational awareness and thus help find,

select and engage the desired target with higher precision rates than without the use of AI solutions. These precision rates, however, can never be perfect: as mentioned earlier, AI systems are ultimately probabilistic by nature and absolute certainty can never be achieved. In a way, the development and deployment of AI-enabled technologies in armed conflict can be described as trying to decrease uncertainty from an inherently uncertain environment (war) using a tool that, by nature, retains a degree of uncertainty in its results.

2.2 Uncertainties and Expectations for Certainty in Targeting: An International Humanitarian Law Perspective

In the light of inevitable uncertainty in armed conflict, there have been a number of attempts to clarify as to how the law applies and should be interpreted in situations where uncertainty prevails and adjacent questions – notably the issue of accuracy in targeting. The conduct of military targeting in armed conflict is indeed framed by a number of rules in international humanitarian law, such as the rule of distinction, proportionality, and precautions in attack, and the use of AI-enabled technologies for such operations is also subject to these rules, as discussed across this thesis. Yet, how the law applies in the light of the inherent uncertainties stemming from the messy nature of conflict is source of much discussion: in the context of military targeting, for example, this uncertainty can affect assessments with regards to the legality of an individual target. [610] While this chapter will unpack, more in detail, how this unfolds in the context of AI development and deployment, the present section will first take stock on current, general discussions with regards to how the law applies in military targeting and whether there is, indeed, a legal expectation for certainty.

[610] See for example Adil Ahmad Haque 'Killing in the Fog of War' (Proceedings of the ASIL Annual Meeting , Volume 106 , 2012) <https://www.cambridge.org/core/journals/proceedings-of-the-asil-annual-meeting/article/abs/killing-in-the-fog-of-war/209F6440BEC661F3DE6F813B07BBED10> accessed 9 December 2022

Military targeting operations involve 'a multifaceted situational assessment when planning, approving or executing attacks.'[611] Agents and variables involved in targeting environments are diverse, interconnected – and at times, interdependent – forming a complex ecosystem and, echoing our previous discussions, inevitably shrouded by uncertainty. While not a new phenomenon, the increasing urbanisation of conflict – where town, cities, and other urban settings are the primary battlegrounds for armed conflicts – add further layers of complexity to war.[612] On the basis that the more complex an environment is, the more uncertainties there will be, it must be assumed that they will significantly affect the way risk and legality assessments will be conducted, and the way IHL will be subsequently applied.

To a certain extent, part of the law surrounding the conduct of hostilities could easily be perceived as almost 'mechanical', perhaps even transcribable into a decision tree or at least, a checklist. This 'dry' nature of IHL allows for clarity and is an effective response to the uncertain and highly variable environments IHL ought to be applied. For example, the rule of distinction has a very specific set of characteristics on what constitutes a 'military objective', i.e., any object targetable under IHL: 'military objectives are limited to those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.'[613]

The definition provided by API, generally argued as reflective of customary law and thus, binding to states non-party to API, provides a list of characteristics that must be met in order for objects to be considered as a military objective and thus, be lawfully targetable. Objects

[611] Michael N Schmitt, Michael Schauss 'Uncertainty in the Law of Targeting: Towards a Cognitive Framework' (2019) 10 Harvard National Security Journal <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3755556> accessed 2 February 2023
[612] Laurent Gisel and others 'Urban warfare: an age-old problem in need of new solutions' (*Humanitarian Law & Policy*, 27 April 2021) <https://blogs.icrc.org/law-and-policy/2021/04/27/urban-warfare/> accessed 2 February 2023
[613] API, art 52(2)

failing to meet these criteria must be considered as a civilian object and thus, cannot be directly targeted. This rather simple, yet comprehensive and context-dependent checklist API provides clarity on a subject bound to be shrouded by uncertainty. For example, with dual-use objects (i.e., objects serving both civilian and military purposes, such as infrastructure, power-generating stations, and technologies including computer hardware and satellites), the characteristics set out by API are clear: even if the object in question is also used for civilian purposes, the moment it meets the criteria for military objectives, then it becomes targetable, even if its military use is secondary.[614] Conversely, if the 'effective contribution to military action' is hard to justify, 'a general label is insufficient' and, as the ICTY asserted, if the criteria for military objectives are not met, their targeting would simply be unlawful.[615] It also means that this assessment is dynamic and subject to changes based on the circumstances ruling at the time: In other words, if the destruction of an object considered as a military objective does not offer 'a definite military advantage' anymore because circumstances have changed, then it is not targetable anymore under IHL.

On paper, the well-established and widely recognized rules surrounding the conduct of hostilities provide a certain clarity. However, they are not, in and of themselves, silver bullets to all the uncertainty that may arise during conflict. In fact, the characteristics set out, for example, in establishing whether an object is a military objective (and subsequently targetable) leave room for interpretation – which will inevitably be informed by facts on the ground available, as well as the uncertainties surrounding them. They also require a degree of subjective assessment coming from whomever is in charge of interpreting the law, for example on the 'effective' nature of the contribution to military action from the destruction of the object. Risk and legality assessments will thus inevitably be affected by uncertainty: decisions are

[614] Sassòli (n 370)
[615] ICTY (n 380)

indeed not only based on evidence and facts, but also 'other complex factors, including perception and human judgment, and their influencing factors.'[616] Uncertainty – including from the lack of certain facts and evidence – will inevitably affect both perception and human judgment. Absolute certainty can never be fully expected.

In this sense, there have been a number of attempts at clarifying the law surrounding military targeting and the level of accuracy expected. The interest particularly grew in the light of technological advancements in precision targeting from the early 2000s, as well as the rise of 'targeted killings' notably in the wake of 9/11 by US forces.[617] As states' doctrines and policies on targeting practices had started to emerge, so did academic discussions that looked into clarifying the law, including who and what is targetable both in international and non-international armed conflict, the permissible choice of means and methods of warfare, the precautionary measures to be taken, and reparation mechanisms, among others.[618] As technological progress is proceeding apace and targeting practices evolved alongside, so did

[616] Unal and others (n 479)

[617] For a detailed discussion on targeted killing under international law, see notably Nils Melzer, *Targeted Killing in International Law* (Oxford, 2008). See also Caitlan McNamara 'Targeting Decision sin the Crosshairs of Humanitarian Law and Human Rights Law' (2014) 5(2) <https://heinonline.org/HOL/Page?handle=hein.journals/creintcl5&div=15&id=&page=&collection=journals> accessed 8 November 2023

[618] In the wake of 9/11, there were notably the US' Joint Doctrine for Targeting: Joint Chiefs of Staff Washington, DC (US), *Joint Doctrine for Targeting, Joint Publication 3-60* (2002) <https://apps.dtic.mil/sti/pdfs/ADA434278.pdf> accessed 9 November 2023 and Joint Warfighting Center & Joint Warfighters Joint Test and Evaluation (US), *Commander's Handbook for Joint Time-Sensitive Targeting* (2002) <https://apps.dtic.mil/sti/pdfs/ADA403414.pdf> accessed 9 November 2023. Another notable case was the Targeted Killings Case in Israel from 13 December 2006, see: 'Israel, The Targeted Killings Case' (*ICRC*) <https://casebook.icrc.org/case-study/israel-targeted-killings-case> accessed 9 November 2023. For examples of commentary, see: Michael N Schmitt, Eric Widmar 'The Law of Targeting' in Paul A.L. Ducheine, Michael N. Schmitt, Frans P.B. Osinga (eds) *Targeting: The Challenges of Modern Warfare* (Springer, 2016) <https://link.springer.com/chapter/10.1007/978-94-6265-072-5_6> accessed 8 November 2023; Michael N Schmitt 'Targeting and Humanitarian Law: Current Issues' (2004) 34 Israel Yearbook on Human Rights <https://brill.com/display/book/edcoll/9789047414070/B9789047414070_s005.xml> accessed 9 November 2023; Michael Elliot 'Where Precision Is the Aim: Locating the Targeted Killing Practices of the United States and Israel within International Humanitarian Law' (2009) 47 Canadian Yearbook of International Law <https://heinonline.org/HOL/LandingPage?handle=hein.journals/cybil47&div=6&id=&page=> accessed 9 November 2023; Michael N Schmitt 'Targeting and International Humanitarian Law in Afghanistan' (2009) 39 Israel Yearbook on Human Rights <https://brill.com/display/book/edcoll/9789047443773/BP000005.xml> accessed 9 November 2023; and Chinwe Patricia Iloka 'Precision Attack and Reparation of the Vulnerable under International Humanitarian Law: An Appraisal' (2022) 7 African Journal of Criminal Law and Jurisprudence <https://heinonline.org/HOL/LandingPage?handle=hein.journals/afcjlocil7&div=21&id=&page=> accessed 9 November 2023

academic research and commentaries which increasingly looked into the implications of new and emerging technologies with regards to targeting practices, such as the targeting of data and, more increasingly so, the use of AI technologies.[619]

There remain, at present, a number of contentious points with regards to the application of the law under certain circumstances that may arise in military targeting operations still very much subject to heated debates both by states themselves, but also in academic discussions. These disagreements can arise for a number of reasons, including differences in national postures, approaches and policies, but also the interpretation and subsequent application of the law[620], as well as its applicability (e.g., the status of private, security contractors sent by states).[621]

One key, contentious question is whether IHL would indeed expect a certain level of accuracy in the conduct of targeting operations. In fact, whether IHL does indeed prescribe, at all, a duty for accurate precision/precision targeting is subject to much debate.[622] Specifically, there are

---

[619] On data, see: Tim McCormack 'International Humanitarian Law and the Targeting of Data' (2018) 94 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1725&context=ils> accessed 9 November 2023 and Noam Lubell 'Lawful Targets in Cyber Operations: Does the Principle of Distinction Apply?' (2013) 89 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1026&context=ils> accessed 9 November 2023; on AI see for example the use of AI to enhance IHL compliance in military targeting: Peter Marguiles 'The Other Side of Autonomous Weapons: Using Artificial Intelligence to Enhance IHL Compliance' in Robert T P Acala, Eric Talbot Jensen (eds) *The Impact of Emerging Technologies on the Law of Armed Conflict* (Oxford University Press, 2019) <https://academic.oup.com/book/32410/chapter-abstract/268714567?redirectedFrom=fulltext> accessed 9 November 2023, as well as to support kinetic targeting: Anastasia Roberts, Adrian Venables 'The Role of Artificial Intelligence in Kinetic Targeting from the Perspective of International Humanitarian Law' (13th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2021) <https://ieeexplore.ieee.org/abstract/document/9468301> accessed 9 November 2023

[620] This also the raises the question of differences in rules of engagement between armed forces, which serve as 'translation' of the boundaries and limitations set by international humanitarian law to frame the conduct of hostilities in the battlefield. See for example: Marc Warren 'The "Fog of Law": The Law of Armed Conflict in Operation Iraqi Freedom' (2010) 86 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1103&context=ils> accessed 10 November 2023

[621] Michael N Schmitt 'The Interpretive Guidance on the Notion of Direct Participation in Hostilities: A Critical Analysis' in Michael N Schmitt, *Essays on Law and War at the Fault Lines* (Springer, 2012) <https://link.springer.com/chapter/10.1007/978-90-6704-740-1_10> accessed 10 November 2023

[622] See most notably Michael Schmitt's work on the law surrounding precision targeting: Michael N Schmitt & Eric W Widmar '"On Target": Precision and Balance in the Contemporary Law of Targeting' (2014) 7(3) Journal of National Security Law and Policy <https://heinonline.org/HOL/LandingPage?handle=hein.journals/jnatselp7&div=18&id=&page=> accessed 9 December 2022 and Michael N Schmitt, John J Merriam 'The Tyranny of Context: Israeli Targeting Practices in Legal Perspective' (2015) University of Pennsylvania Journal of International Law <https://heinonline.org/HOL/LandingPage?handle=hein.journals/upjiel37&div=5&id=&page=> accessed 9 December 2022

two, components to the question of accuracy and precision: 1) Accuracy in target identification (i.e., the level of accuracy of the intelligence collected surrounding the identified target); and 2) Accuracy in target engagement (i.e., the level of accuracy and precision when engaging with the target, generally of the weapon system and ammunition employed).

2.2.1 Accuracy in target identification

The first, relevant issue with regards to accuracy in military targeting is the question of target identification. More specifically, it is about ensuring accuracy in, first, their identity, but also, second, in their legal status as to whether or not they would actually be targetable under applicable laws. Both go hand-in-hand as, given the number of complexities and nuance international humanitarian law provides, a person or an object's status does not always equate with the legality (or, conversely, illegality) of their targeting. Under the rule of distinction, if a person has been identified as a combatant, it does not necessarily always mean that they are targetable, for example if they were *hors de combat*, e.g., by surrendering.[623] Conversely, a person identified as a civilian will not necessarily always be protected if, for example, it has been established that they are directly participating in hostilities at the time of the attack.

The question is therefore – to what extent does the law expect these assessments with regards to the target's identification, and legality, to be accurate; and to what extent is there room for mistakes in assessments with regards to the legality of a target. Further complications may arise with regards to legal assessments and the application of the law on the ground due to uncertainties in the battlefield and variables that may arise and affect the legality (or, conversely, the illegality) of an attack. The veracity and accuracy of these legal assessments will highly depend on the context on the ground; and echoing previous points, the messy and dynamic nature of warfare may sow doubts with regards to the legal status of persons and

[623] Geneva Convention I, art 3; Geneva Convention II; art 3; Geneva Convention III, art 3; Geneva Convention IV, art 3

objects. This is particularly the case for ambiguous statuses, including, yet again, the interpretation and application of the rule of distinction in the context of civilians directly participating in hostilities – not only with regards to ascertaining their status but also the circumstances under which their targeting would be lawful or, conversely, unlawful.[624]

One solution IHL provides is its protective stance in case of doubt: the person or the object is to be treated as of civilian nature that, until proven otherwise, shall remain immune to being the object of direct targeting.[625] This can be drawn both from statutory law and case law. In fact, the Article 50 of API provides a negative definition of civilians that work on the assumption that 'any person' is considered as such unless they fall under the categories provided by Article 4(A)(1), (2), (3), and (6) of the 1949 Geneva Convention III, as well as members of armed forces as defined in Article 43 of API. The same protective approach is provided for the targeting of objects: Article 57 API, in providing the scope for the rule of precautions in attack, also ascertains the need for positive identification and verification that the targeted objective is, first, not a civilian object, nor subject to a special protection or prohibition from attack, 'but are military objectives' as defined in Article 52(2). This protective approach has been further re-ascertained, by the ICTY, which reaffirmed the authority of civilian presumption outlined in Article 50(1) API.[626] It is however important to note and add the caveat that the interpretation and application of such presumption remains subject to much reservation and as such, complexifies and may even put into question its customary nature.[627]

---

[624] Dapo Akande 'Clearing the Fog of War? The ICRC's Interpretative Guidance on Direct Participation in Hostilities' (2010) 59(1) International & Comparative Law Quarterly <https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/abs/ii-clearing-the-fog-of-war-the-icrcs-interpretive-guidance-on-direct-participation-in-hostilities/707794583F0B2CDFFEDDEEF9F5DDCAAF> accessed 2 February 2023

[625] Emanuela-Chiara Gillard 'Sieges, the Law and Protecting Civilians' (2019) Chatham House <https://www.chathamhouse.org/sites/default/files/publications/research/2019-06-27-Sieges-Protecting-Civilians_0.pdf> accessed 19 February 2023

[626] *Prosecutor v Stanislav Galic* (Judgment and Opinion) ICTY IT-98-29-T (5 December 2003) <https://www.icty.org/x/cases/galic/tjug/en/> accessed 19 February 2023 [50]

[627] 'Rule 6: Civilians' Loss of Protection from Attack' (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule6> accessed 3 June 2025

This serves as a useful reminder of the pragmatic nature of IHL: Through its protective stance, the law acknowledges the uncertainties, and doubt in target identification that may arise from conflict, in the light of its messy, dynamic and ever-changing nature. It addresses the issue of inaccuracies by putting the onus, on the attacking party, to demonstrate that the target has indeed been identified as lawful (i.e., either as combatant, civilians directly participating in hostilities, or whether they meet the cumulative criteria set for military objectives); thus reducing the risk of inaccuracy by being protective by default. This approach by the law aligns with IHL's ultimate goal, being to provide the greatest level of protection to civilians while cognizant of military necessity and the realities of war: Doubt will compel commanders to direct their decision-making in the way that is most protective of civilians and civilian objects.

This protective approach in applying the rule of distinction is aligned with the rule of precautions in attack, which must be applied not only during the conduct of the attack, as discussed in the following section, but also at the planning stages, including in accuracy and risk assessments for military targeting operations. In fact, and also as discussed throughout this thesis, the rule of precautions to protect civilian populations and objects against the effects of attack is such that these measures must be done 'to the maximum extent feasible'.[628] This leads to the subsequent question of what states ought to do in this regard: inaccuracies cannot be completely avoided, especially because of the messy nature of war, even though the law provides a protective framework that should, in principle, reduce at the very least the risk of such inaccuracies from happening. In other words, there is the question of standard of care states must meet in order to demonstrate how far their assessments must go – especially so as to strike the right balance between the need to protect civilians and civilian objects on the one hand, and military necessity considerations on the other hand. A number of States have sought to clarify their approach as to how they should grapple with the issue of doubt and certainty in

[628] API, art 58

the light of IHL requirements. For example, the United States' DoD Manual's approach rests assessments on the basis of 'good faith based on the information available to them in light of the circumstances ruling at the time'.[629] The United Kingdom and France had expressed reservations or at least, provided nuances in relation to their application of the rule set by Article 50(1) of Additional Protocol (i.e., presumption of civilian status in case of doubt), specifically stating that such provision only applies in situations of 'substantial doubt', and the commander maintains a 'duty to protect the safety of troops under his command or to preserve his military situation.'[630] In academic scholarship, there have been attempts too at interpreting the acceptable level of doubt under IHL. For instance, the framing of 'deontological targeting' has been proposed, requiring a 'reasonable belief threshold' where the intentional killing of an individual target should be prohibited if the fighter believes the individual is a civilian, 'or if the soldier does not reasonably believe the individual is a combatant.'[631]

While there is no specific, universally agreed provision on how States should ensure accuracy in target identification and in subsequent legal assessments, one possible approach is to look at the feasibility dimension in adopting and implementing precautionary measures prior to, and in the conduct of the attack. Generally, the luxury of time, or the lack thereof, will play a decisive role in determining the standard of care expected from the attacking party in not only identifying the target, but also in assessing it and, subsequently, planning the attack.[632] The factor of time and, more widely, the circumstances surrounding the process will indeed bear great influence over compliance with, and the application of a number of specific and relevant

[629] *Law of War Manual* (n 65) Section 5.4.3.2. as cited in Michael Schmitt, Schauss (n 611)
[630] Haque (n 610) and Julia Gaudreau 'The reservations to the Protocols additional to the Geneva Conventions for the protection of war victims' (2003) 849 International Review of the Red Cross <https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/irrc_849_gaudreau-eng.pdf> accessed 19 November 2024
[631] Haque (n 610)
[632] Michael N Schmitt 'Precision attack and international humanitarian law' (2005) 87(859) International Review of the Red Cross 451 <https://international-review.icrc.org/sites/default/files/irrc_859_3.pdf> accessed 20 November 2023

IHL rules. For example, with regards to the rule of distinction, should the circumstances ruling at the time be such that there is doubt surrounding the legal status of the target (i.e., whether they would be targetable or not under the rule of distinction), and should there be no time to shed light and ascertain their status, the assumption that the target is of civilian nature ought to prevail.

This approach shares the same, fundamental assertion with the others: doubt does not necessarily equate to prohibition. This resonates with the probabilistic nature of AI and their inherent degree of uncertainty. As discussed throughout this chapter, these technologies should not be treated, nor regarded as oracles; hence expecting a 100% certainty would neither be realistic, nor helpful. As such, if, for example, a classifier system identifies a person as a potentially targetable combatant with an 80% confidence rate, this implies a 20% uncertainty with risks of targeting a civilian or a protected person. In theory, in case of doubt, IHL's protective stance is such that the person remains protected until proven otherwise. However, not only would it be non-realistic to expect the model to operate at 100% accuracy; 'proving otherwise' will still maintain an inherent degree of uncertainty. The principle of military necessity adds further layers of complexity: if the armed forces would only be able to engage a target if there is a 100% certainty over its legal status, this would be impractical and could in fact jeopardize military necessity. This issue fundamentally brings to the fore the tensions between military necessity on the one hand, and the protections conferred by IHL on the other hand.

While there is no one definite answer to address these tensions, one approach would be to look at the good faith of those setting the 'acceptable threshold' of acceptable uncertainty as well as the circumstances ruling at the time. In fact, it has been argued that honest mistakes in implementing the rule of distinction could be acceptable in the absence of an intention to

specifically and wilfully attack civilians.[633] This approach could be linked to the 'Rendulic rule', oftentimes cited to address military necessity as a basis for justifying honest mistakes.[634] The Rendulic rule specifically looks at the situation 'as it appeared to the defendant at the time' and what the commander saw as military necessity, an approach that has reportedly been adopted and incorporated into States' rules towards implementing the legal obligations surrounding the conduct of hostilities.[635] At the same time, there is a risk that a commander would try to stretch military necessity 'with a view to trumping more and more positive rules'.[636] IHL application is, ultimately, highly contextual, which then creates challenges to defining an 'acceptable threshold' that would potentially be usable both for the use but also development and testing of military AI. What may be 80% acceptable in one setting may not be so in another. An 80% confidence rate of a person's targetability in trenches may be less controversial than in a densely populated area. This acceptable threshold must thus be treated as a guidance to the commander, who would then need, as part of their responsibility, to assess the context at hand against a number of considerations, including proportionality analyses, varied sources of intelligence, as well as the rules of engagement applicable to the situation in question.

[633] Magdalena Pacholska 'Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective' (2023) 56(1) Israel Law Review 12-17 <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/israel-law-review/article/military-artificial-intelligence-and-the-principle-of-distinction-a-state-responsibility-perspective/972D0662B1207C14656D5128B7B139FA?utm_campaign=shareaholic&utm_medium=copy_link&utm_source=bookmark> accessed 19 October 2024

[634] Nobuo Hayashi 'Honest Errors, the Rendulic Rule, and Modern Combat Decision-Making' (*Articles of War*, 24 October 2023) <https://lieber.westpoint.edu/honest-errors-rendulic-rule-modern-combat-decision-making/> accessed 19 October 2024

[635] Sean Watts 'The Genesis and Significance of the Law of War "Rendulic Rule"' in Nobuo Hayashi, Carola Lingaas (eds) *Honest Errors? Combat Decision-Making 75 Years After the Hostage Case* (2024, Springer) 155-176

[636] Elliot Winter 'Pillars not Principles: The Status of Humanity and Military Necessity in the Law of Armed Conflict' (2020) 25(1) Journal of Conflict and Security Law <https://academic.oup.com/jcsl/article/25/1/1/5722185?login=true> accessed 19 October 2024

2.2.2 Accuracy in target engagement

In addition to accuracy in the identification of the target, its second dimension pertains to accuracy and precision in engaging the target in and of itself. This includes the ability of the means and method of warfare chosen to be accurate in engaging the target, both anticipated and in dynamic targeting settings. In fact, against the aforementioned considerations, discussed in the previous section, with regards to what would be expected in terms of accuracy for the identification of target, there are also a number of legal considerations with regards to the conduct of operations and decisions eventually leading to target engagement. In situations where the target has been identified prior to the conduct of operations, whether it is then expected that the target will be engaged in an accurate way is at stake.

One approach to this question pertains to the choice of means and methods of warfare to engage the target. In line with the customary IHL rule that such choice is, in fact, not unlimited, the attacking party must ensure that, in selecting the means and methods of warfare for engaging the target, the choice is such that it minimises incidental loss of civilian life, injury to civilians, as well as damage to civilian objects. Beyond the choice of accurate weapons systems, this also entails choices and decisions made on the systems surrounding these, including 'robust command, control, communications, computers, intelligence, surveillance, and reconnaissance (C4ISR)'.[637] A number of case studies have indeed shown examples where, even though the weapons employed would indeed generally be considered as accurate in terms of delivery, flaws in the systems and processes surrounding it led to engaging the wrong targets and thus, what many would consider as 'accidental bombing'.[638]

This has, for example, been the case of the bombings of the Médecins Sans Frontières hospital in Kunduz, Afghanistan by the US on 3 October 2015. Many of the facts and thus, the

[637] Schmitt (n 632)
[638] Ibid

subsequent (il)legality of the attack is still subject to much heated debates and contentions to this day; however, essentially, what is certain is that the United States' armed forces were found to have launched an attack by air against a hospital operated by Médecins Sans Frontières. The US military Central Command eventually released a partially redacted report of its investigation of the incident, where they argued that both the ground force and the aircrew were 'unaware the aircrew was firing on a medical facility throughout the engagement' and that this 'tragic incident was caused by a combination of human errors, compounded by process and equipment failures.'[639] These failures did not necessarily pertain to the accuracy of weapons systems embedded in the AC-130 gunship employed for the attack; but rather, at least from a technological perspective, the issues stemmed from failures with regards to the aircraft's communications systems, including its satellite radio.[640] As a result, and in addition to human error, the United States reportedly mistook the Médecins Sans Frontières (MSF) hospital for a prison overrun by the Taliban, and the attack led to the reported deaths of over 30 civilians and over 30 injured.[641] Regardless of where one stands in the facts and subsequent legal conclusions, it is clear, through this example, that accuracy in target engagement would indeed extend beyond the weapon system and/or any other delivery method to engage the identified target.

[639] Central Command (US) 'Summary of the Airstrike on the MSF Trauma Center in Kunduz, Afghanistan, on October 3, 2015; Investigation and Follow-on Actions' (2016) USCENTCOM FOIA 16-0060, 1 <https://www3.centcom.mil/FOIALibrary/cases/16-0060/00.%20CENTCOM%20Summary%20Memo.pdf> accessed 20 November 2023

[640] Jim Garamone 'Centcom Commander: Communications Breakdowns, Human Errors Led to Attack on Afghan Hospital' (*DoD News*, 29 April 2016) <https://www.defense.gov/News/News-Stories/Article/Article/746393/centcom-commander-communications-breakdowns-human-errors-led-to-attack-on-afgha/> accessed 20 November 2023

[641] Peter Marguiles 'Centcom Report on Kunduz Hospital Attack: Accounting for a Tragedy of Errors' (*Lawfare*, 2 May 2016) <https://www.lawfaremedia.org/article/centcom-report-kunduz-hospital-attack-accounting-tragedy-errors> accessed 20 November 2023 and 'Kunduz bombing: US attacked MSF clinic 'in error'' (*BBC News*, 25 November 2015) <https://www.bbc.co.uk/news/world-asia-34925237> accessed 20 November 2023

Beyond ensuring that these means and methods of warfare are not unlawful by nature and/or all use would not be unlawful,[642] their choice and their use must be in compliance with a number of legal considerations. This includes the rule of distinction, proportionality, and precautions in attack, in line with the many discussions held earlier throughout this thesis.

- The rule of distinction applies both in the context of target identification, as well as target engagement. In addition to providing special protection for certain objects such as medical units and transports[643], cultural property, as well as the natural environment[644], which the attacking party must respect at all times, certain groups of individuals also enjoy protection, from attack, under IHL, provided that they are not found to be directly participating in hostilities such as journalists[645], medical personnel[646], as well as prisoners of war, for whom the Third Geneva Convention provides a whole set of rules and protection. It is here therefore expected that the attacking party will exercise accuracy, in engaging targets, in such way that also respects the protection provided to these objects and categories of persons.
- The rule of proportionality, while its application is, as discussed earlier, heavily based on assessments *ex ante*, also provides for the conduct of the attack in itself – notably

[642] For example, the weapon chosen would be able to comply with the rule of distinction by drawing the 'distinction between the civilian population and combatants, or between civilian objects and military objectives, and their effects […] could […] be restricted, either in time or in space, to lawful military targets.' See *Legality of the Threat or Use of Nuclear Weapons* (n 69) [92]

[643] Emanuela-Chiara Gillard 'Seventy Years of the Geneva Conventions: What of the Future?' (2020) Chatham House 2 <https://www.chathamhouse.org/2020/03/seventy-years-geneva-conventions/protection-medical-care-armed-conflict> accessed 20 November 2023

[644] Helen Obregón Gieseken and Vanessa Murphy 'The protection of the natural environment under international humanitarian law: The ICRC's 2020 Guidelines' (2023) 105(924) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/abs/protection-of-the-natural-environment-under-international-humanitarian-law-the-icrcs-2020-guidelines/98B58E81D29737ABB853A350CC50D0C1#:~:text=By%20virtue%20of%20its%20civilian,war%20wrought%20on%20the%20environment> accessed 20 November 2023

[645] Elizabeth Levin 'Journalists as a Protected Category: A New Status for the Media in International Humanitarian Law' (2013) 17 UCLA Journal of International Law and Foreign Affairs <https://www.jstor.org/stable/45302375> accessed 20 November 2023

[646] M Goniewicz, K Goniewicz 'Protection of medical personnel in armed conflicts – case study: Afghanistan' (2013) 39(2) Eur J Trauma Emerg Surg <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3611028/> accessed 20 November 2023

the obligation to cancel or suspend the attack 'if it becomes apparent' that the attack would result in excessive civilian casualties.[647] This therefore extends the proportionality assessment beyond the planning stage and imposes a constant re-evaluation of accuracy and risk assessments even after the chosen means and methods of warfare have been deployed for target engagement.

- The rule of precautions in attack complements the latter, with parties mandated to take the 'necessary precautions to protect the civilian population, individual civilians and civilian objects under their control against the dangers resulting from military operations.'[648] Generally deemed as of customary nature, this rule extends both to the planning stage, as discussed earlier in this thesis, as well as to the actual conduct of the military operation in question 'to the maximum extent feasible'.[649] While the latter puts a certain burden on the attacking party to endeavour, as much as possible, to adopt and implement precautionary measures to spare civilians and civilian objects, it also mandates the defending party to adopt such precautionary measures – in the absence of which, the attacking party may not be prevented from conducting the attack, provided it complies with other applicable rules.[650] In other words, accuracy with regards to target engagement – and with regards to sparing civilian objects and populations – is expected from the attacker who is, however, not the sole party bound with precautionary duties.

Echoing the point on the feasibility of precautionary measures against the effects of an attack, and similar to the previous section, one question that emerges pertains to expectations for

[647] API art 57(2)(b), see also Gillard (n 369)
[648] API, art 58(c)
[649] API, art 58
[650] 'Rule 22: Principle of Precautions against the Effects of Attacks Rule 22. The parties to the conflict must take all feasible precautions to protect the civilian population and civilian objects under their control against the effects of attacks' (*ICRC IHL Database*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule22> accessed 20 November 2023

accuracy – this time, in target engagement – with regards to compliance with the other rules. One particular concept had emerged around the 1990s and early 2000s, particularly in US scholarship and doctrine, but also in the UK: 'effects-based operations.'[651] This approach essentially consists of identifying, analysing, and engaging with the target in such way that is most effective to yield the specific, desired effect consistent with the commander's objectives.[652] This, for example, would consist of disabling communication transmissions from a media facility misused to direct military operations: the desired effect being to neutralise all forms of offending communication and transmission from the facility, the use of carbon filaments on power lines would be more effective – and carry less risks of civilian harm – with regards to the desired objective without carrying a bombardment of the entire media facility.[653] In this sense, accuracy would extend beyond simply engaging with the target: the way it is being engaged with and neutralised is such that it is the most effective for the desired objective, even if it implies a re-evaluation and re-definition of the identified target.

Yet, a number of observations and limitations of such approach can be identified. For example, on the one hand, while the desired effect can be defined such that IHL is a central component to it; conversely, it does not necessarily guarantee that the commander would indeed put the law at the heart of the desired effect, at the risk of putting military victory and hawkish ambitions at the expense of legal compliance. Second, it has been argued that such approach is, simply, a 'fancy acronym for what most people would simply call rational behavior', notably

[651] This concept was indeed particularly prominent in the UK's 2004 Defence White Paper: House of Commons (UK) Library 'The Defence White Paper' (2004) Research Paper 04/71 23 <https://researchbriefings.files.parliament.uk/documents/RP04-71/RP04-71.pdf> accessed 20 November 2023. It also featured in the Air Force (US) Air Force Basic Doctrine Document 1 (2003) 18 <https://www.bits.de/NRANEU/others/END-Archive/USAF-afdd1(2003).pdf> accessed 20 November 2023, as well as a number of NATO documents, see Liselotte Odgaard (ed) *Strategy in NATO* (Palgrave Studies in Governance, Security and Development, 2014) 179 <https://link.springer.com/book/10.1057/9781137382054>accessed 20 November 2023

[652] T. W. Beagle Jr. 'Effects-Based Targeting: Another Empty Promise?' (2001) Air University Press <https://www.jstor.org/stable/pdf/resrep13830.8.pdf> accessed 20 November 2023

[653] Schmitt (n 632)

in the light of the supposedly comprehensive nature of effects-based operations.[654] Regardless of where one stands with regards to the overall authority of this approach, one thing that can be drawn from this with regards to expectations for accuracy is that such assessments must be done holistically, taking into account – to the maximum extent feasible, as the rule of precautions stipulates – the many components and dimensions related to the identified target against the desired effect. In other words, accuracy lies in achieving the desired effect – which, for us, would particularly correspond to legal compliance.

3 Inherent Uncertainties and Expectations for Accuracy in AI Systems: What does the law say?

In the light of how, as discussed in the previous section, the law addresses issues with regards to uncertainties and expectations for accuracy in the context of military targeting, the present section will specifically look at how these considerations apply in the light of AI development to assist with such operations. In fact, as states and companies in the private sector invest heavily in developing AI-enabled capabilities to facilitate the conduct of military operations, the desire for such solutions to foster compliance with international law plays a critical role. This is due to a number of reasons including, but not limited to, chances that AI capabilities will enable greater precision and accuracy in the conduct of military targeting, both in terms of target identification and target engagement.[655] Whether or not these aspirations are, indeed, realistic and promising or, on the contrary, source of risks of IHL violation is obviously dependent on a number of factors, including, among others, the technology's overall reliability, but also the conditions under which it has been developed and tested, as well as the circumstances under which it is being deployed. The present section will dissect, in detail, a

[654] Odgaard (n 651)
[655] Marguiles (n 619)

select number of legal aspects relevant to military targeting and how they apply with regards to the development of AI technologies for military targeting.

3.1 AI-Enabled Technologies: A Compliance Enabler?

As discussed in the previous section, there are a number of ways the rule of distinction applies in the context of military targeting: from the identification of target in the planning stages to target engagement, the rule of distinction shapes many of the commander's decisions as it delineates who is targetable and under which circumstances – including in case of doubt. Yet, in the light of both the messy and dynamic nature of warfare and thus, its inherent uncertainties, as well as recent trends and notably with the increasing urbanisation of warfare and the prevalence of non-international armed conflicts, it is proving, at times, to be difficult to ensure accuracy and thus, compliance with the rule of distinction. The extent to which parties are capable of collecting accurate intelligence for target identification – and subsequently apply it for target engagement – can be met with challenges.

As discussed in the previous section, even if the intelligence available holds the correct information on what is targetable and, conversely, non-targetable, a combination of human error and technological issues (e.g., with the communication system) may lead to the engagement of the wrong target. The MSF Kunduz hospital's GPS coordinates were indeed recorded in the US' no-target list, however the AC-130 attacking aircraft was not able to access the list; and while the aircrew transmitted the hospital's exact coordinates to the Bagram Airbase before launching the attack, the latter did not check its status despite having access to the no-target list.[656]

[656] Françoise Bouchet-Saulnier and Jonathan Whittall 'An environment conducive to mistakes? Lessons learnt from the attack on the Médecins Sans Frontières hospital in Kunduz, Afghanistan' (2018) 100(907-909) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/an-environment-conducive-to-mistakes-lessons-learnt-from-the-attack-on-the-medecins-sans-frontieres-hospital-in-kunduz-afghanistan/53B753AD4843E442AE69EB87A94458F7> accessed 21 November 2023

It is precisely these kinds of human errors and technological limitations that drive the desire by many, primarily states but also the private sector and a handful of academic researchers, to develop AI-enabled solutions to foster accuracy in target identification and engagement and thus, compliance with international law.

As discussed throughout this thesis, AI-enabled technologies hold the potential, if working as intended, to facilitate, or even enable, the conduct of military operations in compliance with the law. One of the key appeals of such technologies is the potential it holds to allow for greater situational awareness and thus, subsequently, greater accuracy in target identification while reducing risks of civilian loss and damage. Such technologies would indeed not only enable the collection of large volumes of data live from the battlefield; they also hold the potential to simulate attacks and conduct subsequent risk assessments and decision-making based on the programme's outputs.[657] Many see the promise held by AI technologies as capable of clearing 'through the fog of war' by enabling greater situational awareness and thus, not only increasing mission success rates through greater accuracy, but also, eventually, greater compliance with the law.[658]

AI-enabled capabilities also hold the potential to foster accuracy and, subsequently, compliance with the law by enabling precision strikes through the weapon system and other means of delivery, if not within the munition in itself. In fact, beyond AI-enabled technologies, there have been a number of research and development endeavours for the military to develop, and

[657] Marguiles (n 619)

[658] Defence companies tend to use this clearing of the fog of war as a selling point for their AI capabilities, see for example: 'New Advanced Technologies Improve Situational Awareness for Land Forces' (*THALES*, 12 September 2022) <https://www.thalesgroup.com/en/united-states/magazine/new-advanced-technologies-improve-situational-awareness-land-forces> accessed 21 November 2023. See also from the US army: 'Machine learning algorithms promise better situational awareness' (*US Army*, 22 June 2020) <https://www.army.mil/article/236647/machine_learning_algorithms_promise_better_situational_awareness> accessed 21 November 2023, as well as Ling He and others 'Artificial intelligence technology in battlefield situation awareness' (VESEP22, 2022) <https://www.e3s-conferences.org/articles/e3sconf/pdf/2022/27/e3sconf_vesep2022_01063.pdf> accessed 21 November 2023, and Nurkin and Siegel (n 201)

eventually deploy, such capabilities – for example with precision-guided munitions.[659] Today, states and the defence industry are looking into integrating AI to further improving the performance and thus, the accuracy of such weapons; one of the most notable examples being the use of AI to enable swarm drones, capable of operating in a coordinated way to perform the task as intended.[660]

3.2 AI's Probabilistic Nature and Inherent Uncertainties: Is There Room for Errors?

While the previous section discussed ways in which AI could foster compliance with IHL, on the flipside of the coin, the issue of these technologies' probabilistic nature arises. As established earlier in this chapter, AI technologies are indeed bound to have a degree of uncertainty; hence a margin of error is to be expected. How this knowledge, and anticipation for error translates into compliance with the law is to be discussed in the present section.

IHL has ways of dealing with uncertainties, an inevitable characteristic of armed conflicts. This includes the protective stance the law takes in case of doubt with regards to the legal status of a target in the context of the rule of distinction. Yet, certain points remain unclear and subject to much debate. One of these is the question of accidental targeting against persons and objects that, under the rule of distinction, would normally be protected. Most of the discussions and proposed approaches to this issue pertains to verifying compliance with the rule of precautions; although this question of 'honest mistakes' that result in the attack of civilians, despite all precautions being taken beforehand, remains subject to much challenging.[661] Yet, in most if

[659] Congressional Research Service (US), 'Defense Primer: U.S. Precision-Guided Munitions' (2024) <https://crsreports.congress.gov/product/pdf/IF/IF11353> accessed 23 November 2024
[660] William H. Boothby 'Highly Automated and Autonomous Technologies' in William Boothby (ed) New Technologies and the Law in War and Peace (Cambridge University Press, 2018) 140-142 <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/product/A4911CB3BC08CC35B7D22B9439FD9C13/core-reader> accessed 21 November 2023
[661] See for example: Pacholska (n 633)

not all cases, states have indeed seemed to refrain themselves from admitting legal responsibility in cases where they argued this was an accident.[662] Even the ICTY generally rests the legality assessment on whether the attacking party took the precautionary measures needed, in the absence of which there may have been a violation of the law: the Tribunal specifically refused to prosecute for an incident where a NATO aircraft was found to have accidentally launched an attack on a convoy of Albanian refugees on the Djakovica-Prizren road, as 'neither the aircrew nor their commanders displayed the degree of recklessness in failing to take precautionary measures which would sustain criminal charges.'[663]

Following the logic of the majority, this means that the rule of precautions will bear heavy weight in addressing the inherent degrees of uncertainty and risks of inaccuracies in the context of AI development for target identification and target engagement. The question of technology development for precision targeting is actually not new: from the early 2000s, there have been discussions on using technologies to allow 'greater battlespace transparency' and thus, situational awareness, as well as to facilitate the conduct of the attack in itself (i.e., target engagement), echoing both aspects of accuracy discussed in the previous section.[664] Building on what has been discussed in the previous section, the availability of precautionary measures (and their degree) for each attack will vary from one another.[665] These depend on a number of decisive factors including, but not limited to, time constraints, the objective and overall effects desired from the attack, military doctrine, as well as the technological solutions available and used for each attack. In this sense, while there is no one-size-fits-all solution to addressing inherent degrees of uncertainties and risks of inaccuracies in AI capabilities for military

[662] Marko Milanovic 'Mistakes of Fact When Using Lethal Force in International Law: Part I' (*EJIL: Talk!*, 14 January 2020) <https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-i/> accessed 21 November 2023
[663] ICTY (n 380)
[664] Schmitt (n 632)
[665] Schmitt (n 611)

targeting, there are a number of ways through which IHL can be used to frame and shape key decisions made in the development of these technologies.

One of such solutions, as discussed in chapter 2, is to restrict target identification capabilities exclusively to the positive identification of combatants, with the assumption that all are civilian and protected until proven otherwise with a high degree of confidence.[666] In the context of the present thesis, it is impossible to provide a specific and precise percentage for this confidence rate, nor it is actually desirable. Again, States have varying risk assessments, and there are many ways through which AI technologies can be developed and deployed to assist with military targeting; the associated levels of risk for inaccuracies will subsequently greatly vary from one another. This approach primarily rests on the protective assumption of IHL in case of doubt with regards to the rule of distinction as a legal basis. Beyond target identification, with regards to developing solutions to facilitate target engagement (e.g., through AI-enabled weapons and delivery systems), another possible approach would be to make high-risk decisions subject to human oversight and decision-making, ensuring traceability and clarity on the commander's accountability. Even though the latter ought to always be upheld, additional verification and validation measures, whether it is in the programme or the interface, will ensure that risks of inaccuracies are reduced as much as possible, in line with the rule of precautions in attack. A number of studies are in fact dedicated to dissecting what the humans need to be able to do, precisely, including the ability to have 'reasonable certainty about the effects' and exercise 'judgment and intent'.[667] In the case of the NSA's Skynet, for example, where a journalist was mistaken as an Al-Qaeda courier based on patterns emitted by their phone's metadata: in situations where this identification of targets is done with plenty of time for the verification of the programme's outputs, and for such decisions involving the lethal

[666] See Chapter 2, Section 3.2.2 on the case for a legal obligation of targets' positive identification.
[667] Boulanin and others (n 558)

engagement of targets, it would in fact be expected that a human commander steps in to validate, or not, the target identified by the programme. Ultimately, clarifying not only where humans could or should intervene, from the development/pre-deployment stages, to decrease risks of inaccuracies or at least, risks of legal 'grey zones' will enable compliance by design, an approach this thesis aims to advocate for in its entirety.[668]

Hence, the issue at stake is not whether there is indeed room for mistakes with regards to the application of the law in military operations, especially with regards to AI deployment: the question is more how much room for mistake there is, in the light of the capabilities and information such technologies offer. Normal accidents are to be expected: they are not a novel problem, which the military has been grappling for years and across technological developments.[669] In this sense, while the degree of accuracy expected from each technology will, as ascertained earlier, depend on a case-by-case basis, AI solutions that have passed testing, evaluation, verification and validation examinations and legal reviews pre-deployment are expected to be used in such ways that risks of inaccuracy in targeting will, in fact, be decreased. It is then up to each deploying state to clarify, for their own legal review processes, the level of risks of inaccuracy they are willing to accept and how they interpret these risks against the applicable laws.[670] This, of course, does not mean that States get a 'free pass' to decide for themselves how accurate their systems ought to be. As discussed earlier, a static acceptable threshold would actually not be desirable in the light of the highly contextual nature

[668] Berenice Boutin, Taylor Woodcock 'Aspects of Realizing (Meaningful) Human Control: A Legal Perspective' in Robin Geiß and Henning Lahmann (eds) *Research Handbook on Warfare and Artificial Intelligence* (2024, Edward Elgar Publishing) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4109202> accessed 21 November 2023

[669] Paul Scharre 'Autonomous Weapons and Operational Risks' (2016) CNAS <https://www.stopkillerrobots.org/wp-content/uploads/2021/09/CNAS_Autonomous-weapons-operational-risk.pdf> accessed 21 November 2023

[670] A number of studies seek to identify key considerations to feed into states' legal review processes, see for example: Dustin A. Lewis 'Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider' (*Humanitarian Law & Policy*, 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 21 November 2023

of IHL interpretation.[671] Furthermore, States ought to remember, in the interpretation of the law, that IHL's essence is not about setting the minimum levels of permissibility to justify destructive acts. Rather, legal interpretation should be done in the spirit of fostering the protection of civilians – the underlying and very purpose of why IHL came to be in the first place.[672]

3.3 The Black Box Issue, Forensics, and Evidence for Investigations of Alleged Violations: An Expectation for Clarity?

One key question that arises from the black box issue, its opacity and the inherent uncertainties that come from it, pertains to the obligation to conduct investigations into alleged violations of the law by the parties to an armed conflict. Tied to this question is whether, more generally, the exercise and implementation of certain rules set in international humanitarian law would require the clarity black box models would not necessarily offer. In other words – when it comes to applying the law and eventually investigating the facts for accountability purposes, does the law require, in certain situations, documented traceability in the commander's decision-making process? And if so, under which circumstances does the law set such requirements?

A number of states and legal sources, both statutory and customary, have recognised the value and the need for investigations of alleged violations of international humanitarian law as 'critical for the proper application' of the law both in international and non-international armed conflict settings.[673] While there is no universally agreed norms on what and how such investigations ought to be undertaken, scholars have specifically researched, in detail, not only

[671] See Section 2.2.1
[672] Droege (n 560)
[673] Noam Lubell, Jelena Pejic, Claire Simmons 'Guidelines on Investigating Violations of International Humanitarian Law: Law, Policy, and Good Practice' (2019) ICRC and the Geneva Academy of International Humanitarian Law and Human Rights <https://www.icrc.org/en/document/guidelines-investigating-violations-ihl-law-policy-and-good-practice> accessed 24 February 2023

the legal foundation justifying such requirement to undertake investigations; but also key considerations and principles that must be taken into account while conducting investigations; as well as the legal and practical challenges that arise from the conduct of investigations of alleged violations of the law in armed conflict.[674] Generally, such investigations will need to be 'effective'. While there is no general consensus on a set definition of what 'effective' means, one proposed approach was that they must be '*capable* of enabling a determination of whether there was a violation of international humanitarian law, or identifying the individual and systemic factors that caused or contributed to an incident, and of laying the ground for any remedial action that may be required'.[675]

Drawing from the proposed approach above on what an effective investigation would look like, the following observations and reflections ought to be laid out. First, the authors of the abovementioned approach put a specific emphasis on the capacity to enable a determination of whether there was, indeed, a violation of IHL. To note that it is not about the capacity to determine, but the capacity to enable a determination; hence the key here is to see the force of what enables such determination – as in, not only the procedures in place for the conduct of the investigation (e.g., processes in place to trigger an investigation; the designation of an independent and impartial investigative authority; processes in place for the assessment; etc.). It is also about the forensics, for example the available documentation from the pre-planning and conduct of operations (and what enables their collection and maintenance in archives); as

[674] Ibid; see also: M. Tidball-Binz and others 'A good practice guide for the use of forensic genetics applied to human rights and international humanitarian law investigations' (2013) 4(1) Forensic Science International: Genetics Supplement Series <https://www.sciencedirect.com/science/article/abs/pii/S1875176813001108> accessed 25 February 2023; Andrea Joy Harrison 'Guidelines on Investigating Violations of International Humanitarian Law and Practical Mechanisms for Accountability - Today and Beyond' (2021) 229(2) Military Law Review <https://heinonline.org/HOL/LandingPage?handle=hein.journals/milrv229&div=16&id=&page=> accessed 25 February 2023; and for a more recent example of scholarly research in the area of investigations in armed conflict, see: Claire Simmons 'Investigations in armed conflict' in Robert Kolb, Gloria Gaggioli and Pavle Kilibarda (eds) *Research Handbook on Human Rights and Humanitarian Law* (Elgar, 2022) <https://www.elgaronline.com/display/book/9781789900972/book-part-9781789900972-29.xml> accessed 25 February 2023

[675] Lubell, Pejic, Simmons (n 673)

well as the collection and analysis of evidence – as feasible as possible in armed conflict situations.[676] Second, violation of IHL is laid out as not only caused (or contributed to) by individual choices, but also 'systemic issues, which the authors of the discussed approach carefully defined as 'if the underlying causes of an incident are likely to have led to or could lead to further incidents.'[677] Another observation to note here is that it is not simply about factors that are likely to have contributed to IHL violations, but also those that could lead to more incidents in the future. In this sense, once a factor has been identified as an underlying cause to an incident, there is a call for assessments as to whether their continued adoption, deployment or use carries the risk of leading to further violations.

How this approach would be applied and operationalized in practice in the light of the black box nature of certain AI-enabled technologies remains a question mark subject of concern shared by many – states and civil society alike.[678] First, on the capacity to enable a determination of whether there was, indeed, a violation of IHL. One could easily make a claim that because of the opacity induced by black box models, this very opacity obstructs this capacity to enable the determination of an alleged violation of the law. However, this claim would merit further reflections – as in, does international humanitarian law actually require a degree of transparency and explainability over (certain) decisions? As discussed earlier in this

[676] The ICRC has, for example, a forensics unit to enable for the search, recovery, examination and identification of the deceased from the battlefield. See: 'How does ICRC use forensic science for humanitarian purposes?' (*ICRC*, 19 May 2022) <https://www.icrc.org/en/document/how-does-icrc-use-forensic-science-humanitarian-purposes> accessed 26 February 2023. Information recovered from such forensic analyses, whether by local authorities or from third, neutral parties such as the ICRC, could arguably be used as part of IHL implementation – for example, see: Anjli Parrin '"How did they die?": Bridging humanitarian and criminal-justice objectives in forensic science to advance the rights of families of the missing under international humanitarian law' (2023) 105(923) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/how-did-they-die-bridging-humanitarian-and-criminaljustice-objectives-in-forensic-science-to-advance-the-rights-of-families-of-the-missing-under-international-humanitarian-law/8A3A55040D256FEA79387512E25F9751> accessed 26 February 2023

[677] Lubell, Pejic, Simmons (n 673)

[678] What black box and the subsequent lack of transparency means for investigations and accountability is a point raised by many participants, both during panel discussions and shared with the author on the occasion of private discussions taken place at the Responsible AI in the Military domain summit, held on 15-16 February 2023 in The Hague, Netherlands

thesis, the black box is in some (if not most) cases inevitable. Some would even claim that there is a certain trade-off between transparency and the sophistication and accuracy of AI: in other words, the more complex the problem to solve is, the more opaque the programme's calculations will need to be. It is not the scope of this thesis to discuss the technical intricacies of AI and deliberate on the accuracy of such claims. Yet, it is important that we do spend some time on how black box models will affect legal assessments and, subsequently, the ability to conduct effective investigations on alleged violations of the law.

For example, let us take the rule of distinction. Beneath its simple, binary surface lies layers and nuances of discussions as to how to assess someone's, or an object's status (i.e., targetable, or non-targetable). For individuals, the determination of a civilian as directly participating in hostilities (i.e., civilians foregoing their protection from direct attacks due to the conduct of direct participation in hostilities) is one particularly contentious subject; while for objects, the determination of an object's status with dual-use functions, both civilian and for military, is also the subject of heated debates. In both cases, the interpretation and application of the law has laid out a number of specific conditions for the individual, or the object in question, to be considered as lawfully targetable.

Regarding the direct participation in hostilities, while the law does not specifically define what conduct and/or actions would constitute such 'direct participation', the ICRC has published its own interpretative guidance on the notion and laid out a number of cumulative, constitutive criteria that would indicate that someone was indeed directly participating in hostilities, specifically on: 1) the effect of the act; 2) the direct causal link between the act and the harm; and 3) the belligerent nexus – i.e., the intent to cause the required threshold of harm to support a party to the conflict and to the detriment of another.[679] The attacker must be able to justify

[679] Melzer (n 336)

that their target's actions met, cumulatively, these three requirements in order to render their attack lawful; this threshold being arguably high in the light of the protective presumption prescribed by the law in case of doubt. In addition, this assessment also requires a degree of qualitative appreciation that must be justified by the attacker. For example, the belligerent nexus requires an element of intent from the civilian in question to support one specific belligerent party and to cause harm to the other. The same goes for assessments of a military objective in the context of dual-use objects. The only difference is that the latter is a bit more set in stone than the former, in the light of API's definition of a military objective somewhat providing some direction on how to address dual-use objects: the objects must, 'by their nature, location, purpose or use make an effective contribution to military action and whose partial or total destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.'[680] The attacker must also, in this case, be able to justify that their target's actions met, cumulatively, these requirements in order to render their attack lawful; yet again with the general assumption that objects are protected in case of doubt. And, again, this assessment presents a number of qualitative judgments that must be made by the attacking party – for example on 'the circumstances ruling at the time' and how an attack on the object would offer a 'definite' military advantage.

In the light of their qualitative nature, the reliance on black box models to undertake these assessments would make it somewhat difficult to justify the attack. In investigative contexts, one could easily argue that in situations where a contentious attack was launched because of a black box model's calculations and results (thus, there is no way in knowing why the system decided to launch an attack specifically on the target – and how the system 'justifies' their targetable status), this opacity stands in the way of conducting 'effective' investigations. This issue makes the case for our anticipatory approach in enabling compliance with international

[680] API, art 52(2)

humanitarian law from the earlier stages of AI-enabled technologies for military targeting. In other words, by addressing this question from 'upstream' on how to conciliate between black box models and the need for qualitative assessments to justify an attack on certain types of targets, this will not only decrease risks of violations of the law, but also risks of the ineffective conduct of investigations further down the line in case of alleged violations.

Coming back to our discussions on compliance with the rule of distinction. The deployment and use of AI to identify civilians directly participating in hostilities is to be expected in the foreseeable future. As mentioned in previous chapters, for example, DARPA has a dedicated project specifically on developing autonomous systems to assist US ground forces in urban settings with detecting hostile forces and establish their positive identification before the troops come in contact with them.[681] As stated by DARPA, the final decisions whether to escalate or reduce response to a perceived threat would remain in the hands of humans; the programme merely provides 'additional intelligence' to inform troops prior to direct engagement.[682] Two things to note: First, given that the programme is designed for deployment and use in urban settings and thus, in environments with high concentrations of civilians, it is very likely that it will also be used to identify civilians directly participating in hostilities. Second, DARPA presents the programme as combining knowledge about 'human behaviors, autonomy algorithms, integrated sensors, multiple sensor modalities, and measurable human responses' in order to differentiate 'between hostile and innocent people' – or, in other words, lawful targets and protected civilians.[683] The operationalization of such complex parameters will inevitably imply the use of black box models. What the programme comes up with in terms of

[681] Russell (n 321)
[682] Ibid
[683] Ibid

final results is one thing – how it gets to the final decision and the intermediate decisions in-between is another.[684]

Say, in case of alleged violations of the law stemming from the results showcased by DARPA's programme, because it led troops to directly target civilians who was erroneously considered as directly participating in hostilities: how would investigations be conducted in relation to the opacity in the system's functioning as to how it got to its final, yet fatal decision? Echoing our discussions earlier, the mere fact that a black box model contributes to (or is the main cause of) such incident should not automatically absolve commanders and their subordinates of any responsibility, and there are ways to circumvent these challenges. Both assessments to establish direct participation in hostilities, and the establishment of a dual-use object as military objective, require the meeting of cumulative criteria: thus, some form of evidence is needed to allow the commander to justify their assessment and their decision to engage the target, including if their decision-making was heavily informed by a black box model. This can be done, for example, by hardwiring some form of explainability 'by design' in such high-risk and high-impact programmes – where the interface allows users to consult (and subsequently re-visit at a later time) the key factors the programme relied on for its final decision. In the case of civilians directly participating in hostilities, users could see the reasons why the programme classified the individual target as such – for example, because the programme detected the individual as bearing arms; they showcased hostility; they were wearing symbols/clothing that suggest allegiance or support to the adversary; etc. Even if there would be a black box in terms of, for example, how the programme ascertained that the individual in question was indeed bearing arms (e.g., Did it recognize a specific weapon? Did the individual carry an object in a certain way that indicated that they were bearing arms?), one way to not make this too much

[684] Paul J. Blazek 'Why We Will Never Open Deep Learning's Black Box' (*Towards Data Science*, 2 March 2022) <https://towardsdatascience.com/why-we-will-never-open-deep-learnings-black-box-4c27cd335118> accessed 26 February 2023

of an issue (for investigatory purposes, at least) would be to couple these assessments with documented evidence – for example, video recordings from the scene of the incident, which will allow investigators to assess the veracity of the system's results. As such, explainability – and the support of other forms of evidence – are key to circumvent issues surrounding black box models for qualitative assessments in military targeting.

The same approach applies for compliance with the rule of proportionality; as in, the assessment as to whether the launch of the attack, which may be expected to cause incidental loss of civilian life/injury/damage to civilian objects, would be excessive in relation to the concrete and direct military advantage anticipated. Here again, there is an issue of qualitative elements that need to be factored in. What would be key is to ensure that, from the design and training stages of the programme, it has been hardwired to provide some form of explanation on the key factors contributing it to calculating the anticipated military advantage and, on the flipside of the coin, how 'excessiveness' is calculated. In addition to, of course, ensuring that the training of these systems has been well guided by clear data on key indicators, from the law, on what would constitute 'excessive' – and that there is accessible documentation on what the training and testing data were, further reinforcing the arguments we made earlier in chapter 2.

One additional consideration pertains to the question of opacity on how the system counterbalances considerations for mission success with limits imposed by the law. It may be that the system has been hardwired to demonstrate key metrics it has used to determine that an object constituted indeed, at the time of the attack, a military objective. However, there may be limitations as to how detailed and 'transparent' can it be with regards to making a decision against competing considerations. If we take, for example, the 1999 NATO Operation Allied Force. It was subject to much scrutiny and criticism by Amnesty International, notably when it comes to NATO's high-altitude bombing, as this tactic would supposedly decrease accuracy

and subsequently increase risks of harm to civilians.[685] However, it has been argued that NATO's choice was justified in order to stay outside the reach of the Yugoslavian air defences' systems and lower flight altitudes may actually decrease accuracy under certain circumstances.[686]

If a similar scenario took place and the choice of flight altitudes is determined by a battle management software (with a commander's supervision), the use of such software may be a good way of calculating and predicting the precision and impact of strikes at a certain place using specific kinds of missiles while ensuring mission success, for example. In other words, the use of such software would be beneficial for the purpose of fulfilling military needs. As long as it meets the needs for mission success (from a strictly military necessity perspective), whether the software uses black box models would not change much. Where there may be contention is if there is no clarity as to how the software sought to counterbalance its calculations on military necessity with humanitarian considerations; hence there are risks that the system and suggests decision-making pathways that would disproportionately cause harm on civilians and civilian objects. In case 'opening' the black box is not possible for such high-risk and high-impact decisions, one way to circumvent legal issues would be to ensure that there is a certain level of human oversight from the design of the software, on these AI-assisted decision-making processes, who could then ensure that the use of such software will not lead to the violation of IHL. This, of course, will also need to be assessed against the circumstances and intended uses, where such human oversight would not constitute a liability more than a strength). This involvement of human operators would need to be facilitated and taken into account in the design of the software's interface – for example with an accessible and user-

[685] 'Federal Republic of Yugoslavia (FRY) /NATO: "Collateral damage" or unlawful killings? Violations of the Laws of War by NATO during Operation Allied Force' (*Amnesty International*, 5 June 2000) <https://www.amnesty.org/en/documents/eur70/018/2000/en/> accessed 29 April 2022

[686] Michael N Schmitt 'Military Necessity and Humanity in International Humanitarian Law: Preserving the Delicate Balance' in Michael N Schmitt, Essays on Law and War at the Fault Lines (Springer, 2012) <https://link.springer.com/chapter/10.1007/978-90-6704-740-1_3> accessed 29 April 2022

friendly interface and adequate training, in such a way that end-users will have the capacity to maintain such oversight. In most extreme cases, one could even argue that in the light of risks that are too high of IHL violation stemming from the black box nature of the system, the only solution to remain in compliance with the law would be to not deploy these systems altogether.

To conclude, there are definite ways to circumvent risks of investigators being kept in the dark by the black box. And it is somewhat arguably a legal duty to find these workarounds. In this sense, while the means may justify the ends (as in, black box models may be permissible as long as the outcome is lawful), it is also in the condition that there is clarity on the situations in which their use would be lawful; the legal boundaries are clear; and that solutions are embedded, from the development and testing phases of the algorithm, to avoid accountability loopholes and ensure legal compliance. Black box models are bound to be deployed and used: their compliance with the law 'by design' is what needs to be ensured.

Conclusion

1. Reflections on the thesis' work and objectives

In voicing their interest and commitment to developing AI solutions to assist with, or even enable military operations, it is clear that we live in an age where, unlike blinding laser weapons, advocating for the overall prohibition of their development and use is neither productive, nor realistic.[687] It is encouraging to see that, unlike the early years of the discussions surrounding military AI in international fora, including when the author first started to reflect on this thesis' scope back in 2017, the general understanding has evolved and gained in nuance: it is indeed much less about whether AI-enabled technologies ought to be prohibited or not, and there is a realisation that beyond 'lethal autonomous weapons systems', which delineate the sole deliberations existing at the multilateral level, the application of AI-enabled technologies in the military domain is much wider and more diverse.[688]

There has been, eventually, the realisation that military AI is not going anywhere: if anything, research and development in this space is proceeding apace, and levels of investment and interest are at their highest. National policies and processes – or even at the international level – seeking to frame states' approaches and strategies with regards to the development, testing, deployment and use of these capabilities are mushrooming. The US have published a number of national strategy documents in this space, while the UK has established a House of Lords Select Committee on AI in Weapon Systems.[689] Beyond national efforts, the Netherlands and the Republic of Korea have both spearheaded the organisation and convening of the inaugural

[687] Louise Doswald-Beck 'New Protocol on Blinding Laser Weapons' (1996) 30 June 1996, International Review of the Red Cross' (1996) 36(312) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross-1961-1997/article/abs/new-protocol-on-blinding-laser-weapons/B72BF4A7260BC6B957EF6786A5D79CFE> accessed 21 November 2023

[688] On the international discussions on LAWS, see: Congressional Research Service (US) 'International Discussions Concerning Lethal Autonomous Weapon Systems' (2021) <https://apps.dtic.mil/sti/pdfs/AD1171922.pdf> accessed 21 November 2023

[689] For US documents, see for example: *Responsible Artificial Intelligence Strategy and Implementation Pathway* (n 35) and US Department of State (2023), 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy' (n 514)

REAIM summit, first held in The Hague in February 2023 and the second edition of which was held in Seoul in September 2024, and which has been co-hosted by Singapore, Kenya, and the UK. In this sense, as the narrative shifts from the early landscape (i.e., either seeking the ban on such capabilities versus the justification of developing such capabilities) to one where of acceptance of the reality, that these technologies are set to play an increasingly important role in the conduct of military operations.

Building on this assumption, it is critical that extensive research is dedicated to clarifying how the law applies, both in the context of the eventual deployment and use of these technologies, but also their design and development. The pre-deployment stages of an AI technology's lifecycle plays such a decisive role in their performance and final outputs; this thesis seeks to leverage the criticality of these early stages to ensure that as these technologies are being developed, tested, and eventually deployed for use, they will be compliant 'by design'. In other words, this thesis sought, throughout its chapters, to clarify how international humanitarian law considerations can be particularly useful – and, actually, essential – in shaping and influencing the way these technologies are being designed and developed to assist with military targeting operations. The realisation for the important of such topic stemmed from a number of reflections the author has had, along with her supervisors, over the years as the debate and discussions on this topic evolved and gained in maturity. While the very initial scope of this thesis sought to examine the legal implications of algorithmic bias in AI for military targeting, the focus eventually shifted to reliability of such technologies for legal compliance in order to encompass the wider context in which discussions on algorithmic bias take place. Reflections in this space eventually led to the conclusion that, in order for reliability to be achieved, an upstream approach to legal compliance ought to be explored.

As such, while this thesis has not sought to provide definitive answers on how such reliability can be achieved, it is hoped that through its four chapters, its readers will grasp a better

understanding on the importance of identifying, and dissecting, relevant international humanitarian law considerations – and analyse ways through which they can be used to inform and shape the development of AI-enabled technologies for military targeting. Given the high stakes – often of life and death – of decisions stemming from the use of such technologies, it is imperative that end-users are capable of using these technologies with the confidence that they have indeed gone through thorough processes, from the very beginning, that will ensure that their intended use will indeed be in compliance with the law. To do so, the author looked at three, particular substantive dimensions to the development of these technologies:

1. **The importance and relevance of training datasets**:

Chapter 2 was focused on examining specific aspects to the quality and quantity of data in the development, training, testing and evaluation of AI for military targeting. This analysis was motivated by two main elements. First, data constitutes a key element to all AI technologies. Not only is data necessary for the very creation of AI technologies; it is also a critical and decisive factor for their performance and overall reliability. Data is, essentially, the lifeblood of AI systems. As such, no discussion on fostering the compliance of AI technologies 'by design' could be held without looking at data. Second, governance discussions surrounding the development of AI in the military domain tend to generally overlook the issue of data. Even more so with regards to legal scholarship. As such, this chapter seeks to fill in this research gap by providing readers with an overview of the key legal considerations surrounding select data practices and dimensions. One key takeaway is that international humanitarian law is very much relevant to the many facets of data, from the correlation between legal compliance and the diversity of data, to the legal implications of proxy data use.

2. **The design and development processes of AI-enabled programmes**:

Beyond data, Chapter 3 sought to examine other aspects related to the design and development of AI systems. Three facets were examined through the lens of international humanitarian law: the legal permissibility of machine learning use for anti-personnel targeting; the role of developers and relevant implications; as well as the need for legal reviews to keep up with ever-learning machines. It is important to recognize that, while many, if not most of the main IHL issues related to military AI are not necessarily new, these technologies still bring specific considerations that would merit further thought. Perhaps there is nothing new under the sun; but clarity may be needed on what could be the unique and novel nuances and considerations AI brings to the debate. As such, this chapter was an invitation, for the reader, to reflect on some of these aspects and how these relate to existing approaches to implementing international humanitarian law. One key takeaway from this chapter is that our assumptions on this topic ought to always be questioned and revisited. What seems like a uniquely crowded space may not necessarily be as unique as we think it is. And what may be common practice with regards to legal reviews for weapons systems may not always be fit-for-purpose with regards to the testing and evaluation of AI technologies.

3. **The probabilistic nature of AI technologies and inherent uncertainties**:

Finally, Chapter 4 sought to address the issue of unknowns and uncertainties that stem out of AI technologies. The military is indeed known for its inclination for certainty, order and predictability, perhaps out of necessity in response to war's inherently messy, complex and unpredictable nature. While AI does not necessarily solve the issues that stem out of it, correct design could help mitigate risks instead of exacerbating them. Yet, it is important to note that as AI is developed to decrease the uncertainties brought by warfare, paradoxically, these technologies are probabilistic by nature and thus bring with them an inherent degree of uncertainty. These are further exacerbated by the black box issue, or the opacity brought by the complex nature of today's models. As such, the chapter sought to examine what these

uncertainties are, and how the law approaches expectations for accuracy and reliability from the development, deployment and subsequent use of these technologies. One key takeaway is that war and AI are both inherently uncertain, and while the law may not provide a one-size-fits-all solution to addressing these uncertainties, context is in fact important and desirable in the light of the intricate nature of conflict.

2. Academic Contribution & Areas for Future Research

The primary aim of this thesis has been to contribute to the general work and efforts on the governance of AI and, ultimately, ensuring the responsible development and deployment of these technologies as they grow in prominence and influence over the conduct of military operations. By formulating the IHL considerations surrounding the development of these technologies for military targeting from the upstream stages of their lifecycle, this research will hopefully demonstrate compliance and innovation are not mutually exclusive and can, in fact, be mutually reinforcing. This work has also sought to draw a number of lessons, and contribute, to the ever-growing scholarship examining the intersection between military AI and international law. Specifically, it is hoped that the following contributions have been made:

- Clarity on key IHL considerations for the development of AI technologies for military targeting: The thesis will have provided the reader with a relatively diverse survey of what are the legal considerations that are relevant, and may be used as a guiding framework to foster compliance 'by design' from the earliest stages of these technologies.
- A more granular overview of entry points to foster legal compliance: The notion of 'responsible AI' may, instinctively, seem to be mainly an issue requiring technical solutions. Yet, what 'responsible' concretely consists of and what are some of the key considerations that would contribute to this responsible behaviour is far from being a purely technical reflection. Whether it is through the lens of social sciences, ethics or,

in our case, international law, it is ultimately a multidisciplinary and collective endeavour. As such, it is hoped that this thesis will have provided lawyers with a more granular overview of specific intervention points to foster legal compliance 'by design'. It is also hoped that this thesis will have provided with the non-legal community with an overview of where and when to consult with legal experts.

- Possible approaches to translating legal requirements into technical solutions: One complaint that the author has heard many times over the past years is that international law is too 'complicated' and too 'abstract' to be operationalized. Perhaps these remarks are somewhat warranted – the author thinks that the legal community may need to invest more into an equivalent of 'science communication', i.e., to inform, raise awareness, and get involved with 'audiences that include, at least in part, people from outside the […] community.'[690] Yet, at the same time, more initiatives ought to be dedicated towards building the literacy of the non-legal community on the most pertinent legal issues. This thesis somewhat hopes to contribute to such efforts in 'translating' legal requirements into practical approaches and considerations, and will pave the way for further work in the formulation of concrete measures to foster compliance.

If, after five years, this study has been relatively successful in unearthing the main, relevant international humanitarian law considerations for the design and development of AI technologies for military targeting, it has raised a number of questions that invite for further reflections and extends far beyond the limited answers this thesis sought to provide. Among the many questions that the author encountered, the following areas would merit further exploration as research and efforts at clarifying the law in this space continuously grow:

[690] 'Science Communication' (*Science Europe*) <https://scienceeurope.org/our-priorities/science-communication/> accessed 19 October 2024

- Data practices for legal compliance: While the present thesis sought to provide an overview of what are the key legal considerations related to select aspects of data, there is ample room (and much need) for further research to look into how specific data practices can be leveraged to foster legal compliance. Echoing the messaging throughout this thesis, it is not simply about establishing the minimum threshold for legality to justify military actions; rather, it is about dissecting specific practices that would make a positive impact and enable 'success stories' on compliance.
- Means to reconcile divergences in IHL interpretation: Echoing the growing number of research on this pre-existing issue, IHL can be interpreted (very) differently, often and in practice at the discretion of each State. These divergences may not necessarily be harmful *per se* to upholding international humanitarian law. Yet, inconsistencies in IHL interpretation could, further down the line, jeopardize its authority and may be problematic in the development and procurement of these technologies. If a State's approach and interpretation of the rule of proportionality is fundamentally different to that of another State; should one of these States procure decision support systems with proportionality assessments functions from the other, this could possibly lead towards issues with the implementation of the law. This point may not necessarily have been directly reflected in this thesis' research, however further research as to how these differences should be addressed could be very much useful for the promotion of IHL in the context of AI in the military domain.
- Good practices to foster legal compliance in sales and procurement: As interest and investment in AI technologies for military targeting skyrocket, States will approach the issue of supply and demand differently. Larger States that have the necessary resources, structures and processes in place for 'in-house' research and development will be in a very different position than 'smaller' States that will depend on foreign procurement if

their goal is to possess such capabilities for their armed forces. In fact, it may be that the former will be in a position of selling their products to the latter. As such, guidance is needed to ensure and promote legal compliance, whether it is for the sales or for the procurement of military AI technologies. Further research in this space, providing States with the likes of a handbook that capture good practices for legal compliance, could potentially make a tremendously positive impact for the promotion of IHL in this space.

- Solutions to reconcile black box models with the duty to investigate: Finally, chapter 4 provided the reader with some reflections regarding the challenges that may arise from the inherent opacity of certain models with regards to the duty to investigate. As reports of military AI use in the battlefield emerge and the contentious responses they attract, calls to investigate possible violations of IHL from their use will emerge. Yet, the black box nature of most of the models may very well stand in the way of the ability to conduct such investigations. Hence, further research aiming to develop solutions that will address these challenges would not only be desirable – they will be necessary when formal investigations will need to occur. At this stage, it is not a question of 'if' such investigations would occur, but 'when'.

This list is, of course, far from being exhaustive. In fact, while this thesis seeks to capture the author's accumulated thoughts of over eight years since she was first introduced to the subject in the summer of 2016, it is by no means the end of it. It is very well the author's intention to keep the conversation going and use this thesis as a foundation to conduct further research in this space. While this thesis seeks to capture the state of affairs and provide some insights as to what the existing international humanitarian law considerations relevant to the development of AI for military targeting are; the next steps should focus on putting this knowledge into practice. Far too often, many obstacles stand in the way of concrete steps and solutions for the

implementation and operationalization of governance principles. It is thus hoped that this thesis can be used as a solid starting point in the author's future work and contribution to addressing this bigger issue, which will seek to focus even further on taking the law 'out of the books' – an ode to her participation in the Jean-Pictet Competition back in 2018.

Beyond personal motivations, it is hoped that this thesis, albeit small, could have a hand in advancing the responsible governance and development of these technologies. At the start of this thesis, the author was already convinced that the deployment and use of AI in the military domain is inevitable and as such, promoting legal compliance would be of outmost importance from the earlier stages of these technologies' lifecycle while they are still 'in the lab'. Five years later, as there is acknowledgement that these technologies are 'out of the lab', this conviction has only been reinforced. As States, international organizations, industries, research institutions and civil society organizations alike seek to grapple with the governance of these technologies, there is also the recognition that demand for their development and procurement will continue to grow, exponentially. While the bigger and more resourceful armed forces have invested in their research and development for years now, so-called 'middle powers' are all looking into acquiring and integrating these technologies into their capabilities. As such, the need for guidance on key legal considerations these States would need to take into account in the development and procurement of these technologies will only grow in scale and importance. As a growing number of scholarship is dedicated to providing such guidance, it is hoped that this thesis can contribute to these efforts.

It is also clear that amidst today's geopolitical landscape, the need for such thinking, and for action, is more pressing than ever. Reports of AI use in armed conflict are surfacing in contexts where civilians are unfortunately paying a hefty price. In fact, as the civilian death tolls grow every day, hostages remain in captivity, and an overwhelming sense of fear and danger loom over entire populations, it is clear that the authority and preservation of international

humanitarian law are under extreme pressure. Reported use-cases of AI in military targeting are adding layers of complexity to this situation. While writing this thesis has been particularly difficult for the author over the past year amidst the everyday reporting of civilian casualties and the sheer destruction brought by these conflicts, the author cannot even imagine how harrowing and devastating it must have been for those directly impacted and affected. As such, if the luxury of this intellectual endeavour could play an even small part to upholding international humanitarian law in such difficult circumstances and foster compliance in the longer run, it is hoped that this thesis is, in fact, making a useful contribution.

Bibliography

1. Books

1.1. Authored books

Aggarwal C C, *Neural Networks and Deep Learning: A Textbook* (Springer 2018)

Andrew C *The Secret World: A History of Intelligence* (Penguin Books, 2018)

Boden M, *Artificial Intelligence and Natural Man* (Basic 1977) 360, as cited in

Cardon A, *Beyond Artificial Intelligence: From Human Consciousness to Artificial Consciousness* (ISTE Ltd. 2018)

Cormen T H and others, *Introduction to algorithms* (3rd edn, The MIT Press 2009)

DeLanda M, *War in the Age of Intelligent Machines* (Zone Books 1991)

Flach P, *Machine Learning: The Art and Science of Algorithms* (1st edn, Cambridge University Press 2012)

Grimm V, Railsback S F *Individual-based Modeling and Ecology* (Princeton University Press 2005)

Hassanien A, Dey N, Elghamrawy S, *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach* (1st edn, Springer International Publishing 2020)

Kochenderfer M J, *Decision Making Under Uncertainty: Theory and Application* (MIT Lincoln Laboratory Series (2015)

Kwik J, *Lawfully Using Autonomous Weapon Technologies* (Springer, 2024)

Melzer N, *Targeted Killing in International Law* (Oxford, 2008)

Mohri M, Rostamizadeh A, Talwalkar A, *Foundations of Machine Learning* (2nd edn, The MIT Press 2018)

Murphy K P, *Machine Learning: A Probabilistic Perspective* (The MIT Press, 2012)

Nikolenko S I, *Synthetic Data for Deep Learning* (Springer, 2022) 4

Noor F A, *Data-Gathering in Colonial Southeast Asia 1800-1900: Framing the Other* (Amsterdam, University Press, 2019) <https://www.aup.nl/en/book/9789463724418/data-gathering-in-colonial-southeast-asia-1800-1900> accessed 28 May 2023

Russell S, *Human Compatible: AI and the Problem of Control* (Allen Lane, 2019)

Scharre P, *Army of None: Autonomous Weapons and the Future of War* (W. W. Norton & Company, 2018)

Solis G D, *The Law of Armed Conflict: International Humanitarian Law in War* (1st edn Cambridge University Press, 2010)

1.2. Edited books

Imlay T C, Toft M D (eds) *The Fog of Peace and War Planning: Military and Strategic Planning under Uncertainty* (Routledge, 2006)

Odgaard L (ed) *Strategy in NATO* (Palgrave Studies in Governance, Security and Development, 2014) 179 <https://link.springer.com/book/10.1057/9781137382054>accessed 20 November 2023

Russell S, Norvig P (eds) *Artificial Intelligence: A Modern Approach* (3rd edn, Pearson 2016)

Schmitt M N (ed) *Tallinn Manual 2.0 in the International Law Applicable to Cyber Operations* (Cambridge University Press 2017)

Tsihrintzis G A, Sotiropoulos D N, Jain L C (eds) *Machine Learning Paradigms* (1st edn, Springer International Publishing 2019)

1.3. Contributions to edited books

Bartneck C and others 'Psychological Aspects of AI' in Bartneck C and others (eds) *An Introduction to Ethics in Robotics and AI* (Springer, 2021) <https://link.springer.com/chapter/10.1007/978-3-030-51110-4_7> accessed 17 March 2023

Bennett S 'System Reliability: A Cold War Lesson' in Adlakha-Hutcheon G, Masys A, *Disruption, Ideation and Innovation for Defence and Security* (Springer 2022) <https://link.springer.com/chapter/10.1007/978-3-031-06636-8_2> accessed 15 November 2024

Berrar D 'Cross-validation' in Shoba Ranganathan and others (eds) *Encyclopedia of bioninformatics and computational biology* (1st edition, vol. 1, 2019) <https://oro.open.ac.uk/96776/> accessed 23 November 2024

Bode I, Nadibaidze A '25 Autonomous Drones' in James Patton Rogers (ed) *De Gruyter Handbook of Drone Warfare* (Vol. 4, De Gruyter, 2024) <https://www.degruyter.com/document/doi/10.1515/9783110742039-025/pdf?licenseType=restricted> accessed 27 September 2024

Boothby W H 'Highly Automated and Autonomous Technologies' in William Boothby (ed) New Technologies and the Law in War and Peace (Cambridge University Press, 2018) 140-142 <https://www-cambridge-

org.uniessexlib.idm.oclc.org/core/product/A4911CB3BC08CC35B7D22B9439FD9C13/core-reader> accessed 21 November 2023

Boutin B, Woodcock T 'Aspects of Realizing (Meaningful) Human Control: A Legal Perspective' in Robin Geiß and Henning Lahmann (eds) *Research Handbook on Warfare and Artificial Intelligence* (2024, Edward Elgar Publishing) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4109202> accessed 21 November 2023

Brown L 'Significant effects of radar on the Second World War' in Blumtritt O, Petzold H and Aspray W (eds) *Tracking the History of Radar* (1994, Institute of Eletrical and Electronics Engineers, Inc.) <https://ethw.org/w/images/0/0e/Tracking_the_History_of_Radar.pdf> accessed 15 November 2024

Liveley G and Thomas S 'Homer's Intelligent Machines: AI in Antiquity' in Cave S and Dihal K (eds) *Imagining AI: How the World Sees Intelligent Machines* (Oxford University Press 2023) <https://global.oup.com/academic/product/imagining-ai-9780192865366?cc=ch&lang=en&#> accessed 11 November 2024

Chengeta T 'Autonomous Weapon Systems: Accountability Gaps and Racial Oppression' in Christopher Sabatini (ed) *Reclaiming Human Rights in a Changing World Order* (Brookings Institution Press, 2022) <https://www.chathamhouse.org/2022/10/reclaiming-human-rights-changing-world-order/9-autonomous-weapon-systems-accountability> accessed 11 October 2024

Coco A and Dias T ''Handle with care': due diligence obligations in the employment of AI technologies' in Robin Geiß and Henning Lahmann (eds) *Research Handbook on Warfare and Artificial Intelligence* (2024, Edward Elgar Publishing) <https://www.elgaronline.com/edcollchap/book/9781800377400/book-part-9781800377400-19.xml> accessed 12 October 2024

Cummings M L 'Automation Bias in Intelligent Time Critical Decision Support Systems' in Harris D and Li W (eds) *Decision Making in Aviation* (1st edn, Routledge, 2015) <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315095080-17/automation-bias-intelligent-time-critical-decision-support-systems-cummings> accessed 19 November 2024

Do Amaral Vieira F, 'The Struggles for Corporate Accountability in the UN: A Global South Perspective' in Ananthavinayagan T V, Shenoy A V (eds) *The Wretched of the Global South: Critical Approaches to International Human Rights Law* (Springer, Springer) <https://link.springer.com/chapter/10.1007/978-981-99-9275-1_12> accessed 12 October 2024

Gibson J 'Death by Data: Drones, Kill Lists and Algorithms' in McKay A, Watson A, Karlshøj-Pedersen M (eds), *Remote Warfare: Interdisciplinary Perspectives* (E-International Relations Publishing 2021) <https://www.e-ir.info/2021/02/18/death-by-data-drones-kill-lists-and-algorithms/> accessed 22 November 2024

Kaufman M 'Chapter 22: AI in policing and law enforcement' in Paul R, Carmel E and Cobbe J (eds), *Handbook on Public Policy and Artificial Intelligence* (Edgar 2024)

<https://www.elgaronline.com/edcollchap-oa/book/9781803922171/book-part-9781803922171-31.xml> accessed 11 November 2024

Marguiles P 'The Other Side of Autonomous Weapons: Using Artificial Intelligence to Enhance IHL Compliance' in Robert T P Acala, Eric Talbot Jensen (eds) *The Impact of Emerging Technologies on the Law of Armed Conflict* (Oxford University Press, 2019) <https://academic.oup.com/book/32410/chapter-abstract/268714567?redirectedFrom=fulltext> accessed 9 November 2023

Pacholska M 'Many Hands in the Black Box: Artificial Intelligence and the Responsibility of International Organizations' (2023) in Rossana Deplano, Antal Berkes, Richard Collins (eds) *Reassessing the Articles on the Responsibility of International Organizations: From Theory to Practice* (2023, Edward Elgar) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4501072> accessed 12 April 2024

Sartor G and Andrea A 'The autonomy of technological systems and responsibilities for their use' in Bhuta N, Beck S, Robin Geiß R (eds) *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press 2016) <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/books/autonomous-weapons-systems/autonomy-of-technological-systems-and-responsibilities-for-their-use/9C47BF93BCF884E3D4735C95509790BD> accessed 14 October 2022

Schmitt M N '„Direct Participation in Hostilities" and 21st Century Armed Conflict' in H Fischer (ed) *Crisis Management and Humanitarian Protection* (BWV 2004) <https://www.uio.no/studier/emner/jus/humanrights/HUMR5503/h09/undervisningsmateriale/schmitt_direct_participation_in_hostilties.pdf> accessed 16 July 2023

—— 'Military Necessity and Humanity in International Humanitarian Law: Preserving the Delicate Balance' in Michael N Schmitt, Essays on Law and War at the Fault Lines (Springer, 2012) <https://link.springer.com/chapter/10.1007/978-90-6704-740-1_3> accessed 29 April 2022

—— 'The Interpretive Guidance on the Notion of Direct Participation in Hostilities: A Critical Analysis' in Michael N Schmitt, *Essays on Law and War at the Fault Lines* (Springer, 2012) <https://link.springer.com/chapter/10.1007/978-90-6704-740-1_10> accessed 10 November 2023

Schmitt M N, Widmar E 'The Law of Targeting' in Ducheine P A L, Schmitt M N, Osinga F P B (eds) *Targeting: The Challenges of Modern Warfare* (Springer, 2016) <https://link.springer.com/chapter/10.1007/978-94-6265-072-5_6> accessed 8 November 2023

Seixas-Nunes A, SJ 'Scapegoats! Assessing the Liability of Programmers and Designers for Autonomous Weapons Systems' in Jan Maarten Schragen (ed) *Responsible Use of AI in Military Systems* (CRC Press, 2024) <https://library.oapen.org/handle/20.500.12657/89843> accessed 8 October 2024

Simmons C 'Investigations in armed conflict' in Robert Kolb, Gloria Gaggioli and Pavle Kilibarda (eds) *Research Handbook on Human Rights and Humanitarian Law* (Elgar, 2022)

<https://www.elgaronline.com/display/book/9781789900972/book-part-9781789900972-29.xml> accessed 25 February 2023

Singh N S, Hariharan S, Gupta M 'Facial Recognition Using Deep Learning' in Vanita Jain, Gopal Chaudhary, M. Cengiz Taplamacioglu, M. S. Agarwal (eds) *Advances in Data Sciences, Security and Applications* (Springer, Singapore 2020)

Thurnher J S 'Feasible Precautions in Attack and Autonomous Weapons' (2018) in Von Heinegg W H, Frau R, Singer T (eds) *Dehumanization of Warfare* (Springer, 2018) <https://doi.org/10.1007/978-3-319-67266-3_6> accessed 15 July 2022

Van Baarda T 'Chapter 4: Ethics in the Royal Netherlands Navy' in Van Baarda T, Verweij D E M (eds) *Military Ethics, The Dutch Approach – A Practical Guide* (Brill | Nijhoff, 2006) <https://brill.com/display/book/edcoll/9789047411253/Bej.9789004154407.i-395_005.xml> accessed 23 November 2024

Visger M 'Garbage In, Garbage Out: Data Poisoning Attacks and their Legal Implications' (2022) in Dickinson L A, Berg E W (eds) *Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold* (The Lieber Studies Series, Oxford Academic, 2024) <https://academic.oup.com/book/55259/chapter-abstract/428635911?redirectedFrom=fulltext> accessed 22 November 2024

Wagner M 'Autonomy in the Battlespace: Independently Operating Weapon Systems and the Law of Armed Conflict', in Saxon D (ed) *International Humanitarian Law and the Changing Technology of War* (2013) (Martinus Nijhoff Publishers, Leiden, Boston 2013) <https://brill.com/edcollbook/title/21680?language=en> accessed 22 November 2024

Watts S 'The Genesis and Significance of the Law of War "Rendulic Rule"' in Hayashi N, Lingaas C (eds) *Honest Errors? Combat Decision-Making 75 Years After the Hostage Case* (2024, Springer)

Zuo M J, Huang J and Kuo W 'Multi-state *k*-out-of-*n* Systems' in Pham H (ed), *Handbook of Reliability Engineering* (Springer 2003) 3

1.4. Older works

Clausewitz, Carl von, *Vom Kriege* (Book 1, 1832)

Sun Tzu, *The Art of War* (5th century BC)

2. Articles

2.1. Online academic journals

Adadi A, Berrada M 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' (2018) 6 IEEE Access <https://ieeexplore.ieee.org/abstract/document/8466590/> accessed 4 January 2023

Akande D ‘Clearing the Fog of War? The ICRC’s Interpretative Guidance on Direct Participation in Hostilities’ (2010) 59(1) International & Comparative Law Quarterly <https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/abs/ii-clearing-the-fog-of-war-the-icrcs-interpretive-guidance-on-direct-participation-in-hostilities/707794583F0B2CDFFEDDEEF9F5DDCAAF> accessed 2 February 2023

Alexander P ‘Exploring Bias and Accountability in Military Artificial Intelligence’ (2022) 7 LSE Law Review <https://heinonline.org/HOL/LandingPage?handle=hein.journals/lselr7&div=21&id=&page=> accessed 7 April 2024

Alsolami F and others ‘Development of Self-Synchronized Drones’ Network Using Cluster-Based Swarm Intelligence Approach’ (2021) 9 IEEE Access <https://ieeexplore.ieee.org/abstract/document/9373310> accessed 17 November 2022

Amoroso D & Tamburrini G ‘Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues’ (2020) 1 Current Robotics Reports <https://link.springer.com/article/10.1007/s43154-020-00024-3> accessed 14 October 2022

—— (2021) ‘Toward a Normative Model of Meaningful Human Control over Weapons Systems’ (2021) 35(2) Ethics & International Affairs <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/ethics-and-international-affairs/article/toward-a-normative-model-of-meaningful-human-control-over-weapons-systems/A3FD9EC4CBD6EA77439211537B94A444> accessed 23 November 2024

Anderson M L ‘Why is AI so scary?’ (2005) 169(2) Artificial Intelligence <https://www.sciencedirect.com/science/article/pii/S0004370205001529> accessed 21 November 2024

Baker R S, Hawn A ‘Algorithmic Bias in Education’ (2022) 32 International Journal of Artificial Intelligence in Education <https://link.springer.com/article/10.1007/s40593-021-00285-9> accessed 7 April 2024

Baran B E, Woznyj H M ‘Managing VUCA: The human dynamics of agility’ (2020) 20 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7439966/> accessed 2 February 2023

Beer Y ‘Humanity Considerations Cannot Reduce War’s Hazards Alone Revitalising the Concept of Military Necessity’ (2015) 26(4) European Journal of International Law <https://academic.oup.com/ejil/article/26/4/801/2599602> accessed 21 May 2021

Beier J M ‘Short circuit: retracing the political for the age of ‘autonomous’ weapons’ (2017) 6(1) Critical Military Studies <https://www.tandfonline.com/doi/full/10.1080/23337486.2017.1384978?needAccess=true> accessed 18 October 2023

Block J ‘A laws of war review of contemporary land-based missile defence system ‘iron dome’’, (2017) 45(2) Scientia Militaria: South African Journal of Military Studies

<https://www.ajol.info/index.php/smsajms/article/view/164585> accessed 24 November 2022

Bo M 'Autonomous Weapons and the Responsibility Gap in light of the *Mens Rea* of the War Crime of Attacking Civilians in the ICC Statute' (2021) 19(2) Journal of International Criminal Justice <https://academic.oup.com/jicj/article-abstract/19/2/275/6181757> accessed 12 October 2024

Boothby B 'And for Such Time as: The Time Dimension to Direct Participation in Hostilities' (2009-2010) 42 N.Y.U. J. Int'l L. & Pol <https://heinonline.org/HOL/LandingPage?handle=hein.journals/nyuilp42&div=27&id=&page=> accessed 16 July 2023

Bossaerts P, Murawski C 'Computational Complexity and Human Decision-Making' (2017) 21(12) Trends in Cognitive Sciences <https://www.sciencedirect.com/science/article/abs/pii/S1364661317301936> accessed 2 February 2023

Boutin B 'State responsibility in relation to military applications of artificial intelligence' (2023) 36(1) Leiden Journal of International Law <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/leiden-journal-of-international-law/article/state-responsibility-in-relation-to-military-applications-of-artificial-intelligence/1B0454611EA1F11A8B03A5D2D052C2BE> accessed 22 November 2024

Budiansky S 'German vs. Allied Codebreakers in the Battle of the Atlantic' (2002) 1(1) International Journal of Naval History, <https://www.ijnhonline.org/wp-content/uploads/2012/01/pdf_budiansky1.pdf> accessed 29 May 2023

Bundy A 'Preparing for the future of Artificial Intelligence' (2017) 32 AI & Society <https://dl.acm.org/doi/abs/10.1007/s00146-016-0685-0> accessed 21 November 2024

Burgis-Kasthala M, Sander B 'Conteporary International Criminal Law After Critique: Towards Decolonial and Abolitionist (Dis-)Engagement in an Era of Anti-Impunity' (2024) 22(1) Journal of International Criminal Justice, <https://academic.oup.com/jicj/article/22/1/127/7695263> accessed 11 October 2024

Burridge B 'Post-Modern Warfighting with Unmanned Vehicle Systems: Esoteric Chimera or Essential Capability?' (2005) 150(5) The RUSI Journal <https://www.tandfonline.com/doi/abs/10.1080/03071840509431879> accessed 16 October 2024

Carabantes M 'Black-box artificial intelligence: an epistemological and critical analysis' (2020) 35 AI & Society <https://link.springer.com/article/10.1007/s00146-019-00888-w> accessed 19 January 2023

Carabotti M and others 'The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems' (2015) 28(2) Annals of Gastroenterology <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4367209/> accessed 19 February 2023

Carminati L 'Behavioural Economics and Human Decision Making: Instances from the Health Care System' (2020) 124(6) Health Policy <https://pubmed.ncbi.nlm.nih.gov/32386789/> accessed 28 October 2022

Cave S and Dihal K 'Race and AI: the Diversity Dilemma' (2021) 34 Philosophy & Technology <https://link.springer.com/article/10.1007/s13347-021-00486-z> accessed 17 March 2023

Chengeta T 'Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law' (2016) 45(1) Denver Journal of International Law and Policy <https://heinonline.org/HOL/P?h=hein.journals/denilp45&i=9> accessed 11 October 2024

Cho N, Park J, Kim M 'The Application of Distributed Synthetic Environment Data to a Military Simulation' (2010) 19(4) Journal of the Korea Society for Simulation <https://koreascience.kr/article/JAKO201016450102376.page> accessed 21 July 2022

Choi J, Lim K 'Identifying machine learning techniques for classification of target advertising' (2020) 6(3) ICT Express <https://www.sciencedirect.com/science/article/pii/S2405959520301090> accessed 16 March 2023

Christie E W and others 'Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability' (2024) 4 AI and Ethics <https://link.springer.com/article/10.1007/s43681-023-00261-0> accessed 14 July 2023

Cinà A E and others, 'Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning' (2023) 55(13) ACM Computing Surveys <https://dl.acm.org/doi/full/10.1145/3585385> accessed 10 July 2024

Cotter M 'Military Necessity, Proportionality and Dual-Use Objects at the ICTY: A Close Reading of the Prlić et al. Proceedings on the Destruction of the Old Bridge of Mostar' (2018) 23(2) Journal of Conflict and Security Law <https://academic.oup.com/jcsl/article-abstract/23/2/283/5069317> accessed 16 July 2023

Cramer H and others 'Assessing and addressing algorithmic bias in practice' (2018) 25(6) Interactions <https://dl.acm.org/doi/abs/10.1145/3278156> accessed 7 April 2024

Dalalah D, Dalalah O M A 'The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT' (2023) 21(2) The International Journal of Management Education <https://www.sciencedirect.com/science/article/abs/pii/S1472811723000605> accessed 18 July 2023

Daoust I, Coupland R, Ishoey R 'New wars, new weapons? The obligation of States to assess the legality of means and methods of warfare' (2002) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/new-wars-new-weapons-the-obligation-of-states-to-assess-the-legality-of-means-and-methods-of-warfare/89B6C45677FB9F7833B9EBF5B1A4B895> accessed 6 October 2022

Deeks A 'Predicting Enemies' (2018) 104(8) Virginia Law Review <https://www.jstor.org/stable/26790717> accessed 17 July 2024

Deeks A, Lubell N and Murray D 'Machine Learning, Artificial Intelligence, and the Use of Force by States' (2019) 10(1) Journal of National Security Law & Policy <https://jnslp.com/wp-content/uploads/2019/04/Machine_Learning_Artificial_Intelligence_2.pdf> accessed 24 May 2020

Diakopoulos N & Koliska M 'Algorithmic Transparency in the News Media' (2017) 5(7) Digital Journalism <https://www.tandfonline.com/doi/abs/10.1080/21670811.2016.1208053?journalCode=rdij20> accessed 23 February 2023

Doswald-Beck L 'New Protocol on Blinding Laser Weapons' (1996) 30 June 1996, International Review of the Red Cross' (1996) 36(312) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross-1961-1997/article/abs/new-protocol-on-blinding-laser-weapons/B72BF4A7260BC6B957EF6786A5D79CFE> accessed 21 November 2023

Droege C 'Get off my cloud: cyber warfare, international humanitarian law, and the protection of civilians' (2012) 94(886) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/get-off-my-cloud-cyber-warfare-international-humanitarian-law-and-the-protection-of-civilians/72114EF6E71757FAB2B37E7DE918B2BB> accessed 17 November 2022

Elliot M 'Where Precision Is the Aim: Locating the Targeted Killing Practices of the United States and Israel within International Humanitarian Law' (2009) 47 Canadian Yearbook of International Law <https://heinonline.org/HOL/LandingPage?handle=hein.journals/cybil47&div=6&id=&page=> accessed 9 November 2023

Farber D A 'Uncertainty' (2011) 99 Georgetown Law Journal <https://heinonline.org/HOL/LandingPage?handle=hein.journals/glj99&div=30&id=&page=> accessed 8 December 2022

Felzmann H and others 'Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns' (2019) 6(1) Big Data & Society <https://journals.sagepub.com/doi/full/10.1177/2053951719860542> accessed 16 July 2023

Florez-Morris M 'Joining Guerrilla Groups in Colombia: Individual Motivations and Processes for Entering a Violent Organisation' (2007) 30(7) Studies in Conflict & Terrorism <https://www.tandfonline.com/doi/abs/10.1080/10576100701385958> accessed 3 February 2022

Floridi L 'Why the AI Hype is Anotehr Tech Bubble' (2024) 37 Philosophy & Technology <https://link.springer.com/article/10.1007/s13347-024-00817-w> accessed 11 November 2024

Forden G, Podvig P, Postol T A ‘False alarm, nuclear danger’ (2002) 37(3) IEEE <https://ieeexplore.ieee.org/document/825657> accessed 19 February 2023

Friedman B and Nissenbaum H ‘Bias in Computer Systems’ (1996)14(3) ACM Transactions on Information Systems <https://dl.acm.org/doi/10.1145/230538.230561> accessed 24 May 2020

Gaudreau J ‘The reservations to the Protocols additional to the Geneva Conventions for the protection of war victims’ (2003) 849 International Review of the Red Cross <https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/irrc_849_gaudreau-eng.pdf> accessed 19 November 2024

Geng Z and others ‘Target Recognition in SAR Images by Deep Learning with Trainign Data Augmentation’ (2023) 23(2) Sensors <https://pubmed.ncbi.nlm.nih.gov/36679740/> accessed 7 April 2024

Geuillet A, Couthouis F, Díaz-Rodríguez N ‘Explainability in deep reinforcement learning’ (2021) 214 Knowledge-Based Systems <https://www.sciencedirect.com/science/article/pii/S0950705120308145> accessed 19 October 2024

Ghassemi M, Oakden-Rayner L, Beam A L ‘The false hope of current approaches to explainable artificial intelligence in health care’ (2021) 3(11) The Lancet Digital Health <https://www.sciencedirect.com/science/article/pii/S2589750021002089> accessed 23 February 2023

Gieseken H O and Murphy V ‘The protection of the natural environment under international humanitarian law: The ICRC's 2020 Guidelines’ (2023) 105(924) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/abs/protection-of-the-natural-environment-under-international-humanitarian-law-the-icrcs-2020-guidelines/98B58E81D29737ABB853A350CC50D0C1#:~:text=By%20virtue%20of%20its%20civilian,war%20wrought%20on%20the%20environment> accessed 20 November 2023

Gill P & Horgan J ‘Who Were the Volunteers? The Shifting Sociological and Operational Profile of 1240 Provisional Irish Republican Army Members’ (013) 25(3) Terrorism and Political Violence < https://www.tandfonline.com/doi/full/10.1080/09546553.2012.664587> accessed 3 February 2022

Glanois C and others ‘A survey on interpretable reinforcement learning’ (2024) 113 Machine Learning <https://link.springer.com/article/10.1007/s10994-024-06543-w> accessed 19 October 2024

Goertzel B ‘Human-level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's *The Singularity Is Near*, and McDermott's critique of Kurzweil’ (2007) 171(18) Artificial Intelligence <https://www.sciencedirect.com/science/article/pii/S0004370207001464> accessed 24 May 2020

Goniewicz M, Goniewicz K 'Protection of medical personnel in armed conflicts – case study: Afghanistan' (2013) 39(2) Eur J Trauma Emerg Surg <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3611028/> accessed 20 November 2023

González J L S and others 'Real-time gun detection in CCTV: An open problem' (2020) 132 Neural Networks <https://www.sciencedirect.com/science/article/abs/pii/S0893608020303361> accessed 17 November 2022

Gryz J, Rojszczak M 'Black box algorithms and the rights of individuals: No easy solution to the "explainability" problem' (2021) 10(2) Internet Policy Review <https://www.econstor.eu/handle/10419/235967> accessed 23 February 2023

Guembe B and others 'The Emerging Threat of Ai-driven Cyber Attacks: A Review' (2022) 36(1) Applied Artificial Intelligence, <https://www.tandfonline.com/doi/full/10.1080/08839514.2022.2037254> accessed 21 March 2024

Guidotti R and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51(5) ACM Computing Surveys <https://dl.acm.org/doi/10.1145/3236009> accessed 18 October 2024

Halevy A, Norvig P and Pereira F 'The Unreasonable Effectiveness of Data' (2009) 24(2) IEEE Intelligent Systems <https://ieeexplore.ieee.org/document/4804817> accessed 21 July 2022

Hammond D N 'Autonomous Weapons and the Problem of State Accountability' (2014-2015) 15 Chicago Journal of International Law <https://heinonline.org/HOL/LandingPage?handle=hein.journals/cjil15&div=26&id=&page=> accessed 20 October 2022

Harrison A J 'Guidelines on Investigating Violations of International Humanitarian Law and Practical Mechanisms for Accountability - Today and Beyond' (2021) 229(2) Military Law Review <https://heinonline.org/HOL/LandingPage?handle=hein.journals/milrv229&div=16&id=&page=> accessed 25 February 2023

Hassan M, Mollick S, Yasmin F 'An unsupervised cluster-based feature grouping model for early diabetes detection' (2022) 2 Healthcare Analytics <https://www.sciencedirect.com/science/article/pii/S2772442522000521> accessed 8 July 2024

Hays Parks W 'Means and Methods of Warfare' 38 George Washington International Law Review <https://heinonline.org/HOL/P?h=hein.journals/gwilr38&i=521> accessed 6 October 2022

Heinze M 'On 'Gun Culture' and 'Civil Statehood' in Yemen' (2014) 4(1) Journal of Arabian Studies <https://www.tandfonline.com/doi/full/10.1080/21534764.2014.920190?needAccess=true> accessed 5 July 2023

Heller K J, ‘The Concept of ‘the Human’ in the Critique of Autonomous weapons’ (2023) 15 Harvard National Security Journal <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342529> accessed 3 June 2025

Heyns C ‘Autonomous weapons in armed conflict and the right to a dignified life: an African perspective’ (2017) 33(1) South African Journal on Human Rights <https://journals.co.za/doi/abs/10.1080/02587203.2017.1303903> accessed 18 November 2022

Hinds J, Williams E J, Joinson A N ‘“It wouldn't happen to me”: Privacy concerns and perspectives following the Cambridge Analytica scandal’ (2020) 143 International Journal of Human-Computer Studies <https://www.sciencedirect.com/science/article/abs/pii/S1071581920301002> accessed 16 March 2023

Hodges J S (1987) ‘Uncertainty, Policy Analysis and Statistics’ (1987) 2(3) Statistical Science <https://projecteuclid.org/journals/statistical-science/volume-2/issue-3/Uncertainty-Policy-Analysis-and-Statistics/10.1214/ss/1177013224.full> accessed 2 February 2023

Hoijtink M, Planqué-van-Hardeveld A ‘Machine Learning and the Platformization of the Military: A Study of Google's Machine Learning Platform TensorFlow’ (2022) 16(2) International Political Sociology <https://academic.oup.com/ips/article/16/2/olab036/6562417> accessed 30 September 2022

Horowitz M C ‘The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons’ (2016) 145(4) Daedalus <https://direct.mit.edu/daed/article/145/4/25/27111/The-Ethics-amp-Morality-of-Robotic-Warfare> accessed 18 November 2022

Horowitz M C, Kahn L ‘Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts’ (2024) 68(2) International Studies Quarterly <https://academic.oup.com/isq/article-abstract/68/2/sqae020/7638566> accessed 19 November 2024

Horowitz M C and Pindyck S ‘What is a military innovation and why it matters’ (2022) 46(1) Journal of Strategic Studies <https://www.tandfonline.com/doi/full/10.1080/01402390.2022.2038572> accessed 16 October 2024

Huelss H ‘Transcending the fog of war? US military ‘AI’, vision, and the emergent post-scopic regime’ (2024) European Journal of International Security <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/european-journal-of-international-security/article/transcending-the-fog-of-war-us-military-ai-vision-and-the-emergent-postscopic-regime/35BCDEE8E28B076BCD597AFDC8976824> accessed 18 October 2024

Huq A Z ‘Racial Equity in Algorithmic Criminal Justice’ (2019) 68(6) Duke Law Journal <https://heinonline.org/HOL/LandingPage?handle=hein.journals/duklr68&div=33&id=&page=> accessed 7 April 2024

Jevglevskaja N ‘Weapons Review Obligation under Customary International Law’ (2018) 94 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1724&context=ils> accessed 22 September 2022

Iloka C P ‘Precision Attack and Reparation of the Vulnerable under International Humanitarian Law: An Appraisal’ (2022) 7 African Journal of Criminal Law and Jurisprudence <https://heinonline.org/HOL/LandingPage?handle=hein.journals/afcjlocil7&div=21&id=&page=> accessed 9 November 2023

Ishowo-Oloko F and others ‘Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation’ (2019) 1 Nature Machine Intelligence <https://www.nature.com/articles/s42256-019-0113-5> accessed 23 February 2023

Jin S and others ‘Garbage detection and classification using a new deep learning-based machine vision system as a tool for sustainable waste recycling’ (2023), 162 Waste Management, <https://www.sciencedirect.com/science/article/abs/pii/S0956053X23001915> accessed 22 March 2024

Joshi K, Patel M I ‘Recent advances in local feature detector and descriptor: a literature survey’ (2020) 9 International Journal of Multimedia Information Retrieval <https://doi.org/10.1007/s13735-020-00200-3> accessed 28 July 2022

Kaloudi N, Li J ‘The AI-Based Cyber Threat Landscape: A Survey’ (2020) 53(1) ACM Computing Surveys <https://dl.acm.org/doi/abs/10.1145/3372823> accessed 18 October 2023

Kaplan A, Haenlein M ‘Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence’ (2019) 62(1) Business Horizons, <https://www.sciencedirect.com/science/article/pii/S0007681318301393> accessed 11 November 2024

Khanna S, Srivastava S, ‘AI Governance in Healthcare: Explainability Standards, Safety Protocols, and Human-AI Interactions Dynamics in Contemporary Medical AI Systems’ (2021) 1(1) Empirical Quests for Management Essences <https://www.researchberg.com/index.php/eqme/article/view/166> accessed 11 November 2024

Khera R, Simon M A, Ross J A ‘Automation Bias and Assistive AI: Risk of Harm From AI-Driven Clinical Decision Support’ (2023) 330(23) JAMA <https://jamanetwork.com/journals/jama/fullarticle/2812931> accessed 19 November 2024

Kilkenny M F, Robinson K M ‘Data quality: “Garbage in – garbage out”’ (2018) 47(3) Health Information Management Journal <https://journals.sagepub.com/doi/pdf/10.1177/1833358318774357> accessed 18 May 2023

Kim H and others, ‘Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study’ (2020) 2(3) Lancet Digital Health <https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30003-0/fulltext?amp=1> accessed 18 July

Koffman J and others 'Uncertainty and COVID-19: how are we to respond?' (2020) 113(6) Journal of the Royal Society of Medicine <https://journals.sagepub.com/doi/pdf/10.1177/0141076820930665> accessed 2 February 2023

Königstorfer F, Thalmann S 'AI Documentation: A path to accountability' (2022) 11 Journal of Responsible Technology <https://www.sciencedirect.com/science/article/pii/S2666659622000208> accessed 12 October 2024

Koshiyama A, Kazim E, Treleaven P 'Algorithm Auditing: Managing the Legal, Ethical, and Technological Risks of Artificial Intelligence, Machine Learning, and Associated Algorithms' (2022) Computer 55(4) <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9755237> accessed 22 September 2022

Kosta E 'Algorithmic state surveillance: Challenging the notion of agency in human rights' (2022) 16(1) Regulation & Governance <https://onlinelibrary.wiley.com/doi/full/10.1111/rego.12331> accessed 3 September 2022

Kronemann B and others, 'How AI encourages consumers to share their secrets? The role of anthropomorphism, personalisation, and privacy concerns and avenues for future research' (2023) 27(1) Spanish Journal of Marketing – ESIC <https://www.emerald.com/insight/content/doi/10.1108/SJME-10-2022-0213/full/html> accessed 17 March 2023

Kuehnhanss C R 'The challenges of behavioural insights for effective policy design' (2019) 38(1) Policy and Society <https://academic.oup.com/policyandsociety/article-abstract/38/1/14/6403979> accessed 28 October 2022

Kuipers M, Prasad R 'Journey of Artificial Intelligence' (2022) 123 Wireless Personal Communications, <https://link.springer.com/article/10.1007/s11277-021-09288-0> accessed 5 January 2024

Kupek E 'How many more? Under-reporting of the COVID-19 deaths in Brazil in 2020' (2021) 26(9) Tropical Medicine & International Health <https://onlinelibrary.wiley.com/doi/full/10.1111/tmi.13628> accessed 14 April 2023

Kurado K, Ludvigson S C, Ng S (2015), 'Measuring Uncertainty' (2015) 105(3) American Economic Review <https://www.aeaweb.org/articles?id=10.1257/aer.20131193> accessed 8 December 2022

LaGrandeur K 'The consequences of AI hype' (2023) 4 AI Ethics <https://link.springer.com/article/10.1007/s43681-023-00352-y> accessed 13 March 2024

Lam R and others, 'Learning skillful medium-range global weather forecasting' (2023) 382(6677) Science <https://www.science.org/stoken/author-tokens/ST-1550/full> accessed 17 July 2024

LeCun Y, Bengio Y, Hiton G ‘Deep learning’ (2015) 521 Nature <https://www.nature.com/articles/nature14539> accessed 24 May 2020

Lee C and others ‘Deep AI military staff: cooperative battlefield situation awareness for commander’s decision making’ (2023) 79 The Journal of Supercomputing <https://link.springer.com/article/10.1007/s11227-022-04882-w> accessed 12 April 2024

Lehr D & Ohm P ‘Playing with the Data: What Legal Scholars Should Learn About Machine Learning’ (2017) 51 UC Davis Law Review <https://heinonline.org/HOL/Page?handle=hein.journals/davlr51&collection=journals&id=667&startid=&endid=732> (accessed 28 April 2022)

Levin E ‘Journalists as a Protected Category: A New Status for the Media in International Humanitarian Law’ (2013) 17 UCLA Journal of International Law and Foreign Affairs <https://www.jstor.org/stable/45302375> accessed 20 November 2023

Liu F, Liu Y, Shi Y ‘Three IQs of AI systems and their testing methods’ (2020) 2020(13) The Journal of Engineering <https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/joe.2019.1135> accessed 17 October 2024

Lloyd S ‘Measures of complexity: a nonexhaustive list’ (2001) 21(4) IEEE Control Systems Magazine <https://ieeexplore.ieee.org/document/939938> accessed 2 February 2023

Longobardo M ‘The Relevance of the Concept of Due Diligence for International Humanitarian Law’ (2019) 37 Wisconsin International Law Journal <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3570423> accessed 7 November 2022

Lovato J, Witte Zimmerman J, ‘Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership and Fairness in AI Generative Art’ (2024) 7(1) Proceedings of the AAAI/ACM Conference on AI, Ethics and Society <https://ojs.aaai.org/index.php/AIES/article/view/31691> accessed 11 November 2024

Lubell N ‘Lawful Targets in Cyber Operations: Does the Principle of Distinction Apply?’ (2013) 89 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1026&context=ils> accessed 9 November 2023

Lv C and others ‘Innovative applications of artificial intelligence during the COVID-19 pandemic’ (2024) 3(1) Infectious Medicine <https://www.sciencedirect.com/science/article/pii/S2772431X24000091> accessed 11 November 2024

Madianou M ‘Nonhuman humanitarianism: when ‘AI for good’ can be harmful’ (2021) 6 Information, Communication & Society <https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.1909100> accessed 11 November 2024

Magrabi F and others ‘Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications’ (2019) 28(01) Yearbook of Medical Informatics,

<https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0039-1677903> accessed 17 October 2024

Maia P 'The Case for Interfaces in International Relations' (2023) 3(3) International Political Sociology <https://academic.oup.com/isagsq/article/3/3/ksad054/7274820?login=false> accessed 16 November 2023

Marotta V and others, 'The Welfare Impact of Targeted Advertising Technologies' (2022) 33(1) Information Systems Research <https://pubsonline.informs.org/doi/abs/10.1287/isre.2021.1024> accessed 16 March 2023

Mauri D 'The Holy See's Position on Lethal Autonomous Weapons Systems: An Appraisal through the Lens of the Martens Clause' (2020) 11(1) Journal of International Humanitarian Legal Studies, <https://brill.com/view/journals/ihls/11/1/article-p116_116.xml?ebody=pdf-60564> accessed 18 November 2022

McClelland J 'The review of weapons in accordance with Article 36 of Additional Protocol I' (2003) 85(850) <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/review-of-weapons-in-accordance-with-article-36-of-additional-protocol-i/61B38D9967B45A0EDE039E13D2024EAA> accessed 23 September 2022

McCormack T 'International Humanitarian Law and the Targeting of Data' (2018) 94 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1725&context=ils> accessed 9 November 2023

McGregor L, Murray D and Ng V 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68(2) International & Comparative Law Quarterly <https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/international-human-rights-law-as-a-framework-for-algorithmic-accountability/1D6D0A456B36BA7512A6AFF17F16E9B6> accessed 3 September 2022

McNamara C 'Targeting Decision sin the Crosshairs of Humanitarian Law and Human Rights Law' (2014) 5(2) <https://heinonline.org/HOL/Page?handle=hein.journals/creintcl5&div=15&id=&page=&collection=journals> accessed 8 November 2023

Meerveld H W and others, 'The irresponsibility of not using AI in the military' (2023) 25 Ethics and Information Technology <https://link.springer.com/article/10.1007/s10676-023-09683-0> accessed 18 October 2024

Mendi A F and others, 'Applications of Reinforcement Learning and its Extension to Tactical Simulation Technologies' (2021) 22(1) International Journal of Simulation: Systems, Science & Technology, <https://ijssst.info/Vol-22/No-1/paper14.pdf> accessed 22 March 2024

Menon D and Ranganathan R, 'A Generative Approach to Materials Discovery, Design, and Optimization' (2022) ACS Omega <https://pubs.acs.org/doi/10.1021/acsomega.2c03264> accessed 27 February 2023

Mero T 'The Martens Clause, Principles of Humanity, and Dictates of Public Conscience' (2000) 94(1) American Journal of International Law <https://www.cambridge.org/core/journals/american-journal-of-international-law/article/abs/martens-clause-principles-of-humanity-and-dictates-of-public-conscience/F55EECE5BED3DDB9D78162DA4509A03A> accessed 18 November 2022

Mihlar F 'Global Models, Victim Disconnect and Demands for International Intervention: The Dilemma of Decoloniality and Transitional Justice in Sri Lanka' (2024) 16(3) Journal of Human Rights Practice, <https://academic.oup.com/jhrp/advance-article-abstract/doi/10.1093/jhuman/huae024/7748714> accessed 11 October 2024

Milan S 'Techno-solutionism and the standard human in the making of the COVID-19 pandemic' (2020) 7(2) Big Data & Society <https://journals.sagepub.com/doi/full/10.1177/2053951720966781> accessed 13 January 2023

Milaninia N 'Biases in machine learning models and big data analytics: The international criminal and humanitarian law implications' (2021) 102(913) International Review of the Red Cross, <https://international-review.icrc.org/articles/biases-machine-learning-big-data-analytics-ihl-implications-913> accessed 18 March 2021

Montgomery M R and others 'Measuring Living Standards with Proxy Variables' (2000) 37(2) Demography <https://link.springer.com/article/10.2307/2648118> accessed 16 April 2021

Neelamegam S, Ramaraj E 'Classification algorithm in Data mining: An Overview' (2013) 3(5) International Journal of P2P Network Trends and Technology <http://www.ijpttjournal.org/volume-3/issue-8/IJPTT-V3I8P101.pdf> accessed 19 November 2020

Pacholska M 'Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective' (2023) 56(1) Israel Law Review <https://www-cambridge-org.uniessexlib.idm.oclc.org/core/journals/israel-law-review/article/military-artificial-intelligence-and-the-principle-of-distinction-a-state-responsibility-perspective/972D0662B1207C14656D5128B7B139FA?utm_campaign=shareaholic&utm_medium=copy_link&utm_source=bookmark> accessed 19 October 2024

Paleyes A, Urma R, Lawrence N D 'Challenges in Deploying Machine Learning: A Survey of Case Studies' (2023) 55(6) ACM Computing Surveys <https://dl.acm.org/doi/10.1145/3533378> accessed 18 October 2024

Pan Y and others, 'The adoption of artificial intelligence in employee recruitment: The influence of contextual factors' (2021) 33(6) The International Journal of Human Resource Management <https://www.tandfonline.com/doi/abs/10.1080/09585192.2021.1879206> accessed 16 July 2023

Panch T, Mattie H, Atun R 'Artificial intelligence and algorithmic bias: implications for health systems' (2019) 9(2) Journal of Global Health <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875681/> accessed 7 April 2024

Parrin A '"How did they die?": Bridging humanitarian and criminal-justice objectives in forensic science to advance the rights of families of the missing under international humanitarian law' (2023) 105(923) International Review of the Red Cross <https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/how-did-they-die-bridging-humanitarian-and-criminaljustice-objectives-in-forensic-science-to-advance-the-rights-of-families-of-the-missing-under-international-humanitarian-law/8A3A55040D256FEA79387512E25F9751> accessed 26 February 2023

Parycek P, Schmid V and Novak A 'Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions' (2024) 15 Journal of the Knowledge Economy <https://link.springer.com/article/10.1007/s13132-023-01433-3> accessed 19 November 2024

Perry W J 'Military technology: an historical perspective' (2004) 26(2-3) Technology in Society <https://www.sciencedirect.com/science/article/abs/pii/S0160791X04000363> accessed 16 October 2024

Popović Z B, Thomas J D 'Assessing observer variability: a user's guide' (2017) 7(3) Cardiovasc Diagn Ther <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440257> accessed 14 July 2022

Porciello J and others, 'Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning' (2020) 2 Nature Machine Intelligence <https://www.nature.com/articles/s42256-020-00235-5> accessed 11 November 2024

Portugal I, Alencar P, Cowan D 'The use of machine learning algorithms in recommender systems: A systematic review' (2018) 97 Expert Systems with Applications, <https://www.sciencedirect.com/science/article/abs/pii/S0957417417308333> accessed 22 March 2024

Price W N II 'Medical AI and Contextual Bias' (2019) 33(1) Harvard Journal of Law & Technology <https://heinonline.org/HOL/LandingPage?handle=hein.journals/hjlt33&div=6&id=&page=> accessed 16 July 2023

Pu G and others 'A hybrid unsupervised clustering-based anomaly detection method' (2021) 26(2) Tsinghua Science and Technology, <https://ieeexplore.ieee.org/abstract/document/9147152> accessed 22 March 2024

Rathburn B C 'Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory' (2007) 51(3) International Studies Quarterly <https://academic.oup.com/isq/article/51/3/533/1795465> accessed 2 February 2023

Renic N and Schwarz E 'Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing' (2023) 37(3) Ethics & International Affairs <https://www.cambridge.org/core/journals/ethics-and-international-affairs/article/crimes-of-dispassion-autonomous-weapons-and-the-moral-challenge-of-systematic-killing/ECDC02C13BA23C432C9B11D84ACBB303> accessed 30 September 2024

Ridzuan F, Zainon W M N W ‘A Review on Data Cleansing Methods for Big Data’ (2019) 161 Procedia Computer Science, <https://www.sciencedirect.com/science/article/pii/S1877050919318885> accessed 22 March 2024

Rogers S K and others ‘Neural networks for automatic target recognition’ (1995) 8(7-8) Neural Networks <https://www.sciencedirect.com/science/article/abs/pii/089360809500050X> accessed 7 April 2024

Rosen C ‘Technosolutionism Isn’t the Fix’ (2020) 22(3) The Hedgehog Review <https://go.gale.com/ps/i.do?id=GALE%7CA645242019&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=15279677&p=AONE&sw=w&userGroupName=anon~fdd2b5a8> accessed 17 March 2023

Rudin C ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’ (2019) 1 Nature Machine Intelligence <https://www.nature.com/articles/s42256-019-0048-x> accessed 23 February 2023

Ryan M ‘In AI We Trust: Ethics, Artificial Intelligence, and Reliability’ (2020) 26 Science and Engineering Ethics <https://link.springer.com/article/10.1007/s11948-020-00228-y> accessed 17 March 2023

Sandhu A, Fussey P ‘The ‘uberization of policing’? How police negotiate and operationalise predictive policing technology’ (2021) 31(1) Policing and Society <https://www.tandfonline.com/doi/full/10.1080/10439463.2020.1803315> accessed 10 July 2024

Saygin A P, Cicekli I, Akman V ‘Turing Test: 50 Years Later’ (2000) 10 Minds and Machines <https://link.springer.com/article/10.1023/A:1011288000451> accessed 20 March 2024

Schmitt M N ‘Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics, (2013) 4 Harvard National Security Journal <https://harvardnsj.org/2013/02/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/> accessed 1 August 2022

—— ‘Precision attack and international humanitarian law’ (2005) 87(859) International Review of the Red Cross <https://international-review.icrc.org/sites/default/files/irrc_859_3.pdf> accessed 20 November 2023

—— ‘Targeting and Humanitarian Law: Current Issues’ (2004) 34 Israel Yearbook on Human Rights <https://brill.com/display/book/edcoll/9789047414070/B9789047414070_s005.xml> accessed 9 November 2023

—— ‘Targeting and International Humanitarian Law in Afghanistan’ (2009) 39 Israel Yearbook on Human Rights <https://brill.com/display/book/edcoll/9789047443773/BP000005.xml> accessed 9 November 2023

Schmitt M N, Merriam J J 'The Tyranny of Context: Israeli Targeting Practices in Legal Perspective' (2015) University of Pennsylvania Journal of International Law <https://heinonline.org/HOL/LandingPage?handle=hein.journals/upjiel37&div=5&id=&page=> accessed 9 December 2022

Schmitt M N, Schauss M 'Uncertainty in the Law of Targeting: Towards a Cognitive Framework' (2019) 10 Harvard National Security Journal <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3755556> accessed 2 February 2023

Schmitt M N & Thurnher J '"Out of the loop": autonomous weapon systems and the law of armed conflict' (2013) 4(2) Harvard National Security Journal <https://harvardnsj.org/wp-content/uploads/sites/13/2013/01/Vol-4-Schmitt-Thurnher.pdf> accessed 7 April 2020

Schmitt M N, Watts S 'Common Article 1 and the Duty to "Ensure Respect"' (2020) 96 Int'l L. Stud. 674 <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=2936&context=ils> accessed 7 February 2024

Schmitt M N & Widmar E W '"On Target": Precision and Balance in the Contemporary Law of Targeting' (2014) 7(3) Journal of National Security Law and Policy <https://heinonline.org/HOL/LandingPage?handle=hein.journals/jnatselp7&div=18&id=&page=, https://link.springer.com/chapter/10.1007/978-94-6265-072-5_6> accessed 9 December 2022

Schraagen J M 'Responsible use of AI in military systems: prospects and challenges' 66(11) Ergonomics <https://doi.org/10.1080/00140139.2023.2278394> accessed 10 May 2024

Scott A C and others 'Explanation Capabilities of Production-Based Consultation Systems' (1977) American Journal of Computational Linguistics <https://aclanthology.org/J77-1006/> accessed 23 February 2023

Selbst A D, Powles J 'Meaningful information and the right to explanation' (2017) 7(4) International Data Privacy Law <https://academic.oup.com/idpl/article/7/4/233/4762325> accessed 23 February 2022

Sharkey A 'Autonomous weapons systems, killer robots and human dignity' (2019) 21 Ethics and Information Technology, <https://link.springer.com/article/10.1007/s10676-018-9494-0> accessed 18 November 2022

Silver D and others 'Mastering the game of Go with deep neural networks and tree search' (2016) 529 Nature, <https://www.nature.com/articles/nature16961> accessed 22 March 2024

Sivakumar S others 'An empirical study of supervised learning methods for breast cancer diseases' (2018) 175 Optik, <https://www.sciencedirect.com/science/article/abs/pii/S0030402618312622> accessed 22 March 2024

Smith G J D ‘The politics of algorithmic governance in the black box city’ (2020) Big Data & Society <https://journals.sagepub.com/doi/pdf/10.1177/2053951720933989> accessed 23 February 2023

Stix C and Maas M M ‘Bridging the gap: the case for an ‘Incompletely Theorized Agreement’ on AI policy’ (2021) 1 AI and Ethics <https://link.springer.com/article/10.1007/s43681-020-00037-w> accessed 11 November 2024

Stottlemyre S A ‘HUMINT, OSINT, or Something New? Defining Crowdsourced Intelligence’ (2015) 28(3) International Journal of Intelligence and Counterintelligence <https://www.tandfonline.com/doi/full/10.1080/08850607.2015.992760> accessed 27 February 2023

Swartout W R ‘XPLAIN: a system for creating and explaining expert consulting programs’ (1983) 21(3) Artificial Intelligence <https://www.sciencedirect.com/science/article/abs/pii/S0004370283800149> accessed 20 January 2023

Thompson D F ‘Moral Responsibility of Public Officials: The Problem of Many Hands’ (1980) 74(4) American Political Science Review <https://doi.org/10.2307/1954312> accessed 12 October 2024

Tomsett R and others ‘Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI’ (2020) 1(4) Patterns <https://www.cell.com/patterns/fulltext/S2666-3899(20)30060-X> accessed 18 October 2024

Tuna O F, Catak F O, Eskil M T, ‘Uncertainty as a Swiss army knife: new adversarial attack and defense ideas based on epistemic uncertainty’ (2023) 9 Complex & Intelligent Systems, <https://link.springer.com/article/10.1007/s40747-022-00701-0> accessed 18 October 2024

Tuyterlaars T and Mikolajczyk K ‘Local Invariant Feature Detectors: A Survey’ (2007) 3(3) Foundations and Trends in Computer Graphics and Vision <https://homes.esat.kuleuven.be/~tuytelaa/FT_survey_interestpoints08.pdf> accessed 28 July 2022

Utesch F and others ‘Towards behaviour based testing to understand the black box of autonomous cars’ (2020) 12 European Transport Research Review <https://etrr.springeropen.com/articles/10.1186/s12544-020-00438-2> accessed 4 January 2023

Va Berkel N and others ‘Human-centred artificial intelligence: a contextual morality perspective’ (2020) 41(3) Behaviour & Information Technology <https://www.tandfonline.com/doi/abs/10.1080/0144929X.2020.1818828> accessed 16 July 2023

Vale D, El-Sharif A, Ali M ‘Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law’ (2022) 2 AI and Ethics <https://link.springer.com/article/10.1007/s43681-022-00142-y> accessed 23 February 2023

Verdiesen I, Santoni de Sio F, Dignum V 'Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight' (2021) 31 Minds and Machines <https://doi.org/10.1007/s11023-020-09532-9> accessed 6 August 2022

Vered M and others 'The effects of explanations on automation bias' (2023) 322 Artificial Intelligence <https://www.sciencedirect.com/science/article/pii/S000437022300098X> accessed 19 November 2024

Vidgen B, Derczynski L 'Directions in abusive language training data, a systematic review: Garbage in, garbage out' (2020) 15(12) PLOS ONE, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0243300> accessed 26 May 2023

Vinge V 'Signs of the singularity' (2008) 45(6) IEEE Spectrum <https://ieeexplore.ieee.org/document/4531467> accessed 21 November 2024

Visvizi A 'Artificial Intelligence (AI) and Sustainable Development Goals (SDGs): Exploring the Impact of AI on Politics and Society' (2022) 14(3) Sustainability <https://www.mdpi.com/2071-1050/14/3/1730> accessed 11 November 2024

Vouros G A 'Explainable Deep Reinforcement Learning: State of the Art and Challenges' (2023) 55(5) ACM Computing Surveys <https://dl.acm.org/doi/10.1145/3527448> accessed 19 October 2024

Wang X and others 'The security of machine learning in an adversarial setting: A survey' (2019) 130 Journal of Parallel and Distributed Computing <https://www.sciencedirect.com/science/article/pii/S0743731518309183> accessed 17 October 2024

Warren M 'The "Fog of Law": The Law of Armed Conflict in Operation Iraqi Freedom' (2010) 86 International Law Studies <https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1103&context=ils> accessed 10 November 2023

Winter E 'Pillars not Principles: The Status of Humanity and Military Necessity in the Law of Armed Conflict' (2020) 25(1) Journal of Conflict and Security Law <https://academic.oup.com/jcsl/article/25/1/1/5722185?login=true> accessed 19 October 2024

Yan L and others 'Promises and challenges of generative artificial intelligence for human learning' (2024) 8 Nature Human Behaviour, <https://www.nature.com/articles/s41562-024-02004-5> accessed 11 November 2024

Yan N, Huang S, Kong C 'Reinforcement Learning-Based Autonomous Navigation and Obstacle Avoidance for USVs under Partially Observable Conditions' (2021) 2021 Mathematical Problems in Engineering, <https://www.hindawi.com/journals/mpe/2021/5519033/> accessed 22 March 2024

Yu A C, Eng J ‘One Algorithm May Not Fit All: How Selection Bias Affects Machine Learning Performance’ (2020) 40(7) RadioGraphics <https://pubs.rsna.org/doi/full/10.1148/rg.2020200040> accessed 12 April 2024

Zhang Q and others ‘Towards cross-task universal perturbation against black-box object detectors in autonomous driving’ (2020) 180 Computer Networks <https://www.sciencedirect.com/science/article/abs/pii/S138912862030606X> accessed 4 January 2023

Zhu L, Majumdar S, Ekenna C ‘An invisible warfare with the internet of battlefield things: A literature review’ (2021) 3 Human Behaviour & Emerging Technologies <https://onlinelibrary.wiley.com/doi/epdf/10.1002/hbe2.231> accessed 27 May 2022

Zhu X and others ‘Do We Need More Training Data?’ (2016) 119 International Journal of Computer Vision <https://link.springer.com/article/10.1007/s11263-015-0812-2> accessed 21 July 2022

Zhuo L, Han D ‘Agent-based modelling and flood risk managemen: A compendious literature review’ (2020) 591 Journal of Hydrology <https://www.sciencedirect.com/science/article/abs/pii/S0022169420310611> accessed 18 September 2024

2.2. Reports

—— ‘Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies’ (2018) UNIDIR Resources No. 9 < https://unidir.org/publication/algorithmic-bias-and-weaponization-increasingly-autonomous-technologies> accessed 24 May 2020

—— ‘Automated Apartheid: How Facial Recognition Fragments, Segregates and Controls Palestinians in the OPT’ (2023) Amnesty International, <https://www.amnesty.org/en/documents/mde15/6701/2023/en/> accessed 29 June 2023

—— ‘Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects’ (2014) Expert Meeting Report, ICRC <https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014> accessed 25 March 2020

—— ‘Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977: Commentary of 1987’ (1987) ICRC <https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=D80D14D84BF36B92C12563CD00434FBD> accessed 21 May 2021

—— ‘The Bombing of Iraqi Cities: Middle East Watch Condemns Bombing Without Warning of Air Raid Shelter in Baghdad’s Al Ameriyya District on February 13’ (1991) Human Rights Watch <https://www.hrw.org/reports/1991/IRAQ291.htm> accessed 4 February 2022

—— ‘The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward’ (2014) (UNIDIR Resources

No. 2 <https://www.unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf> accessed 24 May 2020

Afina Y 'The Global Kaleidoscope of Military AI Governance: Decoding the 2024 Regional Consultations on Responsible AI in the Military Domain' (2024) UNIDIR <https://unidir.org/publication/the-global-kaleidoscope-of-military-ai-governance/> accessed 11 November 2024

Afina Y 'Draft Guidelines for the Development of a National Strategy on AI in Security and Defence: A Policy Brief' (2024) UNIDIR <https://unidir.org/publication/draft-guidelines-for-the-development-of-a-national-strategy-on-ai-in-security-and-defence/> accessed 11 November 2024

Afina Y 'Regional Perspectives on the Application of International Humanitarian Law to Lethal Autonomous Weapon Systems' (2025) UNIDIR <https://unidir.org/publication/regional-perspectives-on-the-application-of-international-humanitarian-law-to-lethal-autonomous-weapon-systems/> accessed 3 June 2025

Afina Y, Grand-Clément S 'Bytes and Battles: Inclusion of Data Governance in Responsible Military AI' (2024) Centre for International Governance Innovation <https://www.cigionline.org/publications/bytes-and-battles-inclusion-of-data-governance-in-responsible-military-ai/> accessed 22 November 2024

Anderljung M and others, 'Frontier AI regulation: Managing emerging risks to public safety' (2023) OpenAI <https://openai.com/research/frontier-ai-regulation> accessed 23 January 2024

Bartlett J, Reynolds L 'The state of the art 2015: A literature review of social media intelligence capabilities for counter-terrorism' (2015) Demos <https://demos.co.uk/wp-content/uploads/2015/09/State_of_the_Arts_2015-1.pdf> accessed 22 March 2024

Blacknell D, Luc Vignaud 'ATR of Ground Targets: Fundamentals and Key Challenges' (2013) EN-SET-172-2013-01, NATO Science & Technology Organization, <https://www.sto.nato.int/publications/STO%20Educational%20Notes/STO-EN-SET-172-2013/EN-SET-172-2013-01.pdf> accessed 12 November 2024

Bo M, Bruun L and Boulanin V 'Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS' (2022) SIPRI <https://www.sipri.org/sites/default/files/2022-10/2210_aws_human_responsibility.pdf> accessed 14 October 2022

Bondar K 'Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare' (2025) Center for Strategic International Studies, <https://www.csis.org/analysis/ukraines-future-vision-and-current-capabilities-waging-ai-enabled-autonomous-warfare> accessed 3 June 2025

Boulanin V and others 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control' (2020) SIPRI <https://www.sipri.org/publications/2020/policy-

reports/limits-autonomy-weapon-systems-identifying-practical-elements-human-control> accessed 6 January 2023

Chahal H, Fedasiuk R, Flynn C 'Messier than Oil: Assessing Data Advantage in Military AI' (2020) CSET Issue Brief, Center for Security and Emerging Technology <https://cset.georgetown.edu/wp-content/uploads/Messier-than-Oil-Brief-1.pdf> accessed 12 April 2024

Chandler K 'Does Military AI Have Gender? Understanding bias and promoting ethical approaches in military applications of AI' (2021) UNIDIR <https://unidir.org/files/2021-12/UNIDIR_Does_Military_AI_Have_Gender.pdf> accessed 7 April 2024

Crootof R 'AI and the Actual IHL Accountability Gap' (2022) CIGI <https://www.cigionline.org/articles/ai-and-the-actual-ihl-accountability-gap/> accessed 12 October 2024

Davison N 'A legal perspective: Autonomous weapon systems under international humanitarian law' (2018) UNODA Occasional Papers No. 30 <https://www.globalgovernancewatch.org/library/doclib/20180208_ICRCAutonomousWeaponSystems.pdf> accessed 6 January 2023

Flournoy M A, Hailes A, Chefitz G 'Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems' (2020) WestExec Advisors <https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf> accessed 10 May 2024

Galindo L, Perset K, and Sheeka F (2021) 'An overview of national AI strategies and policies' (2021) OECD, Going Digital Toolkit Note, No. 14 <https://goingdigital.oecd.org/data/notes/No14_ToolkitNote_AIStrategies.pdf> accessed 21 July 2022

Geiß R 'The International-Law Dimension of Autonomous Weapons Systems' (2015) Friedrich Ebert Stiftung <https://library.fes.de/pdf-files/id/ipa/11673.pdf> accessed 28 October 2022

Gillard E 'Proportionality in the Conduct of Hostilities: The Incidental Harm Side of the Assessment' (2018) Chatham House <https://www.chathamhouse.org/sites/default/files/publications/research/2018-12-10-proportionality-conduct-hostilities-incidental-harm-gillard-final.pdf> accessed 7 May 2021

—— 'Seventy Years of the Geneva Conventions: What of the Future?' (2020) Chatham House <https://www.chathamhouse.org/2020/03/seventy-years-geneva-conventions/protection-medical-care-armed-conflict> accessed 20 November 2023

—— 'Sieges, the Law and Protecting Civilians' (2019) Chatham House <https://www.chathamhouse.org/sites/default/files/publications/research/2019-06-27-Sieges-Protecting-Civilians_0.pdf> accessed 19 February 2023

Goussac N and Pacholska M ‘Towards a Common Understanding of the Application of International Humanitarian Law to Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS)The Interpretation and Application of International Humanitarian Law in Relation to Lethal Autonomous Weapon Systems: Background paper on the views of States, scholars and other experts’ (2025), UNIDIR 10 <https://unidir.org/publication/the-interpretation-and-application-of-international-humanitarian-law-in-relation-to-lethal-autonomous-weapon-systems/> accessed 6 June 2025

Grand-Clément S ‘Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain’ (2023) UNIDIR <https://unidir.org/publication/artificial-intelligence-beyond-weapons-application-and-impact-of-ai-in-the-military-domain/> accessed 11 November 2024

Hartnett G S and others, ‘Operationally Relevant Artificial Training for Machine Learning: Improving the Performance of Automated Target Recognition Systems’ (2020) RAND <https://www.rand.org/pubs/research_reports/RRA683-1.html> accessed 17 June 2023

Hoffman W ‘AI and the Future of Cyber Competition’ (2021) CSET <https://cset.georgetown.edu/wp-content/uploads/CSET-AI-and-the-Future-of-Cyber-Competition-4.pdf> accessed 6 October 2022

Holland Michel A ‘Known Unknowns’ (2021) UNIDIR <https://unidir.org/known-unknowns> accessed 18 May 2021

Holland Michel A ‘Recalibrating assumptions on AI’ (2023) Chatham House <https://www.chathamhouse.org/2023/04/recalibrating-assumptions-ai/about-author> accessed 13 December 2023

Holland Michel A ‘The Black Box, Unlocked: Predictability and Understandability in Military AI’ (2020) UNIDIR, <https://unidir.org/files/2020-09/BlackBoxUnlocked.pdf> accessed 18 October 2024

Horowitz M and Scharre P ‘An Introduction to Autonomy in Weapon Systems’ (2015) CNAS <https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems> accessed 11 November 2024

—— ‘Meaningful Human Control in Weapon Systems: A Primer’ (2015) CNAS <https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf> accessed 22 April 2022

International Committee of the Red Cross ‘International Humanitarian Law and the challenges of contemporary armed conflicts’ (2011) Report for the 31st International Conference of the Red Cross and Red Crescent <https://www.icrc.org/en/doc/assets/files/red-cross-crescent-movement/31st-international-conference/31-int-conference-ihl-challenges-report-11-5-1-2-en.pdf> accessed 24 May 2020

ICRC and Geneva Academy ‘Expert Consultation Report on AI and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts’ (2024) ICRC <https://www.geneva-academy.ch/joomlatools-files/docman-

files/Artificial%20Intelligence%20And%20Related%20Technologies%20In%20Military%20Decision-Making.pdf> accessed 27 September 2024

Janjeva A, Harris A, Byrne J (2022) 'The Future of Open Source Intelligence for UK National Security', RUSI <https://rusi.org/explore-our-research/publications/occasional-papers/future-open-source-intelligence-uk-national-security> accessed 16 June 2022

Janjeva A and others 'The Rapid Rise of Generative AI: Assessing risks to safety and security' (2023) Centre for Emerging Technology and Security, The Alan Turing Institute, Research Report <https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai> accessed 20 December 2023

Jones E, Safak C 'Can algorithms ever make the grade?' Ada Lovelace Institute, 18 August 2020, <https://www.adalovelaceinstitute.org/blog/can-algorithms-ever-make-the-grade/> accessed 27 May 2022

Lewis L, Ilachinski A '(U) Leveraging AI to Mitigate Civilian Harm' (2022) Center for Naval Analyses <https://www.cna.org/reports/2022/02/Leveraging-AI-to-Mitigate-Civilian-Harm.pdf> accessed 12 April 2024

Lewis P and others, 'Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy' (2014) Chatham House <https://www.chathamhouse.org/2014/04/too-close-comfort-cases-near-nuclear-use-and-options-policy> accessed 15 November 2024

Lubell N, Pejic J, Simmons C 'Guidelines on Investigating Violations of International Humanitarian Law: Law, Policy, and Good Practice' (2019) ICRC and the Geneva Academy of International Humanitarian Law and Human Rights <https://www.icrc.org/en/document/guidelines-investigating-violations-ihl-law-policy-and-good-practice> accessed 24 February 2023

Melzer N 'Interpretative Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law' (2009) ICRC <https://shop.icrc.org/interpretive-guidance-on-the-notion-of-direct-participation-in-hostilities-under-international-humanitarian-law-pdf-en.html> accessed 13 July 2022

Nurkin T, Siegel J 'Battlefield Applications for Human-Machine Teaming: Demonstrating Value, Experimenting with New Capabilities, and Accelearting Adoption' (2023) Atlantic Council Scowcroft Center for Strategy and Security, <https://www.atlanticcouncil.org/wp-content/uploads/2023/08/Battlefield-Applications-for-HMT.pdf> accessed 13 April 2024

Pauwels E 'Artificial Intelligence and Data Capture Technologies in Violence and Conflict Prevention: Opportunities and Challenges for the International Community' (2020) Global Center on Cooperative Security <https://www.globalcenter.org/wp-content/uploads/2020/10/GCCS_AIData_PB_H.pdf> accessed 26 November 2020

Rao R, De Melo C, Krim H 'Synthetic Environments for Artificial Intelligence (AI) and Machine Learning (ML) in Multi-Domain Operations' (2021) ARL-TR-9198, DevCom: Army Research Laboratory <https://apps.dtic.mil/sti/trecms/pdf/AD1135395.pdf> accessed 21 March 2024

Roff H M ‘Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons: Briefing Paper for delegates at the Review Conference of the Convention on Certain Conventional Weapons’ (2016) Article 36 <http://www.article36.org/wp-content/uploads/2016/12/Control-or-Judgment_-Understanding-the-Scope.pdf> accessed 24 May 2020

Scharre P ‘Autonomous Weapons and Operational Risks’ (2016) CNAS <https://www.stopkillerrobots.org/wp-content/uploads/2021/09/CNAS_Autonomous-weapons-operational-risk.pdf> accessed 21 November 2023

Schwartz R and others ‘Towards a Standard for Identifying and Managing Bias in Artificial Intelligence’ (2022) NIST Special Publication 1270 <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf> accessed 19 November 2024

Slijper F ‘Slippery Slope: The arms industry and increasingly autonomous weapons’ (2019) PAX for Peace <https://www.paxforpeace.nl/publications/all-publications/slippery-slope> accessed 11 November 2019

Smallegange A and others ‘Big Data and Artificial Intelligence for Decision Making: Dutch Position Paper’ (2018) STO-MP-IST-160, NATO Science & Technology Organization <https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-IST-160/MP-IST-160-PP-1.pdf> accessed 20 May 2020

Sun N ‘Workplace AI in China’ (2024) Chatham House <https://www.chathamhouse.org/2024/07/workplace-ai-china/about-author> accessed 22 November 2024

Unal B ‘Cybersecurity of NATO’s Space-based Strategic Assets’ (2019) Chatham House <https://www.chathamhouse.org/2019/07/cybersecurity-natos-space-based-strategic-assets> accessed 6 October 2022

Unal B and others ‘Uncertainty and complexity in nuclear decision-making’ (2022) Chatham House <https://www.chathamhouse.org/2022/03/uncertainty-and-complexity-nuclear-decision-making/> accessed 28 October 2022

Villani C ‘Donner un Sens à l’Intelligence Artificielle : Pour une Stratégie Nationale et Européenne’ (2018) Mission Parlementaire <https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf> accessed 15 November 2019, 219-221

2.3. Working papers

Araujo R, Fort K, Guest O ‘Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges’ (2024) ArXiv <https://arxiv.org/abs/2410.09219> accessed 11 November 2024

Beagle T W Jr ‘Effects-Based Targeting: Another Empty Promise?’ (2001) Air University Press <https://www.jstor.org/stable/pdf/resrep13830.8.pdf> accessed 20 November 2023

Bommasani R and others ‘On the Opportunities and Risks of Foundation Models’ (2022) v3 ArXiv <https://arxiv.org/abs/2108.07258> accessed 14 November 2024

Bubeck S and others ‘Sparks of Artificial General Intelligence: Early experiments with GPT-4’ (2023) ArXiv <https://arxiv.org/abs/2303.12712> accessed 20 March 2024

De Coninck J ‘The EU’s Human Rights Responsibility Gap – Exhuming Human Rights Impunity of International Organizations’ (2024) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4706514> accessed 11 October 2024

Hagos M T, Kant S ‘Transfer Learning based Detection of Diabetic Retinopathy from Small Dataset’ (2019) ArXiv <https://arxiv.org/abs/1905.07203> accessed 26 May 2022

Milani P A ‘Autonomous Weapon Systems for the Land Domain’ (2020) Submission to the Concept for Robotics and Autonomous Systems 2040 <https://cove.army.gov.au/sites/default/files/autonomous_weapon_systems_for_the_land_domain.pdf> accessed 14 July 2023

Mitchell M and others (2020), ‘Diversity and Inclusion Metrics in Subset Selection’ (2020) ArXiv <https://arxiv.org/pdf/2002.03256.pdf> accessed 5 March 2021

Prabhakaran V and others ‘A Human Rights-Based Approach to Responsible AI’ (2022) ArXiv <https://arxiv.org/abs/2210.02667>

Rhue L ‘Racial Influence on Automated Perceptions of Emotions’ (2018) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765> accessed 27 November 2020

Sastry G and others, ‘Computing Power and the Governance of Artificial Intelligence’ (2024) ArXiv <https://arxiv.org/pdf/2402.08797.pdf> accessed 21 March 2024

Sukhbaatar S and others ‘Training Convolutional Networks with Noisy Labels’ (2014) ArXiv <https://arxiv.org/abs/1406.2080> accessed 21 July 2022

Ustun V and others ‘Adaptive Synthetic Characters for Military Trainig’ (2021) ArXiv <https://arxiv.org/abs/2101.02185> accessed 22 March 2024

2.4. Forthcoming articles and reports

Dear K and Pacholska M, ‘Machine Inhumanity? The Misguided Approach to Human Control in Advanced Weapon Systems’ (forthcoming)

3. Other secondary sources

3.1. Conference papers

Almuzaini A A ‘ABCinML: Anticipatory Bias Correction in Machine Learning Applications’ (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3533211> accessed 7 November 2022

Ballard D, Owsley L ‘Artificial intelligence in the helicopter cockpit of the future’ (IEEE/AIAA 10th Digital Avionics Systems Conference, Los Angeles, USA, 1991) <https://ieeexplore.ieee.org/abstract/document/177154> accessed 29 May 2023

Bender E M and others ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?’ (FAccT 3-10 March 2021, Virtual Event, Canada, 2021) <https://dl.acm.org/doi/10.1145/3442188.3445922> accessed 22 July 2022

Bergman A S, Diab M T ‘Towards Responsible Natural Language Annotation for the Varieties of Arabic’ (ACL Findings 2022) <https://arxiv.org/abs/2203.09597> accessed 7 July 2023

Birhane A and others, ‘The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models’ (FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency 2024) <https://dl.acm.org/doi/abs/10.1145/3630106.3658968à> accessed 17 November 2024

Bryant D, Howard A ‘A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity’ (AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019) <https://dl.acm.org/doi/abs/10.1145/3306618.3314284> accessed 5 July 2023

Carter B and others ‘Overinterpretation reveals image classification model pathologies’ (Conference on Neural Information Processing Systems, NeurIPS 2021) <https://proceedings.neurips.cc/paper/2021/file/8217bb4e7fa0541e0f5e04fea764ab91-Paper.pdf> accessed 21 February 2023

Chapman A and others ‘A Data-driven analysis of the interplay between Criminological theory and predictive policing algorithms’ (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3533071> accessed 28 October 2022

Chen T and others ‘Would Deep Generative Models Amplify Bias in Future Models?’ (IEEE / CVF Computer Vision and Pattern Recognition Conference, Seattle, June 2024) <https://arxiv.org/abs/2404.03242> accessed 12 April 2024

Cho W I and others ‘Towards Cross-Lingual Generalization of Translation Gender Bias’ (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445907> accessed 2 September 2022

Chudzikiewicz J, Furtak J, Zielinski Z ‘Fault-tolerant techniques for the Internet of Military Things’ (IEEE 2nd World Forum on Internet of Things (WF-IoT) Milan, Italy, 2015) <https://ieeexplore.ieee.org/abstract/document/7389104> accessed 20 June 2023

Dabbish L and others, 'Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository' (CSCW '12, Seattle, Washington, 11-15 February 2012) <http://www.jsntsay.com/publications/dabbish-cscw2012.pdf> accessed 15 September 2022

Dang N and De Causmaecker P 'Analysis of algorithm components and parameter: Some case studies' (Learning and Intelligent Optimization: 12th International Conference (LION 12) Kalamata, Greece, 10-15 June 2018) <https://core.ac.uk/download/pdf/196582687.pdf> accessed 16 September 2022

Danks D and London A J 'Algorithmic Bias in Autonomous Systems' (Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, 2017) <https://www.ijcai.org/Proceedings/2017/0654.pdf> accessed 24 May 2020

Dästner K and others 'Classification of Military Aircraft in Real-time Radar Systems based on Supervised Machine Learning with Labelled ADS-B Data' (IEEE Conference, Sensor Data Fusion: Trends, Solutions, Applications, Bonn, 2018) <https://ieeexplore.ieee.org/abstract/document/8547077> accessed 22 March 2024

Ding Y and others 'Impact of Annotator Demographics on Sentiment Dataset Labeling' (Proceedings of the ACM on Human-Computer Interaction CSCW2 Vol. 6, 2022) <https://dl.acm.org/doi/abs/10.1145/3555632> accessed 12 April 2024

Geiss R 'Autonomous Weapons Systems: Risk Management and State Responsibility' (3rd CCW meeting of experts on lethal automous weapons systems (LAWS), Geneva, 11-15 April 2016) <https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/Geiss-CCW-Website.pdf> accessed 14 October 2022

Ghimire A and Edwards J 'From Guidelines to Governance: A Study of AI Policies in Education' (International Conference on Artificial Intelligence in Education, Recife, Brazil, 8-12 July 2024) <https://link.springer.com/book/10.1007/978-3-031-64312-5> accessed 11 November 2024

Gilpin L H and others 'Explaining Explanations to Society' (Proceedings of the NeurIPS 2018 Workshop on Ethical, Social and Governance Issues in AI, Montréal, Québec, Canada, 2018), <https://arxiv.org/abs/1901.06560> accessed 23 February 2023

Globus-Harris I, Kearns M, Roth A 'An Algorithmic Framework for Bias Bounties' (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3533172>

Han Q, Snaiduf D 'Comparison of Deep Learning Technologies in Legal Document Classification' (IEEE International Conference on Big Data, Orlando, 2021) <https://ieeexplore.ieee.org/document/9671486> accessed 7 April 2024

Hanratty K P 'Artificial (military) intelligence: enabling decision dominance through machine learning' (SPIE Defense + Commercial Sensing, Orlando, 2023) <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12538/125380J/Artificial-military-intelligence-enabling-decision-dominance-through-machine-learning/10.1117/12.2663413.short#_=_> accessed 22 March 2024

Haque A A ‘Killing in the Fog of War’ (Proceedings of the ASIL Annual Meeting , Volume 106, 2012) <https://www.cambridge.org/core/journals/proceedings-of-the-asil-annual-meeting/article/abs/killing-in-the-fog-of-war/209F6440BEC661F3DE6F813B07BBED10> accessed 9 December 2022

He L and others ‘Artificial intelligence technology in battlefield situation awareness’ (VESEP22, 2022) <https://www.e3s-conferences.org/articles/e3sconf/pdf/2022/27/e3sconf_vesep2022_01063.pdf> accessed 21 November 2023

Holtzen S and others ‘Inferring human intent from video by sampling hierarchical plans’ (IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2016) <https://ieeexplore.ieee.org/abstract/document/7759242> accessed 4 February 2022

Jacobs A Z, Wallach H ‘Measurement and Fairness’ (FaccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada, 3-10 March 2021) <https://dl.acm.org/doi/abs/10.1145/3442188.3445901> accessed 11 November 2022

Jo E, Gebru T Eun ‘Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning’ (Conference on Fairness, Accountability and Transparency (FAT*), 27-30 January 2020) <https://dl.acm.org/doi/pdf/10.1145/3351095.3372829> accessed 17 June 2022

Kandybko N and others ‘Application of artificial intelligence technologies in the context of military security and civil infrastructure protection’ (AIP Conference Proceedings, Vol. 2476 (1) 2023) <https://pubs.aip.org/aip/acp/article/2476/1/030019/2891077> accessed 12 April 2024

Khan Z, Fu Y ‘One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision’ (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021) <https://dl.acm.org/doi/abs/10.1145/3442188.3445920> accessed 19 April 2023

Klug D and others ‘Trick and Please. A Mixed-Method Study on User Assumptions About the TikTok Algorithm’ (WebSci '21: Proceedings of the 13th ACM Web Science Conference 2021) <https://dl.acm.org/doi/abs/10.1145/3447535.3462512> accessed 17 November 2022

Lee J and others, ‘Revisiting the Importance of Amplifying Bias for Debiasing’ (Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington DC, USA, 7-14 February 2023) <https://ojs.aaai.org/index.php/AAAI/article/view/26748> accessed 12 October 2024

Martin J L ‘Spoken Corporate Data, Automatic Speech Recognition, and Bias Against African American Language: The case of Habitual ‘Be’’ (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445893> accessed 2 September 2022

Moghadam M G ‘Machine learning-assisted performance testing’ (Proceedings of the 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the

Foundations of Software Engineering (ESEC/FSE), New York, 2019) <https://doi.org/10.1145/3338906.3342484> accessed 17 October 2024

O'Donnell M and others 'Roadmap to Implement Artificial Intelligence in Course of Action Development & Effect of Weather Variables on UH-60 Performance' (Proceedings of the Annual General Donald R. Keith Memorial Conference, West Point, USA, 2021) <http://www.ieworldconference.org/content/WP2021/Papers/GDRKMCC_21_61.pdf> accessed 29 May 2023

Pagan N and others, 'A classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems' (3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Boston, October 2023) <https://dl.acm.org/doi/10.1145/3617694.3623227> acccessed 12 April 2024

Park J S and others 'Designingg an Online Infrastructure for Collecting AI Data From People with Disabilities' (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445870> accessed 2 September 2022

Raji I D and others 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing' (FAT* '20, Barcelona, Spain, January 27–30, 2020) <https://dl.acm.org/doi/pdf/10.1145/3351095.3372873> accessed 10 October 2022

Roberts A, Venables A 'The Role of Artificial Intelligence in Kinetic Targeting from the Perspective of International Humanitarian Law' (13th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2021) <https://ieeexplore.ieee.org/abstract/document/9468301> accessed 9 November 2023

Sambasivan N and others 'Re-imagining Algorithmic Fairness in India and Beyond' (FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021) <https://dl.acm.org/doi/10.1145/3442188.3445896> accessed 2 September 2022

Sassòli M 'Legitimate Targets of Attacks under International Humanitarian Law' (Background Paper prepared for the Informal High-Level Expert Meeting on the Reaffirmation and Development of International Humanitarian Law, Cambridge, January 27-29, 2003) <http://humanrightsvoices.org/assets/attachments/documents/Session1.pdf> accessed 16 July 2023

Schmitt M N '"Attack" as a Term of Art in International Law: The Cyber Operations Context' (4th International Conference on Cyber Conflict, 2012) <https://ccdcoe.org/uploads/2012/01/5_2_Schmitt_AttackAsATermOfArt.pdf> accessed 26 June 2023

Singer T V P 'Update to revolving door 2.0: The extension of the period for direct participation in hostilities due to autonomous cyber weapons' (9th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2017) <https://ieeexplore.ieee.org/abstract/document/8240332> accessed 16 July 2023

Speith T ‘A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods’ (FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022) <https://dl.acm.org/doi/10.1145/3531146.3534639> accessed 23 February 2023

Sug H ‘Performance of Machine Learning Algorithms and Diversity in Data’ (MATEC Web of Conferences 210, GSCC, 2018) <https://www.matec-conferences.org/articles/matecconf/pdf/2018/69/matecconf_cscc2018_04019.pdf> accessed 5 March 2021

Xu F and others ‘Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges’ (Natural Language Processing and Chinese Computing 2019) <https://link.springer.com/chapter/10.1007/978-3-030-32236-6_51> accessed 23 February 2023

Yasin J N and others ‘Navigation of Autonomous Swarm of Drones Using Translational Coordinates’ (Advances in Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness, PAAMS Collection 18th International Conference, L'Aquila, Italy, October 7–9, 2020) <https://link.springer.com/chapter/10.1007/978-3-030-49778-1_28> accessed 17 November 2022

Yucer S and others ‘Measuring Hidden Bias within Face Recognition via Racial Phenotypes’ (IEEE/CVF Winter Conference on Applications of Computer Vision, 2022) <https://journals.uc.edu/index.php/isrs/article/view/8006> accessed 12 April 2024

Zhang Y, Liao Q V, Bellamy R K E (2020), ‘Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making’ (FAT* ’20: Proceedings of the Conference on Fairness, Accountability, and Transparency 2020) <https://dl.acm.org/doi/abs/10.1145/3351095.3372852> accessed 19 October 2023

3.2. Theses

Aktas S ‘Applying Deep Learning Methods to Identify Targets in Synthetic Aperture Radar Images’ (Thesis, Naval Postgraduate School 2020) <https://apps.dtic.mil/sti/trecms/pdf/AD1126746.pdf> accessed 7 April 2024

Beckett G P ‘Leveraging Artificial Intelligence and Automatic Target Recognition to Accelerate Deliberate Targeting’ (Thesis, Air War College, Air University 2020) <https://apps.dtic.mil/sti/pdfs/AD1107486.pdf> accessed 29 June 2023

Grooms G B ‘Artificial Intelligence Applications for Automated Battle Management Aids in Future Military Endeavors’ (Thesis, Naval Postgraduate School 2019) <https://apps.dtic.mil/sti/citations/AD1080249> accessed 27 February 2023

Haddix, R G II, ‘An Immersive Synthetic Environment for Observation and Interaction with a Large Volume of Interest’ (Thesis, Air Force Institute of Technology 1993) <https://apps.dtic.mil/sti/citations/ADA262547> accessed 30 May 2023

Montgomery C S ‘Trust in the Machine: AI, Autonomy, and Military Decision Making with Lethal Consquences’ (Report, Certificate Program in Ethics and Emerging Military Technology, U.S. Naval War College Newport 2019) <https://apps.dtic.mil/sti/pdfs/AD1121116.pdf> accessed 12 April 2024

Silvestri A ‘The Revolving Door of Modern Warfare: Civilian Direct Participation in Hostilities’ (Thesis, The University of Western Australia 2022) <https://api.research-repository.uwa.edu.au/ws/portalfiles/portal/223429021/THESIS_DOCTOR_OF_PHILOSOPHY_SILVESTRI_Alessandro_2022_Part_1.pdf> accessed 16 July 2023

3.3. Websites, blogs and magazines

—— ‘A-Size Autonomous Underwater Vehicles’ (*Lockheed Martin*) <https://www.lockheedmartin.com/en-us/products/a-size-autonomous-underwater-vehicles.html> accessed 11 November 2019

—— ‘About CJTF-OIR’ (*Operation Inherent Resolve*) <https://www.inherentresolve.mil/About-CJTF-OIR/> accessed 18 November 2020

—— ‘About DIANA’ (*NATO*) <https://www.diana.nato.int/about-diana.html> accessed 2 February 2024

—— ‘About Us’ (*Campaign to Stop Killer Robots*) <https://www.stopkillerrobots.org/about-us/> accessed 2 September 2022

—— ‘AI for Human Rights’ (*The Alan Turing Institute*) <https://www.turing.ac.uk/ai-human-rights> accessed 3 September 2022

—— ‘AI Next Campaign’ (*DARPA*) <https://www.darpa.mil/work-with-us/ai-next-campaign> accessed 9 January 2020

—— ‘AIP for Defense’ (*Palantir*) <https://www.palantir.com/platforms/aip/> accessed 9 June 2023

—— ‘AlphaDogFight Trials Go Virtual for Final Event’ (*DARPA*) <https://www.darpa.mil/news-events/2020-08-07> accessed 29 September 2022

—— ‘Build AI responsibly to benefit humanity’ (*Google DeepMind*) <https://deepmind.google/about/> accessed 21 November 2024

—— ‘Corporate human rights due diligence – identifying and leveraging emerging practices’ (*OHCHR*) <https://www.ohchr.org/en/special-procedures/wg-business/corporate-human-rights-due-diligence-identifying-and-leveraging-emerging-practices> accessed 7 November 2022

—— ‘Cyber defence’ (*NATO*) <https://www.nato.int/cps/en/natohq/topics_78170.htm> accessed 26 February 2023

—— ‘Data’ (*Cambridge Advanced Learner’s Dictionary & Thesaurus*), <https://dictionary.cambridge.org/dictionary/english/data> accessed 18 November 2020

—— ‘Dataiku Selected As Platform For NATO’s Allied Command Transformation Focusing On AI Projects’ (*Dataiku*, 14 February 2020) <https://pages.dataiku.com/nato> accessed 20 May 2020

—— ‘Defence and security’ (*The Alan Turing Institute*) <https://www.turing.ac.uk/research/research-programmes/defence-and-security> accessed 29 September 2022

—— ‘DOT&E Annual Reports’ (*The Office of the Director, Operational Test and Evaluation*) <https://www.dote.osd.mil/annualreport/> accessed 22 September 2022

—— ‘Federal Republic of Yugoslavia (FRY) /NATO: “Collateral damage” or unlawful killings? Violations of the Laws of War by NATO during Operation Allied Force’ (*Amnesty International*, 5 June 2000) <https://www.amnesty.org/en/documents/eur70/018/2000/en/> accessed 29 April 2022

—— ‘How does ICRC use forensic science for humanitarian purposes?’ (*ICRC*, 19 May 2022) <https://www.icrc.org/en/document/how-does-icrc-use-forensic-science-humanitarian-purposes> accessed 26 February 2023

—— ‘How to manage complexity and realize the value of big data’ (*IBM*, 28 May 2020) <https://www.ibm.com/blog/how-to-manage-complexity-and-realize-the-value-of-big-data/> accessed 16 March 2023

—— ‘IDF Response as Sent to the Guardian’ (*IDF*, 3 April 2024) <https://www.idf.il/189654> accessed 13 April 2024

—— ‘Implement Face Recognition on Autonomous sUAS for Identification and Intelligence-Gathering’ Solicitation Number X21.1, Contract FA864922P0006 (*Department of Defense*) <https://www.sbir.gov/node/2217727> accessed 27 June 2023

—— ‘Internet of Battlefield Things (IoBT) CRA’ (*DEVCOM*) <https://www.arl.army.mil/business/collaborative-alliances/current-cras/iobt-cra/> accessed 27 May 2022

—— ‘Israel, The Targeted Killings Case’ (*ICRC*) <https://casebook.icrc.org/case-study/israel-targeted-killings-case> accessed 9 November 2023

—— ‘LAWS & War Crimes Project: Algorithmic Warfare and Accountability’, <https://lawsandwarcrimesdotcom.wordpress.com/> accessed 6 August 2022

—— ‘Lethal Autonomous Weapon Systems (LAWS)’ (*United Nations Office for Disarmament Affairs*) <https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/> accessed 17 November 2022

—— ‘Losing Humanity: The Case against Killer Robots’ (*Human Rights Watch*, 19 November 2012) <https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots> accessed 18 November 2022

—— ‘Machine learning algorithms promise better situational awareness’ (*US Army*, 22 June 2020) <https://www.army.mil/article/236647/machine_learning_algorithms_promise_better_situational_awareness> accessed 21 November 2023

—— ‘Means of warfare’ (*ICRC*) <https://casebook.icrc.org/glossary/means-warfare> accessed 6 October 2022

—— ‘Mind the Gap: The Lack of Accountability for Killer Robots’ (*Human Rights Watch*, 9 April 2015) <https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots> accessed 14 October 2022

—— ‘Modelling & Simulation for Autonomous Systems Conference’ (*NATO M&S COE*) <https://www.mscoe.org/event/mesas-2023/about-us/> accessed 7 April 2023

—— ‘NATO Recognises Cyberspace as a ‘Domain of Operations’ at Warsaw Summit’ (*CCDCOE*) <https://ccdcoe.org/incyder-articles/nato-recognises-cyberspace-as-a-domain-of-operations-at-warsaw-summit/> accessed 6 October 2022

—— ‘New Advanced Technologies Improve Situational Awareness for Land Forces’ (*THALES*, 12 September 2022) <https://www.thalesgroup.com/en/united-states/magazine/new-advanced-technologies-improve-situational-awareness-land-forces> accessed 21 November 2023

—— ‘OpenAI’s approach to AI and national security’ (*OpenAI*, 24 October 2024) <https://openai.com/global-affairs/openais-approach-to-ai-and-national-security/> accessed 18 November 2024

—— ‘Our Research’ (*DARPA*) <https://www.darpa.mil/our-research> accessed 29 September 2022

—— ‘PAI’s Guidance for Safe Foundation Model Deployment: A Framework for Collective Action’ (Partnership on AI) <https://partnershiponai.org/modeldeployment/> accessed 23 January 2024

—— ‘Principle of Distinction’ (*ICRC*) <https://casebook.icrc.org/law/principle-distinction> accessed 27 November 2020

—— ‘Public Eye: an open-source crowd monitoring solution’ (*City of Amsterdam*) <https://www.amsterdam.nl/innovation/mobility/public-eye/> accessed 29 May 2023

—— ‘Rule 6: Civilians’ Loss of Protection from Attack’ (ICRC IHL Databases) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule6> accessed 3 June 2025

—— ‘Rule 8: Definition of Military Objectives’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule8> accessed 29 September 2024

—— ‘Rule 14: Proportionality in Attack’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule14> accessed 29 September 2023

—— ‘Rule 15: Principle of Precautions in Attack’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule15> accessed 21 May 2021

—— ‘Rule 17: Choice of Means and Methods of Warfare’ (*ICRC IHL Databases)* <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule17> accessed 23 September 2022

—— ‘Rule 22: Principle of Precautions against the Effects of Attacks Rule 22. The parties to the conflict must take all feasible precautions to protect the civilian population and civilian objects under their control against the effects of attacks’ (*ICRC IHL Database*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule22> accessed 20 November 2023

—— ‘Rule 70: Weapons of a Nature to Cause Superfluous Injury or Unnecessary Suffering’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule70> accessed 14 October 2024

—— ‘Rule 71: Weapons That Are by Nature Indiscriminate’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule71> accessed 23 September 2022

—— ‘Rule 89: Violence to Life’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule89> accessed 29 September 2024

—— ‘Rule 97: Human Shields’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/Rule97> accessed 22 November 2024

—— ‘Rule 149: Responsibility for violations of International Humanitarian Law’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule149> accessed 20 October 2022

—— ‘Rule 153: Command Responsibility for Failure to Prevent, Repress or Report War Crimes’ (*ICRC IHL Databases*) <https://ihl-databases.icrc.org/en/customary-ihl/v1/rule153> accessed 12 October 2024

—— ‘Science Communication’ (*Science Europe*) <https://scienceeurope.org/our-priorities/science-communication/> accessed 19 October 2024

—— ‘Summary of NATO’s revised Artificial Intelligence (AI) Strategy’ (*NATO*, 10 July 2024) <https://www.nato.int/cps/en/natohq/official_texts_227237.htm> accessed 12 November 2024

—— ‘Surrender’ (*ICRC*) <https://casebook.icrc.org/glossary/surrender> accessed 13 July 2022

—— ‘The Convention on Certain Conventional Weapons’ (*United Nations Office for Disarmament Affairs*) <https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/> accessed 23 September 2022

—— ‘The IDF’s Use of Data Technologies in Intelligence Processing’ (*IDF*, 18 June 2024) <https://www.idf.il/en/mini-sites/idf-press-releases-regarding-the-hamas-israel-war/june-24-pr/the-idf-s-use-of-data-technologies-in-intelligence-processing/> accessed 10 July 2024

—— ‘The Toolkit for Responsible Artificial Intelligence Innovation in Law Enforcement’ (*UNICRI*) <https://unicri.it/topics/Toolkit-Responsible-AI-for-Law-Enforcement-INTERPOL-UNICRI> accessed 11 November 2024

—— ‘Uncertainty’ (*Cambridge Dictionary*) <https://dictionary.cambridge.org/dictionary/english/uncertainty> accessed 2 February 2023

—— ‘Unmanned Aircraft Systems’ (*Airbus*) <https://www.airbus.com/defence/uav.html> accessed 5 December 2019

—— ‘What Alan Tuirng means to us’ (*The Alan Turing Institute*, 23 June 2019) <https://www.turing.ac.uk/blog/what-alan-turing-means-us> accessed 29 May 2023

—— ‘What is a synthetic environment?’ (*Improbable Defence*) <https://defence.improbable.io/what-is-a-synthetic-environment/> accessed 29 May 2023

—— ‘What is unsupervised learning?’ (IBM), <https://www.ibm.com/topics/unsupervised-learning> accessed 22 March 2024

Ackerman E ‘Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms’ (*IEEE Spectrum*, 4 August 2017) <https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms> accessed 4 May 2021

Afina Y ‘Rage Against the Algorithm: the Risks of Overestimating Military Artificial Intelligence’ (*Chatham House*, 27 August 2020) <https://www.chathamhouse.org/2020/08/rage-against-algorithm-risks-overestimating-military-artificial-intelligence> accessed 13 March 2024

Afina Y ‘Intelligence is dead: long live Artificial Intelligence’ (*Chatham House*, 14 July 2022) <https://www.chathamhouse.org/2022/07/intelligence-dead-long-live-artificial-intelligence> accessed 5 February 2024

Afina Y, Inverarity C ‘Predictions and Policymaking : Complex Modelling Beyond COVID-19’ (Chatham House, Expert Comment, 2020) <https://www.chathamhouse.org/expert/comment/predictions-and-policymaking-complex-modelling-beyond-covid-19> accessed 24 May 2020

Agarwal S ‘How a big blue van from 1986 paved the way for self-driving cars’ (*Digital Trends*, 6 February 2022) <https://www.digitaltrends.com/cars/first-self-driving-car-ai-navlab-history/> accessed 20 January 2023

Alayrac J and others, 'Tackling multiple tasks with a single visual language model' (*Google DeepMind*, 28 April 2022) <https://www.deepmind.com/blog/tackling-multiple-tasks-with-a-single-visual-language-model> accessed 15 September 2022

Altman S 'Planning for AGI and beyond' (*OpenAI*, 24 February 2023) <https://openai.com/blog/planning-for-agi-and-beyond> acccessed 23 March 2023

Bajema N 'Will AI Steal Submarines' Stealth?' (*IEEE Spectrum*, 16 July 2022) <https://spectrum.ieee.org/nuclear-submarine> accessed 6 October 2022

Blazek P J 'Why We Will Never Open Deep Learning's Black Box' (*Towards Data Science*, 2 March 2022) <https://towardsdatascience.com/why-we-will-never-open-deep-learnings-black-box-4c27cd335118> accessed 26 February 2023

Bode I 'Falling under the radar: the problem of algorithmic bias and military applications of AI' (*Humanitarian Law & Policy*, 14 March 2024) <https://blogs.icrc.org/law-and-policy/2024/03/14/falling-under-the-radar-the-problem-of-algorithmic-bias-and-military-applications-of-ai/> accessed 9 April 2024

Bode I, Bhila I 'The problem of algorithmic bias in AI-based military decision support systems' (*Humanitarian Law & Policy*, 3 September 2024) <https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based-military-decision-support-systems/> accessed 28 September 2024

Bowen B E 'The Application of Force and Strategy in Sun Tzu and Clausewitz' (*E-International Relations*, 16 December 2010) <https://www.e-ir.info/2010/12/16/the-application-of-force-and-strategy-in-sun-tzu-and-clausewitz/> accessed 2 February 2023

Burt A 'How to Fight Discrimination in AI' (*Harvard Business Review*, 28 August 2020) <https://hbr.org/2020/08/how-to-fight-discrimination-in-ai> accessed 23 April 2021

Cambria E, White B 'Jumping NLP Curves: A Review of Natural Language Processing Research' (*IEEE Computational Intelligence Magazine*, 11 April 2014) <http://sentic.net/jumping-nlp-curves.pdf> accessed 22 July 2022

Cappel L M 'What is the difference between Deep Learning and Reinforcement Learning?' (*Big Data Made Simple*, 2019) <https://bigdata-madesimple.com/what-is-the-difference-between-deep-learning-and-reinforcement-learning/> accessed 24 May 2020

Card D 'The "black box" metaphor in machine learning' (*Towards Data Science*, 5 July 2017), <https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0> accessed 11 January 2020

Castelvecchi D 'Can we open the black box of AI?' (*Nature*, 5 October 2016) <https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731> accessed 19 January 2023

Čevora G 'How Discrimination occurs in Data Analytics and Machine Learning: Proxy Variables' (*Towards Data Science*, 5 February 2020) <https://towardsdatascience.com/how-

discrimination-occurs-in-data-analytics-and-machine-learning-proxy-variables-7c22ff20792> accessed 16 April 2021

Cummings M 'Lethal Autonomous Weapons: Meaningful human control or meaningful human certification?' (*Technology and Society*, 23 December 2019) <https://technologyandsociety.org/lethal-autonomous-weapons-meaningful-human-control-or-meaningful-human-certification/> accessed 23 November 2024

Defense Digital Service 'Code.mil: An Open Source Initiative at the Pentagon' (*Medium*, 13 March 2017) <https://medium.com/@DefenseDigitalService/code-mil-an-open-source-initiative-at-the-pentagon-5ae4986b79bc#.i78how76u> accessed 29 September 2022

Droege C 'War and what we make of the law' (*Humanitarian Law & Policy*, 18 July 2024) <https://blogs.icrc.org/law-and-policy/2024/07/18/war-and-what-we-make-of-the-law/> accessed 18 October 2024

Ekelhof M 'Autonomous weapons: Operationalizing meaningful human control' (*Humanitarian Law & Policy*, 15 August 2018) <https://blogs.icrc.org/law-and-policy/2018/08/15/autonomous-weapons-operationalizing-meaningful-human-control/> accessed 22 April 2022.

Gisel L and others 'Urban warfare: an age-old problem in need of new solutions' (*Humanitarian Law & Policy*, 27 April 2021) <https://blogs.icrc.org/law-and-policy/2021/04/27/urban-warfare/> accessed 2 February 2023

Goyal A 'Benefits of Land Registry Digitization' (*World Bank Blogs*, 17 April 2012) <https://blogs.worldbank.org/en/digital-development/benefits-of-land-registry-digitization> accessed 9 July 2024

Greipl A R 'Artificial intelligence in urban warfare: opportunities to enhance the protection of civilians?' (2023) 61(2) *The Military Law and the Law of War Review*, <https://doi.org/10.4337/mllwr.2023.02.03> accessed 12 April 2024

Hayashi N 'Honest Errors, the Rendulic Rule, and Modern Combat Decision-Making' (*Articles of War*, 24 October 2023) <https://lieber.westpoint.edu/honest-errors-rendulic-rule-modern-combat-decision-making/> accessed 19 October 2024

Hefron R 'Air Combat Evolution (ACE)' (*DARPA*) <https://www.darpa.mil/program/air-combat-evolution> accessed 29 September 2022

Heyns C, Casey-Maslen S, Probert T 'The Definition of an "Attack" under the Law of Armed Conflict' (*Articles of War*, 3 November 2020) <https://lieber.westpoint.edu/definition-attack-law-of-armed-conflict-protection/> accessed 26 June 2023

Hsu J 'Army Trains AI to Identify Faces in the Dark: The U.S. Army Research Laboratory has developed a dataset of faces to train facial recognition that works in darkness' (*IEEE Spectrum*, 9 March 2021) <https://spectrum.ieee.org/army-trains-ai-to-identify-faces-in-the-dark#toggle-gdpr> accessed 17 June 2022

Koehrsen W ‘Modeling: Teaching a Machine Learning Algorithm to Deliver Business Value’ (*Medium*, 15 November 2018), <https://towardsdatascience.com/modeling-teaching-a-machine-learning-algorithm-to-deliver-business-value-ad0205ca4c86> accessed 24 May 2020

Lewis D A ‘Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider’ (*Humanitarian Law & Policy*, 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 21 November 2023

Marguiles P ‘Centcom Report on Kunduz Hospital Attack: Accounting for a Tragedy of Errors’ (*Lawfare*, 2 May 2016) <https://www.lawfaremedia.org/article/centcom-report-kunduz-hospital-attack-accounting-tragedy-errors> accessed 20 November 2023

Marr B ‘How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read’ (*Forbes*, 21 May 2018) <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=471cf14c60ba> accessed 16 March 2023

Milanovic M ‘Mistakes of Fact When Using Lethal Force in International Law: Part I’ (*EJIL: Talk!*, 14 January 2020) <https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-i/> accessed 21 November 2023

Olejnik L ‘Smartphones Blur the Line Between Civilian and Combatant’ (*WIRED*, 6 June 2022) <https://www.wired.com/story/smartphones-ukraine-civilian-combatant/> accessed 22 July 2022

Palmer M ‘AI’s crucial role onboard drones for ISR capability improvement’ (*Shield AI*, September 2022) <https://sentientvision.com/ais-crucial-role-onboard-drones-for-isr-capability-improvement/> accessed 29 June 2023

Parikh D ‘Learning Paradigms in Machine Learning’ (*Medium*, 7 July 2018) <https://medium.datadriveninvestor.com/learning-paradigms-in-machine-learning-146ebf8b5943> accessed 21 November 2024

Pellandra A, Henningsen G ‘Predicting refugee flows with big data: a new opportunity or a pipe dream?’ (*UNHCR Blogs*, 6 January 2022) <https://www.unhcr.org/blogs/predicting-refugee-flows-with-big-data-a-new-opportunity-or-a-pipe-dream/> accessed 10 July 2024

Postma F ‘US Soldiers Expose Nuclear Weapons Secrets Via Flashcard Apps’ (*Bellingcat*, 28 May 2021) <https://www.bellingcat.com/news/2021/05/28/us-soldiers-expose-nuclear-weapons-secrets-via-flashcard-apps/> accessed 22 November 2024

Ramchandani P ‘Random Forests and the Bias-Variance Tradeoff’ (*Medium*, 10 October 2018) <https://towardsdatascience.com/random-forests-and-the-bias-variance-tradeoff-3b77fee339b4?gi=6cdd6a5535b8> accessed 28 September 2024

Reed S and others ‘A Generalist Agent’ (*Google DeepMind*, 12 May 2022) <https://deepmind.google/discover/blog/a-generalist-agent/> accessed 11 November 2024

Russell B ‘Urban Reconnaissance through Supervised Autonomy (URSA)’ (*Defense Advanced Research Projects Agency*) <https://www.darpa.mil/program/urban-reconnaissance-through-supervised-autonomy> (accessed 1 December 2020)

Saltini A ‘To avoid nuclear instability, a moratorium on integrating AI into nuclear decision-making is urgently needed: The NPT PrepCom can serve as a springboard’ (*ELN*, 28 July 2023) <https://www.europeanleadershipnetwork.org/commentary/to-avoid-nuclear-instability-a-moratorium-on-integrating-ai-into-nuclear-decision-making-is-urgently-needed-the-npt-prepcom-can-serve-as-a-springboard/> accessed 21 March 2024

Singer G ‘Advancing Machine Intelligence: Why Context Is Everything’ (*Towards Data Science*, 10 May 2022, <https://towardsdatascience.com/advancing-machine-intelligence-why-context-is-everything-4bde90fb2d79> accessed 22 July 2022

Sonnenburg J & Sonnenburg E ‘Gut Feelings–the "Second Brain" in Our Gastrointestinal Systems [Excerpt]’ (*Scientific American*, 1 May 2015) <https://www.scientificamerican.com/article/gut-feelings-the-second-brain-in-our-gastrointestinal-systems-excerpt/> accessed 19 February 2023

Speier R H ‘Hypersonic Missiles: A New Proliferation Challenge’ (*The RAND Blog*, 29 March 2018) <https://www.rand.org/blog/2018/03/hypersonic-missiles-a-new-proliferation-challenge.html> accessed 18 October 2023

Stanley-Lockman Z, Christie E H ‘An Artificial Intelligence Strategy for NATO’ (*NATO Review*, 25 October 2021) <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html> accessed 2 February 2024

Stewart R, Hinds G ‘Algorithms of war: The use of artificial intelligence in decision making in armed conflict’ (*Humanitarian Law & Policy*, 24 October 2023) <https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict/> accessed 12 April 2024

Sullivan D ‘An overview of our rater guidelines for Search’ (*Google The* Keyword, 19 October 2021) <https://blog.google/products/search/overview-our-rater-guidelines-search/> accessed 17 November 2022

Ton-That H ‘The Myth of Facial Recognition Bias’ (*ClearviewAI*, 28 November 2022) <https://www.clearview.ai/post/the-myth-of-facial-recognition-bias> accessed 17 November 2024

Veisdal J ‘The Birthplace of AI’ (*Medium*, 12 September 2019) <https://medium.com/cantors-paradise/the-birthplace-of-ai-9ab7d4e5fb00> accessed 24 May 2020

Whetham D, Payne K ‘AI: In Defence of Uncertainty’ (*Defence-In-Depth*, 9 December 2019) <https://defenceindepth.co/2019/12/09/ai-in-defence-of-uncertainty/> accessed 19 February 2023

3.4. News

—— ‘СБУ встановила, що російські командири вторглися в Україну, керуючись картами з минулого століття’ [SBU found that Russian Commanders Invaded Ukraine using Maps from the Last Century] (*Security Service of Ukraine*, 18 July 2022) <https://ssu.gov.ua/novyny/sbu-vstanovyla-shcho-rosiiski-komandyry-vtorhlasia-v-ukrainu-keruiuchys-kartamy-z-mynuloho-stolittia> accessed 9 June 2023

—— ‘Gaza tower housing Al Jazeera office destroyed by Israeli attack’ (*Al Jazeera*, 15 May 2021), <https://www.aljazeera.com/news/2021/5/15/building-housing-al-jazeeera-office-in-gaza-hit-by-israeli-strike> accessed 21 May 2021

—— ‘IBM used Flickr photos for facial-recognition project’ (*BBC*, 13 March 2019) <https://www.bbc.co.uk/news/technology-47555216> accessed 21 May 2021

—— ‘Israel’s Shin Bet spy service uses generative AI to thwart threats’ (*Reuters*, 27 June 2023) <https://www.reuters.com/technology/israels-shin-bet-spy-service-uses-generative-ai-thwart-threats-2023-06-27/> accessed 28 June 2023

—— ‘Kunduz bombing: US attacked MSF clinic 'in error'’ (*BBC News*, 25 November 2015) <https://www.bbc.co.uk/news/world-asia-34925237> accessed 20 November 2023

—— ‘Russian military received outdated maps of Ukraine prior to invasion’ (*Militarnyi*, 18 July 2022) <https://mil.in.ua/en/news/russian-military-received-outdated-maps-of-ukraine-prior-to-invasion/> accessed 9 June 2023

—— ‘Sébastien Lecornu lance la stratégie ministérielle sur l’intelligence artificielle’ (*Ministère des Armées*, 8 March 2024) <https://www.defense.gouv.fr/actualites/sebastien-lecornu-lance-strategie-ministerielle-lintelligence-artificielle> accessed 18 November 2024

—— ‘SKYNET: Courier Detection via Machine Learning’ (*The Intercept*, 8 May 2015) <https://theintercept.com/document/skynet-courier> accessed 25 November 2022

Abraham Y ‘‘Lavender’: The AI machine directing Israel’s bombing spree in Gaza’ (*+972 Magazine*, 3 April 2024) <https://www.972mag.com/lavender-ai-israeli-army-gaza/> accessed 13 April 2024

Abraham Y ‘‘A mass assassination factory’: Inside Israel’s calculated bombing of Gaza’ (*+972 Magazine*, 30 November 2023) <https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/> accessed 13 April 2024

Akram F, Keath L ‘Israel strikes Haza home of Hamas leader, destroys AP office’ (*AP*, 16 May 2021) <https://apnews.com/article/israel-west-bank-gaza-middle-east-israel-palestinian-conflict-7974cc0c03897b8b21e5fc2f8c7d8a79> accessed 21 May 2021

Bergengruen V ‘How Tech Giants Turned Ukraine Into an AI War Lab’ (*TIME*, 8 February 2024) <https://time.com/6691662/ai-ukraine-war-palantir/> accessed 13 April 2024

Biddle S ‘U.S. Military Makes First Confirmed OpenAI Purchase for War-Fighting Forces’ (*The Intercept*, 25 October 2024) <https://theintercept.com/2024/10/25/africom-microsoft-openai-military/> accessed 11 November 2024

Chen S ‘China’s military lab AI connects to commercial large language models for the first time to learn more about humans’ (*South China Morning Post*, 12 January 2024) <https://www.scmp.com/news/china/science/article/3248050/chinas-military-lab-ai-connects-commercial-large-language-models-first-time-learn-more-about-humans> accessed 7 April 2024

Conger K ‘Google Is Helping the Pentagon Build AI for Drones’ (*Gizmodo*, 6 March 2018) <https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533> accessed 29 September 2022

Coulter M ‘NATO targets AI, robots and space tech in $1.1 billion fund’ (*Reuters*, 18 June 2024) <https://www.reuters.com/technology/artificial-intelligence/nato-targets-ai-robots-space-tech-11-billion-fund-2024-06-18/> accessed 12 November 2024

Currier C, Greenwald G, Fishman A ‘U.S. Government Designated Prominent Al Jazeera Journalist as “Member of Al Qaeda’ (*The Intercept*, 8 May 2015) <https://theintercept.com/2015/05/08/u-s-government-designated-prominent-al-jazeera-journalist-al-qaeda-member-put-watch-list/> accessed 19 November 2020

Dave P ‘Exclusive: Ukraine has started using Clearview AI’s facial recognition during war’ (*Reuters*, 14 March 2022) <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearview-ais-facial-recognition-during-war-2022-03-13/> accessed 27 June 2023

Dave P, Jeffrey Dastin J ‘Exclusive: Ukraine has started using Clearview AI’s facial recognition during war’ (*Reuters*, 14 March 2022) <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearview-ais-facial-recognition-during-war-2022-03-13/> accessed 17 June 2022

Defence Science and Technology Laboratory ‘Dstl AI success with AUKUS’ (*GOV.UK*, 15 December 2023) <https://www.gov.uk/government/news/dstl-ai-success-with-aukus> accessed 13 April 2024

Druziuk Y ‘A Citizen-like chatbot allows Ukrainians to report to the government when they spot Russian troops – here’s how it works’ (*Business Insider*, 18 April 2022) <https://www.businessinsider.com/ukraine-military-e-enemy-telegram-app-2022-4?r=US&IR=T> accessed 22 July 2022

Defence Science and Technology Laboratory ‘AUKUS trial advances AI and autonomy collaboration’ (*GOV.UK*, 5 February 2024) <https://www.gov.uk/government/news/aukus-trial-advances-ai-and-autonomy-collaboration> accessed 13 April 2024

Garamone J ‘Centcom Commander: Communications Breakdowns, Human Errors Led to Attack on Afghan Hospital’ (*DoD News*, 29 April 2016) <https://www.defense.gov/News/News-Stories/Article/Article/746393/centcom-commander-

communications-breakdowns-human-errors-led-to-attack-on-afgha/> accessed 20 November 2023

Goertzel B 'Artificial Superintelligence Could Arrive by 2027, Scientist Predicts' (*Futurism*, 7 March 2024) <https://futurism.com/artificial-superintelligence-agi-2027-goertzel> accessed 29 March 2024

Grylls G 'Kyiv outflanks analogue Russia with ammunition from Big Tech' (*The Times*, 24 December 2022) <https://www.palantir.com/assets/xrfr7uokpv1b/1Fw2bFxYXmu3RWX7FvssB9/e64d19b6f042bda3d2a61e4fc43ea6ec/TheTimes.pdf> accessed 29 May 2023

Hao K 'How Facebook got addicted to spreading misinformation' (*MIT Technology Review*, 11 March 2021) <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/> accessed 21 May 2021

Harper J 'Anduril, Palantir awarded contracts for Army robotic combat vehicle software system integration' (*DefenseScoop*, 3 April 2024) <https://defensescoop.com/2024/04/03/anduril-palantir-diu-army-contract-robotic-combat-vehicle-software/> accessed 13 April 2024

Heaven W D 'Predictive policing algorithms are racist. They need to be dismantled.' (*MIT Technology Review*, 17 July 2020) <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/> accessed 10 July 2024

Hern A 'Fitness tracking app Strava gives away location of secret US army bases' (*The Guardian*, 28 January 2018) <https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases> accessed 1 February 2022

Hill K 'Microsoft Plans to Eliminate Face Analysis Tools in Push for 'Responsible A.I.' (*The New York Times*, 21 June 2022) <https://www.nytimes.com/2022/06/21/technology/microsoft-facial-recognition.html> accessed 1 July 2022

Judah T 'How Kyiv was saved by Ukrainian ingenuity as well as Russian blunders' (*Financial Times*, 10 April 2022) <https://www.ft.com/content/e87fdc60-0d5e-4d39-93c6-7cfd22f770e8> accessed 22 April 2022

Kayser-Bril N 'Google apologizes after its Vision AI produced racist results' (*Algorithm Watch*, 7 April 2020) <https://algorithmwatch.org/en/story/google-vision-racism/> accessed 24 May 2020

Knight M 'Russian commander killed while jogging may have been tracked on Strava app' (*CNN World*, 12 July 2023) <https://edition.cnn.com/2023/07/11/europe/russian-submarine-commander-killed-krasnador-intl/index.html> accessed 10 July 2024

Lariosa A 'HD HHI And Palantir Reveal Tenebris Unmanned Surface Vessel' (*Naval News*, 16 May 2024) <https://www.navalnews.com/naval-news/2024/05/hd-hhi-and-palantir-reveal-tenebris-unmanned-surface-vessel/> accessed 27 September 2024

Lawson A ‘Google to buy nuclear power for AI datacentres in ‘world first’ deal’ (The Guardian, 15 October 2024) <https://www.theguardian.com/technology/2024/oct/15/google-buy-nuclear-power-ai-datacentres-kairos-power> accessed 6 June 2025

McFadden R D ‘WAR IN THE GULF: Iraq; IRAQIS ASSAIL U.S. AS RESCUE GOES ON’ (*The New York Times*, 15 February 1991) <https://www.nytimes.com/1991/02/15/world/war-in-the-gulf-iraq-iraqis-assail-us-as-rescue-goes-on.html> accessed 16 July 2023

McKernan B and Davies H ‘‘The machine did it coldly’: Israel used AI to identify 37,000 Hamas targets’ (*The Guardian*, 3 April 2024) <https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes> accessed 13 April 2024

Meaker M ‘Ukraine’s War Brings Autonomous Weapons to the Front Lines’ (*WIRED*, 24 February 2023) <https://www.wired.com/story/ukraine-war-autonomous-weapons-frontlines/> accessed 26 February 2023

Milmo D ‘OpenAI raises $6.6bn in funding, is valued at $157bn’ (*The Guardian*, 2 October 2024) <https://www.theguardian.com/technology/2024/oct/02/openai-raises-66bn-in-funding-is-valued-at-157bn> accessed 11 November 2024

Parikh D ‘Learning Paradigms in Machine Learning’ (*Medium*, 7 July 2018) <https://medium.com/datadriveninvestor/learning-paradigms-in-machine-learning-146ebf8b5943> accessed 24 May 2020

Parvaz D ‘Journalists allege threat of drone execution by US’ (*Al Jazeera*, 2 April 2017) <https://www.aljazeera.com/news/2017/4/2/journalists-allege-threat-of-drone-execution-by-us> accessed 25 November 2022

Pellerin C ‘Project Maven to Deploy Computer Algorithms to War Zone by Year’s End’ (*DOD News*, 21 July 2017) <https://www.defense.gov/News/News-Stories/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/> accessed 16 June 2023

Perrigo B ‘Exclusive: Google Contract Shows Deal with Israel Defense Ministry’ (*TIME*, 12 April 2024) <https://time.com/6966102/google-contract-israel-defense-ministry-gaza-war/> accessed 13 April 2024

Porter J ‘UK ditches exam results generated by biased algorithm after student protests’ (*The Verge*, 17 August 2020) <https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications> accessed 10 July 2024

Rowan D ‘DeepMind: inside Google’s super-brain’ (*WIRED*, 22 June 2015) <https://www.wired.com/story/deepmind/> accessed 20 March 2024

Saballa J 'US Army Seeking AI System That Predicts Enemy Actions' (*The Defense Post*, 11 July 2023) <https://www.thedefensepost.com/2023/07/11/us-army-ai-system/> accessed 14 July 2023

Satariano A and Mozur P 'Facial Recognition Powers 'Automated Apartheid' in Israel, Report Says' (*The New York Times*, 1 May 2023) <https://www.nytimes.com/2023/05/01/technology/israel-palestine-facial-recognition.html> accessed 29 June 2023

Savitz E J 'Nvidia CEO Sees Artificial General Intelligence Within 5 Years – With a Big Caveat' (*Barron's*, 20 March 2024), <https://www.barrons.com/livecoverage/nvidia-gtc-ai-conference/card/nvidia-ceo-sees-artificial-general-intelligence-within-5-years-with-a-big-caveat-e2NjhpHnNlqI1qv3QvQw> accessed 20 March 2024

Shachtman N 'Exclusive: Computer Virus Hits U.S. Drone Fleet' (*Wired*, 7 November 2011) <https://www.wired.com/2011/10/virus-hits-drone-fleet/> accessed 9 April 2020

Solon O 'Facial recognition's 'dirty little secret': Millions of online photos scraped without consent' (*NBC News*, 12 March 2019) <https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921> accessed 17 June 2022

Van Esland S L, O'Hare R 'COVID-19: Imperial researchers model likely impact of public health measures' (*Imperial College London*, 17 March 2020) <https://www.imperial.ac.uk/news/196234/covid19-imperial-researchers-model-likely-impact/> accessed 24 May 2020

Vergun D 'DARPA Aims to Develop AI, Autonomy Applications Warfighters Can Trust' (*DOD News*, 27 March 2024) <https://www.defense.gov/News/News-Stories/Article/Article/3722849/darpa-aims-to-develop-ai-autonomy-applications-warfighters-can-trust/> accessed 12 November 2024

Wakabayashi D and Shane S 'Google Will Not Renew Pentagon Contract That Upset Employees' (*The New York Times*, 1 June 2018) <https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html> accessed 27 February 2023

Winfield A 'Artificial intelligence will not turn into a Frankenstein's monster' (*The Guardian*, 10 August 2014) <https://www.theguardian.com/technology/2014/aug/10/artificial-intelligence-will-not-become-a-frankensteins-monster-ian-winfield> accessed 20 March 2024

Xiang C 'OpenAI Used Kenyan Workers Making $2 an Hour to Filter Traumatic Content from ChatGPT' (*VICE*, 18 January 2023) <https://www.vice.com/en/article/wxn3kw/openai-used-kenyan-workers-making-dollar2-an-hour-to-filter-traumatic-content-from-chatgpt> accessed 16 June 2023

Yen H 'Blinken hasn't seen any evidence on AP Gaza building strike' (*AP*, 17 May 2021), <https://apnews.com/article/middle-east-israel-business-israel-palestinian-conflict-government-and-politics-abd641af1607fbae7f49e1cce7dbc49e> accessed 21 May 2021

Zetter K ‘So, the NSA Has an Actual Skynet Program’ (*WIRED*, 8 May 2015) <https://www.wired.com/2015/05/nsa-actual-skynet-program/> accessed 19 November 2020

3.5. Presentations, statements and other interventions at conferences

Ambassador Petit C, Permanent Representative of France to the Conference on Disarmament, Remarks at the UNIDIR Side-Event on ‘Guidelines for the development of national strategies on AI in security and defence’ (New York City, 25 October 2024)

ICRC, Statement delivered at the CCW GGE on LAWS, ‘Autonomous weapons: The ICRC recommends adopting new rules’ (Geneva, 3-13 August 2021) <https://www.icrc.org/en/document/autonomous-weapons-icrc-recommends-new-rules> accessed 19 December 2022

Lubell N, Public evidence for the Artificial Intelligence in Weapons Systems House of Lords Select Committee, ‘AI in Weapon Systems Committee’ (London, 23 March 2023) <https://committees.parliament.uk/oralevidence/12931/pdf/> accessed 29 May 2023

Murray D, Public evidence for the Artificial Intelligence in Weapons Systems House of Lords Select Committee, ‘AI in Weapon Systems Committee’ (London, 23 March 2023) <https://committees.parliament.uk/oralevidence/12931/pdf/> accessed 29 May 2023

Nakamitsu I, Under-Secretary-General and High Representative for Disarmament Affairs, United Nations Office for Disarmament Affairs at the REAIM session on ‘AI, Or Not AI? – Debunking Commonly-Held Assumptions on Military AI’ (The Hague, 15 February 2023)

Permanent Mission of the Holy See to the United Nations and Other International Organizations in Geneva at the 2021 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS) of the Convention on Certain Conventional Weapons, Statement, ‘Item 5(b) Characterization of the systems under consideration’ (Geneva, 4 August 2021) <https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2021/gge/statements/4Aug_HolySee.pdf> accessed 6 January 2023

Wareham M, Statement, ‘Statement on Lethal Autonomous Weapons Systems to the CCW Annual Meeting’ (Geneva, 16 November 2022) <https://www.hrw.org/news/2022/11/16/statement-lethal-autonomous-weapons-systems-ccw-annual-meeting-0> accessed 6 January 2023

Young S, Presentation for the NDIA National Test & Evaluation Conference, ‘Autonomy Test & Evaluation Verification & Validation (ATEVV) Challenge Area’ (3 March 2016) <https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2016/Test/Young.pdf> accessed 22 September 2022

3.6. Social media

Israel Defense Forces (@IDF) ‘Yesterday, we targeted an important base of operations for Hamas' military intel in Al Jala Tower in Gaza. The base gathered intel for attacks against

Israel, manufactured weapons & positioned equipment to hamper IDF operations. (1/3)' (*Twitter*, 16 May 2021 07:14 AM) <https://twitter.com/IDF/status/1393797067273867266> accessed 21 May 2021

Martin M (@ThreshedThought) 'Russian SIMs roaming in Ukraine. Gives you a good idea of the concentration of their forces.' (*Twitter*, 12 May 2022 11:48 AM) <https://twitter.com/ThreshedThought/status/1524687923676954624> accessed 9 June 2022.

Palantir, *Palantir AIP | Defense and Military* (YouTube, 25 April 2023) <https://www.youtube.com/watch?v=XEM5qz__HOU> accessed 27 September 2024