Autonomous Cyber Capabilities under International Law

Edited by
Rain Liivoja
Ann Väljataga

CCDCOE

Autonomous Cyber Capabilities under International Law

Rain Liivoja and Ann Väljataga (Eds.)

CCDCOE
NATO COOPERATIVE
CYBER DEFENCE
CENTRE OF EXCELLENCE

Autonomous Cyber Capabilities under International Law

ISBN (print): 978-9916-9565-2-6
ISBN (pdf): 978-9916-9565-3-3

2021 © NATO CCDCOE Publications
NATO CCDCOE Publications
Filtri tee 12, 10132 Tallinn, Estonia
Phone: +372 717 6800
E-mail: publications@ccdcoe.org
Web: www.ccdcoe.org
Cover design & content layout: Stuudio Stuudio
Language editing: Isabelle Peart

Table of Contents

AUTHORS AND EDITORS

Russell Buchan is Senior Lecturer in International Law at the University of Sheffield, UK. He has published widely in the field of public international law, including three monographs: *International Law and the Construction of the Liberal Peace* (Hart, 2013), *Cyber Espionage and International Law* (Hart, 2018) and *Regulating the Use of Force in International Law: Stability and Change* (Edward Elgar Publishing, 2021). He is also Co-Editor in Chief of the *Journal of International Humanitarian Legal Studies*.

Thomas Burri is a professor of international law and European law at the University of St. Gallen in Switzerland. His research investigating AI and autonomous systems for almost a decade has been widely published.

Damian Copeland is a legal practitioner whose expertise and doctoral studies are in the Article 36 legal review of weapons, specifically focused on weapons and systems enhanced by Artificial Intelligence. He is a weapons law expert with over thirty years military service, including multiple operational deployments where he has extensive experience in the application of operational law in support of military operations.

Ashley Deeks is the E James Kelly Jr–Class of 1965 Research Professor at the University of Virginia Law School and the Director of its National Security Law Center. She serves on the US State Department's Advisory Committee on International Law and the Board of Editors for the American Journal of International Law. She is a Senior Fellow at the Miller Center and a senior contributor to *Lawfare*.

Abhimanyu George Jain is a PhD candidate at the Graduate Institute of International and Development Studies and a research associate at the LAWS & War Crimes Project.

Samuli Haataja is a lecturer at Griffith Law School, Griffith University. His research explores law and emerging technologies with a focus on cyberspace and public international law. He published his book *Cyber Attacks and International Law on the Use of Force: The Turn to Information Ethics* with Routledge in 2019, and he has published in various international law and technology journals. He is also a member of the Program on the Regulation of Emerging Military Technologies (PREMT) and the Institute of Electrical and Electronics Engineers Society on Social Implications of Technology (IEEE SSIT).

Eric Talbot Jensen is the Robert W. Barker Professor of Law at Brigham Young University where he teaches and writes in the areas of Public International Law, National Security Law, the Law of Armed Conflict and Criminal Law. Prior to becoming a professor, he worked as a legal advisor to United States military commanders while deployed to Bosnia, Kosovo, Macedonia and Iraq.

Dustin A Lewis is the Research Director at the Harvard Law School Program on International Law and Armed Conflict. With a focus on public international law sources and methodologies, Mr. Lewis leads research into several wide-ranging contemporary challenges concerning armed conflict. Among his current areas of focus, Mr. Lewis heads the research for the project on 'International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns'.

Rain Liivoja is Associate Professor and Deputy Dean (Research) at the University of Queensland Law School, where he leads the Law and the Future of War Research Group. He is also a Senior Fellow with the Lieber Institute for Law and Land Warfare at the United States Military Academy at West Point, and an Affiliated Research Fellow of the Erik Castrén Institute of International Law and Human Rights at the University of Helsinki. Rain co-edits the *Journal of International Humanitarian Legal Studies*. He has previously served on Estonian delegations to multilateral meetings on humanitarian law and arms control.

Peter Margulies is a Professor of Law at Roger Williams University School of Law in Rhode Island, where he teaches National Security Law and International Law. He is a co-author of the treatise, *National Security Law: Principles and Policy* (with Geoffrey S. Corn, Eric Talbot Jensen and Jimmy Gurule) (2d ed. 2019). Professor Margulies' recent articles include, *Risk and Rights in Transatlantic Data Transfers: EU Privacy Law, U.S. Surveillance, and the Search for Common Ground* (with Ira Rubinstein), Connecticut Law Review (forthcoming 2021), https://ssrn.com/abstract=3786415.

Tim McFarland is a Research Fellow in the Law and the Future of War Research Group at the University of Queensland in Australia. He completed his PhD studies at Melbourne Law School. His research focuses on the legal aspects of autonomous technologies, in particular of utilising increasingly autonomous weapon systems in armed conflict.

Michael N Schmitt is Professor of International Law at the University of Reading in the United Kingdom. He is also Senior Fellow at the NATO Cooperative Cyber Defence Centre of Excellence; Francis Lieber Distinguished Scholar at West Point's Lieber Institute; Charles H Stockton Distinguished Scholar at the US Naval War College; Distinguished Scholar at the University of Texas' Strass Center for International Security and Law; and Director of Legal Affairs for Cyber Law International. He directed the *Tallinn Manual* project from 2009–2017.

Tanel Tammet is a professor at the School of Information Technologies of the Tallinn University of Technology. He has a Ph.D from the Chalmers University of Technology/ University of Gothenburg and an M.Sc in applied mathematics from the University of Tartu. His research focuses on the theory, implementations and applications of automated and commonsense reasoning. He has led cross-sectoral projects on AI in cyber security in cooperation with e.g. European Defence Agency, European Space Agency and US and Estonian Cyber Commands.

Alec Tattersall is a serving member of the Royal Australian Air Force (RAAF). This service has resulted in extensive involvement with emerging technologies and weapons reviews including: undertaking interim and final weapons reviews across a range of traditional and emerging technologies; analysis of national weapons review systems; establishing traditional and novel weapons review processes, and application of weapons law through numerous operational postings and deployments.

Daniel Trusilo is a PhD student at the University of St Gallen. He previously served as a member of the US armed forces, *inter alia*, in Iraq. Most recently he worked as a humanitarian advisor to the military for the US Agency for International Development.

Nicholas Tsagourias is Professor of International Law at the University of Sheffield and Visiting Professor at the Paris School of International Affairs, SciencesPo. He is the editor with Russell Buchan of the *Handbook of International Law and Cyberspace* (2nd edn, Elgar 2021).

Ann Väljataga is a legal researcher at NATO Cooperative Cyber Defence Centre of Excellence, she holds an LLM in Law and Technology and has a background in digital rights advocacy. Her areas of research include the concurrent applicability of IHL and IHRL to cyber operations, legal reviews of novel cyber capabilities, national cyber security strategies and cyber security of space assets.

ABBREVIATIONS

ACC	autonomous cyber capabilities
ACD	active cyber defence
AI	artificial intelligence
AP I	Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts
AUMF	Authorization for Use of Military Force (US)
AWS	autonomous weapon system
C2	command and control
CCDCOE	NATO Cooperative Cyber Defence Centre of Excellence
CCW	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
CGC	Cyber Grand Challenge
CIA	Central Intelligence Agency (US)
CIWS	close-in weapon system
CNAS	Center for a New American Security
COVID-19	Coronavirus Disease 2019
C-RAM	counter rocket, artillery and mortar system
CRS	cyber reasoning system
DARPA	Defense Advanced Research Projects Agency (US)
DCiD	digital precision strike suite collateral damage estimation
DDoS	distributed denial-of-service attack
DIB	Defense Innovation Board (US)
DNI	Director of National Intelligence (US)
DoD	US Department of Defence
DSB	Defense Science Board (US)
ECHR	European Court of Human Rights
EJIL	European Journal of International Law
ENMOD	environmental modification
F2T2EA	Find, Fix, Track, Target, Engage, Assess *(targeting of hostile forces)*
FCAS	Future Combat Air System
FISA	Foreign Intelligence Surveillance Act (US)
FISC	Foreign Intelligence Surveillance Court (US)
FY	financial year
GGE LAWS	Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems
GGE	Group of Governmental Experts
GPS	Global Positioning System
GTI	Global Threat Intelligence
GUI	graphical user interface
HPCR	Program on Humanitarian Policy and Conflict Research at Harvard University
IAP	Internet Access Provider'
ICC	International Criminal Court
ICD	industrial control device
ICJ	International Court of Justice
ICL	international criminal law
ICR	individual criminal responsibility
ICRC	International Committee of the Red Cross

ICS	industrial control system
ICTR	International Criminal Tribunal for Rwanda
ICTY	International Criminal Tribunal for the Former Yugoslavia
IDS	intrusion detection system
IGE	International Group of Experts
IHL	international humanitarian law
IHRL	international human rights law
ILC	International Law Commission
IP	Internet protocol
IPS	intrusion prevention system
IPS	intrusion protection system
ISIS	Islamic State of Iraq and Syria
ISR	intelligence, surveillance, and reconnaissance
IT	information technology
JADOCS	joint automated deep control system
JAG	Judge Advocate General
LAWS	lethal autonomous weapons systems
LOAC	law of armed conflict
LPWS	Land-Based Phalanx Weapon System
MISP	Malware Information Sharing Platform
MMAR	manage, monitor, automate and respond
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NDAA	National Defense Authorization Act (US)
NGO	non-governmental organisation
NSA	National Security Agency (US)
PLC	programmable logic controller
RPA	remotely piloted aircraft
SCADA	supervisory control and data acquisition
SIEM	security information and event management
SOAR	security orchestration, automation and response
SOC	security operations centre
STIX	Structured Threat Information eXpression
TAXII	Trusted Automated eXchange of Indicator Information
TTP	tactics, techniques and procedures for carrying out military operations
UAV	unmanned aerial vehicle
UK	United Kingdom
UN GGE	United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security
UN	United Nations
UNGA	United Nations General Assembly
UNIDIR	United Nations Institute for Disarmament Research
US	United States
USG	United States Government
WPR	War Powers Resolution (US)

ACKNOWLEDGEMENTS

This volume was produced during a global public health emergency, which in one way or another affected the health and wellbeing of everyone involved. Especially with that in mind, the editors gratefully acknowledge the interest, insight and effort of all the contributors to this volume, not only manifested in their own chapters, but also in the thorough engagement with those of others. Special thanks are also due to Isabelle Peart for her diligent copyediting and citation checking and to Maarja Naagel from whom the idea for the book originated. Some chapters of this book were also published in a slightly different form in *International Law Studies*.[1]

1 See Peter Margulies, 'Autonomous Weapons in the Cyber Domain: Balancing Proportionality and the Need for Speed' (2020) 96 International Law Studies 394; Ashley Deeks, 'Will Cyber Autonomy Undercut Democratic Accountability?' (2020) 96 International Law Studies 464; Michael N. Schmitt, 'Autonomous Cyber Capabilities and the International Law of Sovereignty and Intervention' (2020) 96 International Law Studies 549; Eric Talbot Jensen, 'Autonomy and Precautions in the Law of Armed Conflict' (2020) 96 International Law Studies 577; Russell Buchan and Nicholas Tsagourias, 'Autonomous Cyber Weapons and Command Responsibility' (2020) 96 International Law Studies 645.

Chapter 1

Cyber Autonomy and International Law: An Introduction

Ann Väljataga and Rain Liivoja

I
THE STATE OF THE DISCUSSION

In international law circles, conversations about cyber operations and autonomous (military) systems have proceeded on parallel tracks. This is somewhat counterintuitive, given that these enquiries share much of their technological, legal and strategic context.

At times, cyber considerations have received a mention in the debates within the Group of Governmental Experts on Lethal Autonomous Weapons Systems ('GGE LAWS').[1] Notably, the GGE agreed in 2017 that, when developing weapons systems with autonomous functionality, States must consider, *inter alia*, 'non-physical safeguards (including cyber security

1 Formally, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, which has been convened at the direction 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 (adopted 10 October 1980, entered into force 2 December 1983) 1342 UNTS 137.

against hacking or data spoofing)'.[2] The idea of an autonomous cyber weapon, however, has been studiously avoided. At other times, like in the drafting process of United States Department of Defense Directive on Autonomy in Weapons Systems,[3] cyber capabilities have been consciously put to one side because of time-constraints and the risk of adding complexity to the already entangled subject matter.[4] Generally speaking, debates over autonomous weapons systems have been more engaged with the anticipated kinetic effects of the technologies and the understanding, preservation or reconceptualisation of human judgment or control.[5]

Amidst the discussions on cyber operations within the United Nations Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security, the potential of autonomous features to 'significantly reduce the predictability of the [information and communications technologies] … and thus constitute a source for anxiety and mistrust' has been 'noted',[6] but as of now it has not been considered to alter the fundamental legal questions posed by the cyber domain. *Tallinn Manual 2.0*, on the other hand, acknowledges the capacity for autonomous operation of 'software agents' and 'worms' when defining those terms,[7] but does not specifically interrogate the operational or legal implications of this autonomous functionality.

This siloing is all the more peculiar considering that cyber capabilities 'contain an inherent tendency towards autonomous functionality',[8] as they are programmed ahead of time to perform a particular task. In practice, highly autonomous features have been integrated into cyber capabilities for more than a decade. In 2010, it was discovered that a worm dubbed Stuxnet had infiltrated the supervisory control and data acquisition ('SCADA')-systems of an Iranian uranium enrichment

2 'Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems' (23 October 2018) UN Doc CCW/GGE.1/2018/3 ('2018 GGE LAWS Report') [21(e)].

3 US Department of Defense, Directive 3000.09: Autonomy in Weapons Systems (8 May 2017, incorporating change 1).

4 Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (WW Norton & Company 2018) 227–8.

5 See, eg, 'Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems' (25 September 2019) UN Doc CCW/GGE.1/2019/3.

6 'Chair's Summary of Informal Intersessional Consultative meeting of the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security' (December 2019) [56] <https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/01/200128-OEWG-Chairs-letter-on-the-summary-report-of-the-informal-intersessional-consultative-meeting-from-2-4-December-2019.pdf>.

7 Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017) 567 and 568 ('*Tallinn Manual 2.0*').

8 Alec Tattersall and Damian Copeland, 'Reviewing Autonomous Cyber Capabilities', this volume, ch 10, section I.C.

facility. Although the worm needed human assistance to access the computer system at the facility, from there on it was able to manipulate data without receiving real-time instructions from its creators.[9] The most notorious piece of malware in the history of cyber security was, therefore, an autonomous cyber capability. Arguably — depending on one's definition of a 'weapon'[10] — Stuxnet was also the first highly autonomous weapon.[11]

The use of artificial intelligence ('AI') techniques, such as machine learning, can increase the level of autonomy of cyber capabilities. It will also amplify the speed, power, and scale of future cyber operations.[12] The need to prepare for AI-enabled cyber conflict was communicated with exceptional clarity by the US National Security Commission on Artificial Intelligence in its final report published in March 2021. Besides reiterating the importance of developing, testing and deploying 'AI-enabled cyber defences', the report urged the US government to 'promulgate a declaratory policy that addresses the use of AI in cyber operations'.[13]

II
FRAMING THE ISSUES

This edited volume aims to merge the discourses on the application of international law to cyber operations and autonomous systems. To that end, it explores if and how international law differentiates between 'embodied' and 'disembodied' autonomous systems (that is, cyber-physical systems and software, respectively),[14] what to consider

9 See, eg, Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (WW Norton 2018) 214–15.

10 Further on whether cyber capabilities can be seen as weapons, see, eg, Jeffrey T Biller and Michael N Schmitt, 'Classification of Cyber Capabilities and Operations as Weapons, Means, or Methods of Warfare' (2019) 95 International Law Studies 179; Tattersall and Copeland (n 8) section II.3.1.

11 See, eg, Jason Healey, 'Stuxnet and the Dawn of Algorithmic Warfare' (*Huffington Post*, 16 April 2013) <www.huffingtonpost.com/jason-healey/stuxnet-cyberwarfare_b_3091274.html> ('Stuxnet … appears to be the first autonomous weapon with an algorithm, not a human hand, pulling the trigger'). Further on whether cyber capabilities can be seen as weapons, see, eg, Jeffrey T Biller and Michael N Schmitt, 'Classification of Cyber Capabilities and Operations as Weapons, Means, or Methods of Warfare' (2019) 95 International Law Studies 179; Tattersall and Copeland (n 8) section II.3.1.

12 See, eg, James Johnson and Eleanor Krabill, 'AI, Cyberspace, and Nuclear Weapons' (*War on the Rocks*, 31 January 2020) <https://warontherocks.com/2020/01/ai-cyberspace-and-nuclear-weapons/>.

13 National Security Commission on Artificial Intelligence, 'Final Report' (2021) <https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf > 283.

14 For a discussion of the distinction between embodied and disembodied autonomous systems, see Daniel Trusilo and Thomas Burri, 'Ethical Artificial Intelligence: An Approach to Evaluating

when applying the principles of international law to cyber operations involving autonomous functionality, and how to establish responsibility and accountability. In 2019, NATO CCDCOE published an exploratory working paper on these issues.[15] The working paper highlighted, *inter alia*, the relevance to autonomous cyber capabilities of questions around the element of intent in prohibited intervention, and the *mens rea* in international responsibility and liability schemes, as well as the capacity of autonomous systems to assess the severity of attacks and to implement precautionary measures. This volume contains a more in-depth examination of these and many other issues.

The book adopts a broad conceptualisation of autonomy, which is not limited to highly sophisticated, self-governing and AI-enabled solutions. Rather, autonomous operation is taken to simply mean the ability of a system to perform some task without requiring real-time interaction with a human operator.[16] On this view, autonomy exists on a continuum and is function-specific. Hence, a system can have a high degree of autonomy in some function while at the same time having a low level of autonomy or none whatsoever in other functions.[17] Defining autonomy broadly ensures that the widest possible range of legal implications is considered, not only the problems that may be associated with, for example, human-like 'artificial general intelligence'.

Regardless of the degree of autonomy or any other particularity of the hypothetical or existing systems mentioned in this book, a few common propositions guide the legal analysis. First of all, autonomous capabilities are seen as operationally desirable because they can allow systems to outperform humans, for example in terms of speed or precision,[18] or give them the ability to analyse large datasets. Second, surpassing human performance in certain respects implies that in these respects autonomous systems cannot be subjected to real-time human intervention. Indeed, the whole purpose of autonomous functionality — to reiterate, increased speed, precision and data-processing capability — would likely be defeated by having a human operator second-guessing the system

Disembodied Autonomous Systems', this volume, ch 4.

15 Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) 10 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>.

16 ibid 10.

17 ibid 7–11; Tim McFarland, 'The Concept of Autonomy', this volume, ch 2; Law and the Future of War Research Group, 'Autonomy' (University of Queensland Law School, 2 October 2020) <https://law.uq.edu.au/research/future-war/autonomy>.

18 See, eg, Paul C Ney Jr, 'Keynote Address at the Israel Def. Forces 3rd International Conference on the Law of Armed Conflict' (*Lawfare*, 28 May 2019) <https://www.lawfareblog.com/defense-department-gen-eral-counsel-remarks-idf-conference>.

every step of the way. Third, the decision to use a specific autonomous capability in specific circumstances is nevertheless a judgment attributable to a human actor. Fourth, human actors can be held individually responsible for the consequences of such decisions. A major implication of the last two propositions is that humans (as well as the States and international organisations who they serve) are legal actors for the purposes of complying with international law, and that relevant conduct must be capable to being discerned, attributed, understood and assessed.[19]

Autonomous cyber capabilities can be categorised in a number of ways: passive *versus* active, offensive *versus* defensive, and so on. Such categorisations may be helpful in highlighting different properties of specific cyber capabilities by contrasting them to others but, of course, such taxonomical exercises have no formal legal significance. Interestingly, from a legal perspective, offensive cyber capabilities are not necessarily the most problematic. They are generally single-use bespoke capabilities,[20] which means that the circumstances of their use, and their intended and anticipated effects, can be studied in some detail, and specific legal advice can be provided. Conversely, autonomous cyber defence capabilities designed to conduct proactive operations in adversary networks have the potential to cause the most disruption and raise the most complicated legal issues. These systems exhibit a high degree of autonomy not only in selecting and engaging targets, but also in identification of threats, the sources thereof and choosing the optimal means and timing of response. While the most sophisticated cyber reasoning systems have demonstrated such capabilities at an emergent state, the current technological advances are still first and foremost addressing the more passive, but technically no less intricate, types of cyber defence, such as intrusion and anomaly detection.[21] What presents a legal challenge is that such systems must potentially be able to operate within different legal frameworks (for example, both in situations where the law of armed conflict applies and does not apply) and scenarios (for example, when the right to self-defence is or is not engaged).

A highly autonomous proactive cyber capability, though still rather rare in practice, offers up challenges for legal analysis, since it does not lend itself to simple analogies, and requires careful consideration of how the law regulates both cyber operations and the use of autonomous

19 See Dustin A Lewis, 'Preconditions for Applying International Law to Autonomous Cyber Capabilities', this volume, ch 2.
20 Tattersall and Copeland (n 8) section II.1.2.
21 See Tanel Tammet, 'Autonomous Cyber Defence Capabilities', this volume, ch 3.

systems. From the cyber environment such a capability inherits a special sort of covertness, which makes its use particularly likely to escape democratic, executive or judicial oversight and authorisation.[22] Also, it propagates easily, quickly and at a minimal cost. A system with a high degree of autonomy would interact with its environment, without ongoing external supervision, while ideally remaining in the framework of the higher-level goals that it has been programmed to pursue.

III
INTERNATIONAL LEGAL OBLIGATIONS

States have unequivocally confirmed the application of existing international law to cyber operations[23] and to the use of autonomous weapons systems.[24] There is little doubt that international law is both relevant and applicable to the use of autonomous cyber capabilities. In many instances, as several contributors to this book point out, autonomy adds complexity to the application of existing rules, but does not necessarily create legal vacuums or render existing rules ineffectual or obsolete. However, like other major technological advancements, autonomous cyber capabilities may demand new interpretations of rules initially designed for entirely different historical circumstances and technological paradigms.

Questions are sometimes raised about the ability of autonomous systems to comply with the law, especially with rules that that require evaluative judgments, such as the principle of proportionality (whether in the context of *jus ad bello* or *jus in bellum*). But this seems to get things backward. The more precise question is whether *humans*, along with States and international organisations, as bearers of obligations under international law, are able to comply with the law whilst using autonomous systems.

22 See Ashley Deeks, 'Will Cyber Autonomy Undercut Democratic Accountability?', this volume, ch 5.

23 'Final Substantive Report of the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security' (10 March 2021) A/AC.290/2021/CRP.2, [34] ('States reaffirmed that international law, and in particular the Charter of the United Nations, is applicable and essential to maintaining peace and stability and promoting an open, secure, stable, accessible and peaceful ICT environment. In this regard, States were called upon to avoid and refrain from taking any measures not in accordance with international law, and in particular the Charter of the United Nations. …')

24 2018 GGE LAWS Report (n 2) [21(a)] ('International humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems').

If the system has the technological ability to perform some legally required assessment, the operator may choose to entrust the system with making that assessment.[25] Where the system cannot perform the assessment, compliance with the law would require human decision-making ahead of time, real-time human intervention, appropriate environmental constraints, or most likely some combination of the above. In any case, the onus is on the human to use technology in a way that complies with the law.

Autonomous systems, especially those relying on current AI techniques, have their limitations. Significantly, such systems may be unintelligible, brittle and biased, resulting in misperformance, which reduces desirable effects or increases adverse effects.[26] Many of these concerns are often discussed under the general heading of 'unpredictability' of autonomous systems. Avoiding, minimising or mitigating the risk of unpredictable or biased behaviour, and dealing with the consequences of such behaviour, are important technological and regulatory problems.

Part of any technological risk reduction strategy would involve rigorous testing, to ensure that the system performs as intended, and an assessment of the ability of the system to be used in a lawful manner. Legal review processes — such as that contemplated for weapons, means and methods of warfare by Article 36 of Additional Protocol I to the Geneva Conventions[27] — take on a particular significance where the operation of a system is to a greater degree pre-determined by its design features than by direct human intervention at the use stage.[28] Meanwhile, the readiness of national weapons review procedures to address new technologies, including autonomous or cyber capabilities, has been questioned by, among others, United Nations Institute for Disarmament Research ('UNIDIR') and International Committee of the Red Cross ('ICRC').[29] Indeed, the methodology of conducting such

25 See, generally, Eric Talbot Jensen, 'Autonomy and Precautions in the Law of Armed Conflict', this volume, ch 9.

26 See Peter Margulies, 'A Moment in Time: Autonomous Cyber Capabilities, Proportionality, and Precautions', this volume, ch 8.

27 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3, ('AP I') art 36.

28 See, eg, United Kingdom, 'Statement for the General Exchange of Views' (LAWS GGE, 9 April 2018) [7].

29 ICRC, 'International Humanitarian Law and the Challenges of Contemporary Armed Conflicts' (October 2019) 29 <https://rcrcconference.org/app/uploads/2019/10/33IC-IHL-Challenges-report_EN.pdf>; 'Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon System' (11 April 2016) <https://www.icrc.org/en/download/file/21606/ccw-autonomous-weapons-icrc-april-2016.pdf>; UNIDIR, 'The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches — A Primer' (2017) 19 <www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches-en-689.pdf> .

reviews may need a rethink.[30] For example, it may no longer be possible to maintain a strict separation between the so-called 'weapons law' (which deals with means of warfare abstractly by reference to their normal or intended use) and the so-called 'targeting law' (which deal with the use of lawful means of warfare in specific circumstances). This is particularly true for offensive cyber capabilities, which, as already noted, tend to be purpose-built for particular operations.

The taking of precautionary measures to reduce harm to civilians and civilian objects becomes ever more significant as technological capabilities evolve. However, even the obligation to take precautions under *jus in bello*[31] does arguably not require constant involvement of human judgement.[32] The focus shifts from direct human intervention at the deployment stage to the role of States, commanders, developers and other actors far removed from the actual moment of deployment, but better positioned to foresee and influence the behaviour of the autonomous cyber capability. Furthermore, while the taking of feasible precautions is an explicit legal obligation under *jus in bello*, such a duty arguably also forms part of *jus ad bellum* and the law on countermeasures.[33]

The lack of predictability may in extreme cases lead to a situation where an autonomous cyber capability behaves in a manner that was not intended or even foreseen by the operator, resulting in some injury or harm to States or individuals, which the law seeks to avoid. With respect to the application of some rules of international law, such lack of intent is immaterial. Notably, the international law rule most susceptible to breaches through cyber operations — the obligation to respect the sovereignty of other States[34] — does not refer to knowledge or intent. If an autonomous cyber capability deployed by a State breaches the sovereignty of another State, there is no need to enquire whether such a breach was intentional — an inadvertent breach of sovereignty is nevertheless a breach of sovereignty.[35]

30 Tattersall and Copeland (n 8); see also, more generally, Gary D Brown and Andrew O Metcalf, 'Easier Said Than Done: Legal Reviews of Cyber Weapons' (2014) 7 Journal of National Security Law & Policy 115.
31 AP I arts 57–58.
32 Eric Talbot Jensen 'Autonomy and Precautions in the Law of Armed Conflict', this volume, ch 9.
33 Margulies (n 26).
34 See *Tallinn Manual 2.0* (n 7) rule 4.
35 See Michael N Schmitt 'Autonomous Cyber Capabilities and the International Law of Sovereignty and Intervention', this volume, ch 7; see also Liivoja, Naagel and Väljataga (n 15) 19.

IV
INTERNATIONAL LEGAL RESPONSIBILITY

The question of the exact degree to which it is possible to predict and manage the outcomes of an operation using autonomous systems is at the heart of the legal discussion. But, as mentioned above, the issue is somewhat less acute with respect to State responsibility because the relevant primary rules often do not specify a mental element. Also, it is debatable to what extent the mistake of fact doctrine could be accepted as a defence, especially if autonomous technology is seen as inherently unpredictable.

In the context of State responsibility, the most problematic issue might be the perennial difficulty of attribution, but this is equally true for all cyber operations, irrespective of the degree of autonomous functionality in the capabilities used.[36] Otherwise, the doctrine of State responsibility would seem to be quite capable of addressing increases in autonomy. Even the prospect of granting some degree of legal personality to autonomous systems would not appear to fundamentally disrupt the exiting law.

Distinct from the law of State responsibility, in international criminal law, knowledge and intent are decisive concepts. The mental state of operators with respect to harm caused by autonomous systems is more likely to be negligence, recklessness or *dolus eventualis*, which the existing international criminal law paradigm does not address sufficiently. Also, there are difficulties identifying the perpetrator of the *actus reus* of a war crime where a large number of individuals have an impact on the behaviour of a cyber capability. However, the realities of trying war crimes may, in some cases, mitigate these challenges. Furthermore, systematic misuse of a technology by a State might well be better addressed by the law of State responsibility than international criminal law.[37] The responsibility of commanders presents some additional challenges. It is not controversial that commanders can be held criminally liable if they use an autonomous capability to commit the *actus reus* of a crime with the requisite intent or knowledge, or if they control the will of another

36 Samuli Haataja, 'Autonomous Cyber Capabilities and Attribution in the Law of State Responsibility', this volume, ch 11.

37 Abhimanyu George Jain, 'Autonomous Cyber Capabilities and Individual Criminal Responsibility for War Crimes', this volume, ch 12.

person who in turn satisfies the elements of a crime; likewise, commanders can be held criminally liable if they assist in the commission of a crime by another person. The applicability of the doctrine of command responsibility to the relationship between a person and an autonomous capability is more controversial, however, but could hold some potential with the necessary adjustments and interpretative refinements.[38]

V
BY WAY OF CONCLUSION

It is difficult to sum up a book that contains chapters as diverse and thoughtful as the ones that contributors to this book have offered. Perhaps it could be said, with apologies to Mark Twain, that the reports of the sky falling have been greatly exaggerated. Autonomous functionality in cyber capabilities increases the complexity of the legal assessment of the performance and effects of such capabilities. But it is doubtful whether a complete regulatory paradigm shift would be necessary or desirable. A better understanding of the way in which the law could be interpreted to apply to those capabilities would, however, be helpful. From that perspective, it would be beneficial to maintain and deepen an interface between discussions about the legal implications of the use of cyber capabilities and autonomous capabilities, as the overlap of these technological paradigms is only likely to intensify. If it will prove to inspire an active dialogue, greater understanding and convergence, the book at hand has fulfilled its first and foremost purpose.

38 Russell Buchan and Nicholas Tsagourias, 'Autonomous Cyber Weapons and Command Responsibility', this volume, ch 13.

Concepts and Frameworks

Chapter 2

The Concept of Autonomy

Tim McFarland [1]

I
INTRODUCTION

This chapter discusses the notion of autonomy as it applies to software and cyber-physical systems. The purpose is to identify and explain those aspects of autonomous cyber capabilities which bear some significance to the application of relevant bodies of international law. In that respect, the chapter expands upon the outline of various conceptions of autonomy presented in a working paper produced by the NATO Cooperative Cyber Defence Centre of Excellence.[2]

In the debate about regulating the development and use of autonomous systems, much confusion has resulted from different commentators employing, either explicitly or implicitly, different criteria for describing a system as 'autonomous', causing their analyses to vary considerably depending on the criteria adopted. Rather than risk adding to that confusion, the analysis herein takes a more direct approach. The first substantive section below (Section II) presents a study of the efforts by developers of software and cyber-physical systems to impart certain

1 The author wishes to thank all who provided feedback on earlier drafts of this chapter, in particular Thomas Burri, whose insightful observations significantly improved the manuscript.

2 Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) 7–11 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>.

capabilities to those systems, including an overview of the capabilities being sought and the technologies being employed in their pursuit. Based on that material, Section III defines and discusses two legally relevant properties of autonomous software systems. Sections IV and V discuss the ways in which the roles of human and software system may vary in operations involving autonomous software.

The discussion is framed largely in abstract terms. It is not a detailed case study of a specific software application, development technology or device, although numerous examples are employed where appropriate. Rather, it addresses an abstract phenomenon, being the capacity of a software application or a cyber-physical system to operate autonomously, however that capacity may be achieved and whatever the physical form (or lack thereof) of the system involved. Indeed, much of this discussion is as applicable to a study of autonomous robots as it is to a study of autonomous software systems: a software-based control system built into a robot and constrained to interact with the outside world through the hardware components of that robot can do so autonomously just as a software-only autonomous system can interact directly with the software and hardware entities in its environment. Accordingly, frequent reference is made to autonomous 'systems', rather than specifically to software, robots, machines, weapons, or other devices. An abstract approach is useful, even necessary, given that the specific technologies which enable autonomy in artificial systems are developing rapidly. It is desirable that the findings presented here, and any legal reasoning based on them, remain useful as the underlying technologies progress.

II
TECHNICAL ASPECTS OF AUTONOMY[3]

'Autonomous', in the context of cyber capabilities, is not a term selected by lawyers or philosophers; it was selected by scientists and engineers to describe a desired outcome of their work on software and hardware

3 Parts of the discussion in this and subsequent sections are adapted from an earlier article by the present author which discussed autonomy in robotic control systems; see Tim McFarland, 'Factors Shaping the Legal Implications of Increasingly Autonomous Military Systems' (2015) 97 International Review of the Red Cross 1313.

systems. It is a property of a technological system, a degree of which has been achieved in some systems in use today and greater degrees of which are the goal of research and development programs in relevant fields. This investigation therefore begins with a brief study of the development outcome which the term 'autonomous' was selected to describe. Though most participants in the debate about regulating military use of autonomous systems are by now well aware of the essential concepts involved and the capabilities of existing systems, it bears going back to the factual basics in order to establish common ground, before extending the discussion to the implications of autonomy for software systems and further into its legal consequences.

The essence of autonomy in a software context is, as in other contexts, self-regulation or self-governance. It is a concept which may be viewed in two ways: that the system in question generates the rules by which it operates in its environment, and that no other entity generates the rules by which the system operates. Those two sides of the autonomy coin are equally important and equally worthy of further discussion, given the confusion they have caused at various stages of the debate about the regulation of autonomous military systems. They are discussed below in terms of two relationships: that between the system and its task or environment, and that between the system and its operator.

A SYSTEM-ENVIRONMENT RELATIONSHIP

Abbass, Scholz and Reid provide a useful conceptualisation of autonomy in a technical context:

> Foundationally, autonomy is concerned with an agent that acts in an environment. However, this definition is insufficient for autonomy as it requires persistence (or resilience) to the hardships that the environment acts upon the agent. An agent whose first action ends in its demise would not demonstrate autonomy. The themes of autonomy then include agency, persistence and action. … Action may be understood as the utilisation of capability to achieve intent, given awareness.[4]

4 Hussein A Abbass, Jason Scholz and Darryn J Reid, 'Foundations of Trusted Autonomy: An Introduction' in Hussein A Abbass, Jason Scholz and Darryn J Reid (eds), *Foundations of Trusted Autonomy* (Springer 2018) 1.

Likewise, Franklin and Graesser explain that

> An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.[5]

('Agent', in software terms, refers to a software entity which acts autonomously in an environment in pursuit of some goal for which it was designed.[6])

In other words, a software entity is autonomous when it possesses both an encoded representation of a goal (described by Abbass, Scholz and Reid above as an 'intent'; essentially, a representation of the task which the person or organisation operating the software wants it to complete) and the ability to act within its environment in furtherance of that goal. Acting within an environment in furtherance of a goal in turn requires 'awareness' (the ability to sense the environment and changes therein), 'capability' (a facility for effecting desirable changes in the environment and resisting or correcting undesirable changes) and, implicitly, a means to select the capability which best serves the software's purpose in response to a given change in the environment (where that selection process may be characterised as reasoning, choice, decision-making, and so on). Overall, autonomous behaviour may be viewed as the process of aligning the software entity's awareness with its goal:

> If 'capability' is defined as anything that changes the agent's awareness of the world (usually by changing the world), then the error between the agent's awareness and intent drives capability choice in order to reduce that error. Or, expressed compactly, an agent seeks achievable intent.[7]

Notably, that conceptualisation of autonomy lacks sharply defined thresholds, and so invites consideration of whether all software which is created for a purpose and which can detect and respond to changes in its environment without human intervention may be considered to be autonomous. On the one hand, the behaviours described above (pursuing goals, sensing

5 Stan Franklin and Art Graesser, 'Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents' in Jörg P Müller, Michael J Wooldridge and Nicholas R Jennings (eds), *Intelligent Agents III: Agent Theories, Architectures, and Languages* (Springer 1997) 25.
6 ibid 24.
7 Abbass, Scholz and Reid (n 4) 1.

and effecting changes in an environment) are quite obviously matters of degree, abilities which may be exhibited in different measure by different systems in different environments. On the other hand, many authors take the view that autonomous systems are a rigidly defined class of system which may be distinguished from non-autonomous or merely automated systems on the basis of objective criteria. This chapter rejects that latter view,[8] and chooses to focus directly on the underlying capability, being a capacity for self-management in some degree. It is acknowledged, however, that the focus of regulatory interest is on systems with higher degrees of autonomy, and that is also the focus of this chapter.

The notion of a device which can automatically respond to changes in its environment in order to fulfil its purpose is much older than software or robots. James Watt's centrifugal governor for steam engines, dating to approximately 1788, is often cited as an example.[9] Likewise, a thermostat, whether implemented in software or hardware, exhibits the essence of this behaviour. Indeed, any device which employs a closed loop control system of some sort might arguably qualify. In the context of weapon systems, land mines are sometimes described as exhibiting a degree of autonomous behaviour in that, once emplaced, they are able to 'select' targets (via a pressure sensor or other trigger mechanism) and 'attack' (explode) without further human intervention.[10]

However, the use of 'autonomous' as a description of a self-governing device is much more recent and, in practice, the term is not generally used in reference to simple devices like governors and thermostats. Its use came about with efforts to develop systems which can perform their tasks unattended in increasingly complex circumstances (whether that complexity be in the task or the environment). While there is no precise threshold, the term is generally associated with self-governing machines whose task requires higher levels of 'algorithmic and hardware sophistication'[11] and the ability to operate in the face of uncertainty:

> Autonomy is more or less understood as a requirement for operating in complex environments that manifest uncertainty; without uncertainty relatively straightforward automation will do, and

8 For more detail, see Tim McFarland, *Autonomous Weapon Systems and the Law of Armed Conflict* (Cambridge University Press 2020) ch 3.

9 See, eg, HW Dickinson, *James Watt: Craftsman and Engineer* (Cambridge University Press 2010) 153ff.

10 See, eg, Kenneth Anderson, Daniel Reisner and Matthew Waxman, 'Adapting the Law of Armed Conflict to Autonomous Weapon Systems' (2014) 90 International Law Studies 386, 388.

11 Darryn J Reid, 'An Autonomy Interrogative' in Hussein A Abbass, Jason Scholz and Darryn J Reid (eds), *Foundations of Trusted Autonomy* (Springer 2018) 365.

> indeed the autonomy is generally seen here as being predicated on some form of environmental uncertainty.[12]

Thus, in terms that are perhaps more fitting for legal purposes, a self-governing system is more likely to be described as 'autonomous' where humans are not reasonably able to precisely foresee the exact sequence of steps that the system must take in order to complete its assigned task (or, equivalently, cannot foresee all events that will transpire when the system is activated). The term is used when the high level goal of deploying the system is defined in advance but not necessarily every low level step that will be completed in its pursuit. That is, some reliance is placed upon an autonomous system to select the appropriate response to changes in its environment from among the possible responses supported by its capabilities.

Some advanced software systems being developed today, known by the recently coined term 'cyber reasoning systems'[13] ('CRS') demonstrate this quality. Referring primarily to systems such as those which have been deployed in DARPA's 'Cyber Grand Challenge', CRS 'combine various tools, techniques and expert knowledge to create fully autonomous systems that perform automated vulnerability detection, exploit generation and software patching in binary software without human intervention'.[14] They comprise multiple sub-systems with both offensive and defensive roles. These sub-systems search for vulnerabilities in adversaries' systems, develop tools for exploiting those vulnerabilities and conduct attacks against them while simultaneously searching for and repairing vulnerabilities in friendly systems under their protection and intercepting attacks against those systems launched by adversaries.

Deployed in competition with other CRS (or, hypothetically, against any intelligent adversary), human operators could not intervene to manage the use of each of those capabilities. The reasoning required for that purpose had to be encoded into each CRS such that it could respond in an appropriate way to each change in its environment during the competition: 'CRSs had to make strategic decisions throughout the game: Which

12 Abbass, Scholz and Reid (n 4) 7.

13 See, eg, Raytheon Technologies, 'Cyber Reasoning Systems: Automating the Detection and Patching of Vulnerabilities' <https://www.raytheon.com/cyber/capabilities/reasoning> accessed 3 November 2020.

14 Teresa Nicole Brooks, 'Survey of Automated Vulnerability Detection and Exploit Generation Techniques in Cyber Reasoning Systems' (2019) 857 Advances in Intelligent Systems & Computing 1083, 1083. It is notable that the quoted passage uses 'autonomous' and 'automated' interchangeably. This is consistent with the view that both words refer to the same underlying capacity of a machine or software system for self-management.

binaries to patch? Which patches to deploy? Which teams to attack, and with which exploits? Where should limited resources be spent?'[15]

B SYSTEM-OPERATOR RELATIONSHIP

The motives for developing autonomous systems vary widely. Often, they amount to a desire to exceed human capabilities in some way, such as to perform a task with greater speed, accuracy, precision or over a longer period (speed being a particular concern in cyber contexts, where response times might need to be on the order of milliseconds). The goal may alternatively be to remove humans from dangerous situations or hostile environments. Whatever the specific operational concern, the underlying technical need is for a system that can manage its own operation in a relevant way, rather than rely on interaction with a human operator. This section discusses the nature of the system-operator relationship in respect of autonomous systems and the technical means by which it is achieved. The two subsections cover the basic and more complex aspects of autonomous software respectively.

1 *Basic Aspects of Autonomous Software*

The 'user-facing' characteristics of autonomous systems are perhaps the most significant for the purposes of a legal analysis. They are captured in some definitions employed in technical and military operational studies. For example, Lin, Bekey and Abney define autonomy as

> [t]he capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time.[16]

For reasons explained below, the phrase 'without any form of external control' should be read carefully, in the sense of 'operator interaction'. A more succinct definition is provided by Goodrich and Schulz:

15 Thanassis Avgerinos and others, 'The Mayhem Cyber Reasoning System' (2018) 16(2) IEEE Security & Privacy 52, 56.

16 Patrick Lin, George Bekey, and Keith Abney, 'Autonomous Military Robotics: Risk, Ethics, and Design' (California Polytechnic State University, Ethics + Emerging Sciences Group, 20 December 2008) 4 <http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA534697>.

> A system with a high level of autonomy is one that can be neglected for a long period of time without interaction.[17]

(Implicitly, this refers to neglect while the system is operating; in other words, the system can operate for extended periods without requiring human attention.) Both of these definitions make it clear that a capacity for autonomous operation is reflected in the relationship between an autonomous system and its operator. That relationship is not severed; it remains one in which responsible humans (such as the system's designer and operator) exert, or have exerted, a form of control over the system by defining the task to which it is set and the manner in which it is able to interact with its environment. According to the United States Department of Defense:

> Autonomy is a capability (or a set of capabilities) that enables a particular action of a system to be automatic or, within programmed boundaries, 'self-governing.'[18]

Two phrases in that definition, 'particular action of a system' and 'within programmed boundaries', are important. That a system constructed by humans is able to operate without human interaction does not mean there are no human-imposed restrictions on the system's behaviour. Autonomous control essentially describes an approach to constraining or guiding the behaviour of a system in circumstances where direct or real-time human interaction is infeasible or undesirable. In a robotics context (although equally applicable to software-only systems):

> *Autonomous* means having the power for self government. *Autonomous controllers* have the power and ability for self governance in the performance of control functions. They are composed of a collection of hardware and software, which can perform the necessary control functions [on behalf of the operator], without external intervention, over extended time periods.[19]

17 Michael A Goodrich and Alan C Schultz, 'Human-Robot Interaction: A Survey' (2007) 1 Foundations and Trends in Human-Computer Interaction 203, 217.

18 US Department of Defense, 'DoD Directive No. 3000.09: Autonomy in Weapon Systems' (21 November 2012) 1. Note that this Directive provides that the policies it outlines do not apply to cyber systems. However, this chapter takes the position that the cited definition of the abstract notion of 'autonomy' is as applicable to pure software systems as it is to software-driven hardware.

19 Panos J Antsaklis, Kevin M Passino and S J Wang, 'An Introduction to Autonomous Control Systems' (1991) 11(4) IEEE Control Systems 5 (emphasis in original).

A software application, whether or not it exhibits any degree of autonomous behaviour, is a tool. It is an implement used by a person or group to accomplish some goal. Its operation must, therefore, be directed by the operator toward that goal. Although autonomous systems are often described as being able to operate 'without external control', as in some of the definitions given above, that can be misleading in a discussion of their legal characterisation. A software application's lack of interaction with an operator while it is running does not mean that its behaviour has not been defined by a person. Rather, it means that the intended behaviour was defined in advance of the software's activation and is then enforced by the code itself.

The behaviour of an autonomous system ultimately depends upon actions of people in relevant positions, notably its designer and operator, due to the nature of computers and software. Autonomous software entities (to return the focus to the main subject matter of this book) are essentially sets of human-written instructions executed by human-constructed computing devices, including ordinary general-purpose computers, control units governing industrial processes, networking equipment such as routers or switches, 'internet of things' devices that one might not readily recognise as computers, and so on. Although they may be highly specialised in design and purpose, such devices are nevertheless forms of ordinary stored-program computers, the defining characteristic of which is that instructions entered by a human programmer are stored in the machine's memory and drawn upon to govern its operation. Barring a major technological shift, tomorrow's autonomous systems will employ essentially the same technology.

The fact that even very complex programs are just sets of pre-defined instructions is often obscured in discussions about sophisticated autonomous behaviour, and indeed it is not always apparent to an observer that a complex system operating without human interaction is merely executing instructions rather than behaving independently of human influence. This is at least partly attributable to the use of software instructions which define how the software should respond directly to changes in its environment rather than only to instructions from a human operator. For example, even systems with only very simple and limited capabilities are often driven by programs with instructions of the form

```
if <X happens> then <do action A> else <do action B>
```

If 'X' is something other than an input directly from an operator then such instructions can create the impression that the system itself is

'choosing' between two alternative courses of action, when in fact the choice was made in advance by the person or organisation responsible for writing the program; the expression of that party's will was merely waiting within the system's memory for the previously determined trigger to be detected. For example, a hypothetical cyber weapon might be encoded with a description of a specific type of database system which contains sensitive information belonging to an adversary, along with instructions describing a process for seeking, identifying and damaging systems matching that description. If such a cyber weapon detects a candidate system while scanning a network, an instruction like

```
if <signature of detected system matches encoded description
of target> then <connect to and damage the system> else
<keep searching>
```

might create the appearance of the cyber weapon itself selecting targets, when actually the targets and the conditions under which they would be attacked were selected in advance by the system developers.

This reference to computing technology is included here because repeated references to autonomous systems having the capacity for 'choice' or 'truly autonomous' operation in the regulatory debate so far are potentially misleading. No computer is able to choose for itself whether or not to run a program stored in its memory, or to exercise discretion about whether or not to execute a particular instruction within a program. Any such appearance of choice can only be the result of other instructions embedded in software. Fundamentally, the only function of a computer is to run whatever software is installed on it.

Autonomy, in a technical sense, is simply the ability of a system to behave in a desired manner, or achieve the goals previously imparted to it by its operator, without needing to receive the necessary instructions from outside itself on an ongoing basis. It is, of course, of most significance where the desired behaviour requires the system to respond to changes in its environment or to operate in circumstances wherein humans might be unable to intervene. For simple tasks in well-understood environments, that might be achievable with a simple, static step-by-step set of instructions. As one example, many firewalls fit this description. A firewall is a special-purpose computing device which is positioned at the edge of a network along the path taken by network traffic (data sent between computers) travelling into and/or out from that network. Its task is to examine each piece of traffic that attempts to pass through and decide, according

to a set of programmed requirements, whether that traffic is to be allowed through or blocked, in order to protect the network to which the firewall belongs. Firewalls are very common devices, being positioned at the edges of most corporate and government networks as well as being embedded within many consumer devices including personal computers and home internet routers. In the ordinary course of events, users are unlikely to need to interact with them, or even know they are present. However, the relatively static nature of the task undertaken by most firewalls places them outside the scope of autonomous systems which are of regulatory interest.

For more complex tasks, or tasks done in less predictable environments, autonomous operation might require that more advanced capabilities be encoded: to detect changes in the environment, to select a course of action from several possibilities in response to those changes, perhaps to recognise when a goal is not achievable, and so on. Some intrusion prevention systems ('IPS') might fall into this category. An IPS is a network security system which might incorporate several simpler sub-systems capable of performing a range of security-related functions along with some logic to control and co-ordinate those sub-systems according to the needs of the network's operator.[20] For example, an IPS might include a firewall along with the ability to assess whether network activity might be malicious, reconfigure the firewall to block that activity, and perhaps repair damage caused by the malicious activity, such as by removing virus-infected email attachments or similar measures, all without requiring human intervention. Regardless of the complexity involved, though, autonomous software systems remain merely computer programs, written by human beings.

Understanding software autonomy as a form of control rather than as the absence of control is a necessary step toward identifying its legal implications. That it is a form of control is relatively easy to see when the behaviour of the system corresponds directly to software instructions entered by a human programmer. It is more difficult to see in the case of advanced software which, beyond simply operating without human intervention, may appear to exhibit some behaviour which has not been explicitly programmed by a person. Objections to development of highly autonomous military systems based on fears that they may select the wrong target or otherwise act in undesirable ways generally refer either explicitly or implicitly to this type of system.[21]

20 Paloalto Networks, 'What is an Intrusion Prevention System?' <https://www.paloaltonetworks.com/cyberpedia/what-is-an-intrusion-prevention-system-ips> accessed 20 December 2020.

21 Which is not to imply that such fears are the only basis of objections to autonomous military systems.

2 *More Advanced Aspects of Autonomous Software*

Autonomous cyber weapons, like many other military systems for which autonomy is seen as an advantage, must be able to complete complex tasks in hostile and dynamic environments against adversaries who are able to learn and adapt. The exigencies of combat operations arguably make adaptability a more critical requirement for military systems than for those in civilian applications, as well as a greater challenge. The control functions of an autonomous software entity must be able to ensure the entity operates at a sufficiently high standard when there is a very high degree of uncertainty in the environment in which it operates. Behaviour of adversaries, active and passive defences, damage to systems and networks on which the entity operates and other events may all interfere with the operation of a cyber weapon such that some corrective action is needed outside of what might have previously been encountered in the course of the task being undertaken. In the case of an autonomous system, that corrective action must be initiated by the system itself rather than by a human operator. That is, when the system encounters a change in its environment such that the algorithm the system is using is no longer suitable, the system must be able to adapt that algorithm in order to achieve its goal. This type of capability can be found in some software systems in use today. For example, some radar systems offer 'constant false alarm rate' detection, wherein a radar system can adjust its own behaviour to compensate for varying levels of background noise and interference which might otherwise mask the presence of a target.[22] Likewise, some computer worms and viruses employ 'polymorphic' code, or code which can rewrite itself without changing its core functionality, in order to evade security systems which might have been configured (or have adapted themselves) to detect the worm or virus in its previous form.[23] Broadly, a software system which is able to alter its behaviour in response to changing circumstances is referred to as a 'self-adaptive system', or as software which employs an 'adaptive algorithm'. It is one source of the behaviours which define the software systems that are of most interest in the context of a discussion about the legal implications of software autonomy. Essentially,

22 Christian Wolff, 'False Alarm Rate' (*Radartutorial.eu*) <https://www.radartutorial.eu/01.basics/False%20Alarm%20Rate.en.html> accessed 20 December 2020.

23 Trend Micro, 'Polymorphic virus' <https://www.trendmicro.com/vinfo/us/security/definition/Polymorphic-virus> accessed 20 December 2020.

> [s]elf-adaptive software evaluates its own behavior and changes behavior when the evaluation indicates that it is not accomplishing what the software is intended to do, or when better functionality or performance is possible.[24]

Importantly, such adaptation does not alter 'what the software is intended to do' (its purpose as defined by its human designers), although it may alter the low level steps that are taken in the course of fulfilling that purpose.

When utilising adaptive algorithms which enable a system to tune its own behaviour, the system may be operating as it was designed to even if the precise rules by which it is operating at a given time were not explicitly provided by a human operator (and may not even be precisely known to a human operator). Essentially, adaptive software employs higher level logic built into the software itself to generate whatever lower level operative rules are required as circumstances change. That higher level logic represents the operator's intent, and by altering its behaviour according to those higher level rules, the system is behaving in accordance with that intent.

Although adaptive techniques enable a system to alter its behaviour to an extent, their usefulness is limited by complexity. They rely on the system designer having a high degree of *a priori* knowledge about the system, its task and the environmental changes and disruptions that might be encountered, such that those factors can be mathematically modelled and represented in the software. In highly complex, poorly understood or unpredictable environments, or where the task to be completed is complicated, or even where the software itself is very complicated, it is not necessarily feasible to construct such a model in sufficient detail. In that case, another class of algorithm is likely to be employed.

Nonparametric algorithms, or those which do not rely on detailed mathematical models of the task or environment, are 'based on the use of more general models trained to replicate desired behaviour using statistical information from representative data sets.'[25] That is, the software is provided with data representing situations that might be encountered in its intended operating environment, along with the desired responses to

24 Robert Laddaga, Paul Robertson, and Howie Shrobe, 'Introduction to Self-adaptive Software: Applications' in Robert Laddaga, Paul Robertson, and Howie Shrobe (eds), *Self-Adaptive Software: Applications* (Springer 2003) 1.

25 Anthony Zaknich, *Principles of Adaptive Filters and Self-learning Systems* (Springer 2005) 3.

those stimuli, and it is 'trained' to generalise from the provided training data to arrive at an algorithm that will be effective in practice. This is a diverse field which draws on a range of techniques that enable software to operate in environments that are too complex or unpredictable, or about which too little is known, to be susceptible to the mathematical modelling required by traditional techniques. Generally, this 'intelligent' software works by emulating various aspects of biological cognitive processes, based on the premise that biological entities are often able to operate effectively with incomplete knowledge, in complex and ambiguous environments.[26] The specific techniques employed are many, and the details are beyond the scope of this text; the most well-known techniques, which may be employed separately or in combination, are perhaps neural networks, fuzzy logic and genetic algorithms.[27] The relevant advantage which all such techniques afford to the system designer is that they do not require detailed foreknowledge of all combinations of circumstances which the software entity may encounter once it is in operation. They allow the system designer to employ heuristics, approximation techniques and optimisation techniques to adapt the software's behaviour to circumstances which cannot be precisely foreseen.

A related issue is that of systems that continue to 'learn' after being put into operation.[28] Learning, in this context, refers to the process of finding a generalised model which accounts for a set of observations, so that the model can be employed when similar observations are made in the future.[29] Rather than just responding to unexpected changes in its environment, a learning system is one that can improve its abilities over time by adjusting its 'rules' according to accumulated experiential knowledge; that is, allow information such as the performance of the system at previous tasks to be retained and used to tune behaviour in future tasks. Online learning (being learning that happens after a system is put into operation, as opposed to offline learning which happens during a development phase) is a considerably more ambitious control technique that is useful when the complexity or uncertainty of a situation prevents *a priori* specification of an optimal algorithm.

26 See, eg, Katalin M Hangos, Rozália Lakner and Miklós Gerzson, *Intelligent Control Systems: An Introduction with Examples* (Kluwer 2004) 1.

27 See, eg, M Jamshidi, 'Tools for Intelligent Control: Fuzzy Controllers, Neural Networks and Genetic Algorithms' (2003) 361 Philosophical Transactions: Mathematical, Physical and Engineering Sciences 1781.

28 DA Linkens and HO Nyongesa, 'Learning Systems in Intelligent Control: An Appraisal of Fuzzy, Neural and Genetic Algorithm Control Applications' (2002) 143(4) IEE Proceedings — Control Theory and Applications 367.

29 William Bialek, Ilya Nemenman and Naftali Tishby, 'Predictability, Complexity, and Learning' (2001) 13 Neural Computation 2409.

It is another mechanism by which the rules by which a system operates at a given time may not be rules explicitly provided by a human operator. As with adaptive algorithms, though, generation of those rules according to the higher-level learning process is the behaviour intended by the system's operator. The learning process which governs the overall behaviour of the software must be considered to represent the operator's intent.

'Intelligent' software generally applies techniques from the broader field of artificial intelligence ('AI').[30] AI aims to understand the factors that make intelligence possible, and to employ that knowledge in creation of artificial systems that can operate in ways which, if observed in living beings, would be considered intelligent. That is, systems that can respond appropriately to changing circumstances and goals, take appropriate actions when provided with incomplete information and, if needed, 'learn' from experience.

Despite the complexity of software which relies on these advanced algorithms, it is not fundamentally distinct from simpler automated or manual processes. They are all still means of achieving some human-defined goal. Regardless of its complexity, autonomous software amounts to a set of instructions guiding a system toward such a goal. Those instructions may endow a system with a capacity for complex actions and responses, including the ability to operate effectively in response to new information encountered during operations which may not be precisely foreseeable to a human operator. However, that does not constitute independence from human control in any sense. Rather, it is best seen as control applied in a different way, in advance rather than in real time.

III
AUTONOMY AS A PROPERTY OF A SYSTEM

Autonomy is a property of a technological system which may be realised by diverse means. It does not connote the presence of a specific technology nor a particular type of device nor a certain behaviour. It is,

30 See generally, Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd edn, Prentice Hall 2009).

fundamentally, the ability of a system to perform the task assigned to it, whatever that may be, with less interaction with a human operator than a manual system would require. Implicitly, that is achieved by enabling the system to interact directly with its environment rather than have it refer every decision to a human. For the purposes of investigations in non-technical fields, software autonomy primarily affects the relationship between the system and its human operator, not the nature of the system's task nor the precise manner in which it performs that task. Two points in particular are relevant in an investigation of the legal characterisation of autonomous systems.

First, autonomous systems will perform their assigned tasks in place of human-operated manual systems, but the outcomes will not necessarily differ from those which would have been achieved had the tasks been done manually by humans. There is nothing in the concept of software autonomy that supports an inference that an autonomous system must necessarily perform a task in a different manner than would a human or team of humans performing the same task manually. Of course, one of the motivations for developing increasingly autonomous systems is to achieve superior outcomes. The persistence of systems that do not require constant human interaction; the ability to quickly integrate data from many sources; the capacity for greater speed, accuracy or precision; and the ability to take greater risk than could be taken via a manual approach; among other benefits of autonomous systems, will certainly aid both military and other operations. However, such differences, while very important operationally, are somewhat peripheral to the legal aspects of autonomy. Remotely piloted aircraft ('RPA'), for example, already allow for a high level of persistence without necessarily exhibiting any capabilities associated with a high level of autonomy and without raising the same legal questions. In assessing the legal implications of a particular development path, or a particular set of technologies, the focus must be kept on the capability of interest rather than on other capabilities that may be present in the same system. In the case of autonomous software, it is not useful to attempt to attribute specific behaviours to an application merely on the basis of it being described as having autonomy; all that one can reliably say on that basis is that the human operator's direct involvement in part or all of the software's performance of its assigned task will be reduced or removed. The mere fact of reassigning a task from a human to a computer does not necessarily alter the performance of that task.

Second, it is incorrect to describe autonomous systems as being 'independent'[31] systems that operate 'without human control'.[32] The relationship between human and software is not severed, it is only modified. Choices made by developers in the design stage will shape the behaviour of the systems they create from then on, for both technical and operational reasons. On a technical level, as explained above, everything that an autonomous system does (barring malfunctions and interference) is ultimately the result of executing a set of software instructions written by human developers. On an operational level, it is an obvious practical necessity that an autonomous system be constrained to behave consistently with its purpose. In a military context, a cyber weapon is only one tool in the hands of a State's armed forces. Its use must be in accordance with higher level plans and established procedures as well as with the capabilities and practices of other units and support structures, and the autonomous system's role would often be only one component in a larger coordinated effort. Mission planners and commanders will set goals and impose constraints which must be satisfied, and significant human effort might be expended gathering intelligence and otherwise facilitating the operation. For example, the well-known Stuxnet worm, which was used to disable centrifuges at the Natanz nuclear facility in Iran, has been described as possessing a degree of autonomous capability:

> Considering that there was very good chance that no Internet connectivity would be available (only access to the internal network), Stuxnet developers put all of its logic in the code without the need of any external communication. As such the Stuxnet was an autonomous goal-oriented intelligent piece of software capable of spreading, communicating, targeting and self-updating ….[33]

Although much is unknown about the use of Stuxnet, security researchers from Symantec have outlined a possible attack scenario based on their analysis of the worm's capabilities, which demonstrates a plausible degree of human involvement:

31 Markus Wagner, 'Taking Humans Out of the Loop: Implications for International Humanitarian Law' (2011) 21(2) Journal of Law, Information & Science 155, 159.

32 Gary E Marchant and others, 'International Governance of Autonomous Military Robots' (2011) XII Columbia Science & Technology Law Review 272, 273.

33 Stamatis Karnouskos, 'Stuxnet Worm Impact on Industrial Cyber-Physical System Security' in *IECON 2011 — 37th Annual Conference of the IEEE Industrial Electronics Society: Proceedings* (IEEE 2011) 4492.

> First, the attackers needed to conduct reconnaissance. As each PLC [programmable logic controller; the device targeted by Stuxnet] is configured in a unique manner, the attackers would first need the ICS's [industrial control system; the network environment in which the PLCs exist] schematics. These design documents may have been stolen by an insider or even retrieved by an early version of Stuxnet or other malicious binary. Once attackers had the design documents and potential knowledge of the computing environment in the fawcility, they would develop the latest version of Stuxnet. … Attackers would need to setup a mirrored environment that would include the necessary ICS hardware, such as PLCs, modules, and peripherals in order to test their code. The full cycle may have taken six months and five to ten core developers not counting numerous other individuals, such as quality assurance and management. In addition their malicious binaries contained driver files that needed to be digitally signed to avoid suspicion. The attackers compromised two digital certificates to achieve this task. The attackers would have needed to obtain the digital certificates from someone who may have physically entered the premises of the two companies and stole them, as the two companies are in close physical proximity. To infect their target, Stuxnet would need to be introduced into the target environment. This may have occurred by infecting a willing or unknowing third party, such as a contractor who perhaps had access to the facility, or an insider.[34]

In another, albeit robotics-related, example, Boothby describes the level of human involvement that would be required in conducting an attack with an autonomous aircraft:[35]

> A flight plan will have been prepared and filed by a person who will decide on the geographical area that is to be searched, the time period within which the search may take place, the areas where the aircraft may loiter and for how long, and that person will programme these important requirements into the flight control software. The platform will be fuelled by a person thus

34 Nicolas Falliere, Liam O Murchu and Eric Chien, 'W32.Stuxnet Dossier' (v1.4, Symantec Security Response, February 2011) 3 <https://archive.org/details/w32_stuxnet_dossier>.

35 William Boothby, 'How Far Will the Law Allow Unmanned Targeting to Go?' in Dan Saxon (ed), *International Humanitarian Law and the Changing Technology of War* (Martinus Nijhoff 2013) 56.

> defining the maximum endurance of the mission. Operational planners will decide what weapons will be carried and how they are to be fused, and stores will be loaded by people before take-off. The sensors on which the autonomous aspect of the mission depends will have been designed and built by people and will be controlled by similarly designed software. Those designers and/or the mission planners will have prescribed the level of mechanical recognition that is to be achieved before an object is recognised as a target and, thus, before an attack will be undertaken.

It may be assumed that analogous tasks would be performed by humans in relation to other operations involving autonomous software. In these ways a human hand always provides some degree of guidance despite a possible lack of direct supervision. The Defense Science Board of the United States Department of Defense (writing about autonomous vehicles) expresses the dependence of autonomous systems on humans more generally:

> It should be made clear that all autonomous systems are supervised by human operators at some level, and autonomous systems' software embodies the designed limits on the actions and decisions delegated to the computer. Instead of viewing autonomy as an intrinsic property of an unmanned vehicle in isolation, the design and operation of autonomous systems needs to be considered in terms of *human-system collaboration*.[36]

IV
HUMAN-SYSTEM COLLABORATION

Despite that, on the technical level, autonomous capabilities are inherently a matter of degree, it is common in the non-technical literature on autonomous systems to attempt to categorise particular systems according to one taxonomy or another. One popular classification scheme, often cited in discussions about autonomous weapon systems, distinguishes

36 Defense Science Board, 'The Role of Autonomy in DoD Systems' (US Department of Defense Task Force Report, July 2012) 1–2.

between 'automatic', 'automated', and 'autonomous' systems. These terms have been used somewhat differently by different authors, but the essential distinctions are as follows:[37]

- 'Automatic' refers to very simple devices which perform well-defined tasks and may have the ability to respond in pre-set ways to external stimuli. Systems which can operate unattended but which have little or no ability to receive and act on feedback from their environment are sometimes described as 'automatic'.

- 'Automated' may be used synonymously with 'automatic' but may also refer to systems which follow more complex sets of rules in normal operation and in responding to disturbances, such that they can perform more complex tasks or operate in more complex environments. Examples include automated telephone support lines that can respond in limited ways to various queries, or some existing weapon systems.

- The varying uses of the term 'autonomous' among authors reflects the uncertainty that surrounds the nature of these new technologies. The general view is that autonomous systems go beyond automated systems in some way, but the precise criteria vary. Some authors describe systems as 'autonomous' when they exhibit some ability to adapt their own behaviour in response to changing circumstances.[38] Others use the term to indicate that some threshold level of complexity in the system, its task or its operating environment has been reached.[39] Still others say autonomous systems are those with some degree of 'independence' from their human operators.[40] A further subdivision within this category is between 'semi-autonomous' and 'fully autonomous' systems. The claimed difference is that fully autonomous systems are those which are designed to operate

37 Paul Scharre, 'Autonomous Weapons and Operational Risk' (Center for a New American Security, February 2016) 12 <https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk>.

38 Kenneth Anderson and Matthew C Waxman, 'Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can' (Hoover Institution, 9 April 2013) 6 <https://www.hoover.org/sites/default/files/uploads/documents/Anderson-Waxman_LawAndEthics_r2_FINAL.pdf>.

39 Rebecca Crootof, 'The Killer Robots Are Here: Legal and Policy Implications' (2015) 36 Cardozo Law Review 1837, 1854.

40 Chantal Grut, 'The Challenge of Autonomous Lethal Robotics to International Humanitarian Law' (2013) 18(1) Journal of Conflict & Security Law 5.

> entirely without human involvement once activated while semi-autonomous systems would require some form of active human involvement in relation to some or all functions.[41] A range of views have been expressed over whether fully autonomous systems would entirely disallow human involvement, or simply not require it, and about the extent of human involvement required in semi-autonomous systems.

Some commentators create finer distinctions within each of those levels depending on the degree of necessity of the human operator's contribution, the realistic possibility for successful human intervention once the system is deployed, and so forth.[42]

Attempts to define taxonomies of autonomy are further complicated by differing views on whether categories should be based on the degree and type of human interaction with the system, or the complexity of the system and its behaviour. The two variables are both plausible bases for categorisation (if one believes that categorisation is appropriate), but each captures only one aspect of autonomy, and any simple discrete taxonomy fails to reflect the fact that the levels of autonomy exhibited by existing and proposed systems may be expected to vary in complex ways. Capacities for autonomous operation vary widely, as do the ways in which tasks are allocated between an operator and an autonomous system, and the behaviour of a system may be expected to change according to both the specific task being performed and the current state of the environment in which the system is operating. Establishing the relative degrees of control exercised by a human operator and a software system in respect of a particular action for legal or other purposes may be a complex process. The Defense Science Board offers this view of the variability of degrees of autonomy from a cognitive science perspective:

> Cognitively, system autonomy is a continuum from complete human control of all decisions to situations where many functions are delegated to the computer with only high-level supervision and/or oversight from its operator. Multiple concurrent functions may be needed to evince a desired capability, and subsets of functions may require a human in the loop, while other functions can

41 See, eg, US Department of Defense (n 18) 13, 14.

42 See, eg, Frank O Flemisch and others, 'The H-Metaphor as a Guideline for Vehicle Automation and Interaction' (NASA Technical Memorandum 003-212672, December 2003) <https://ntrs.nasa.gov/citations/20040031835>.

> be delegated at the same time. Thus, at any stage of a mission, it is possible for a system to be in more than one discrete level simultaneously.[43]

The complexity of defining a degree of system autonomy is well demonstrated by consideration of the various dimensions along which autonomous behaviour may vary.

Rather than simply requiring more or less guidance from a human operator, a software entity may require guidance in different forms:[44] the software might determine available options in some situation and rely on a human to select one; the software might recommend a particular option or not; it might begin to undertake a course of action and give an operator a chance to override that choice; it might complete the whole task and report back (or not) afterwards.

A human may be required to play the role of a hands-on 'operator' in some cases and a hands-off 'supervisor' in others. They may alternatively be more of a 'collaborator', sharing tasks with an autonomous entity, with the allocation of specific tasks being negotiated between them, or perhaps controlled by a third party.[45] In a collaborative scenario, either the software or the human might have direct control of a specific task at a specific time with the other party assisting.

Just as the activities of an autonomous software entity would generally form one part of a larger coordinated operation, autonomous capabilities are likely to exist in specific sub-systems performing specific functions rather than be applied to the system as a whole. A cyber weapon might be trusted to locate and identify potential targets autonomously but be required to seek human confirmation before attacking them. The level of autonomy displayed by an entity might therefore vary according to the specific task being performed during an operation,[46] enlivening the possibility that a system may be operating at more than one 'level' of autonomy simultaneously, with respect to different tasks.

43 Defense Science Board (n 36) 4.

44 See, eg, Thomas B Sheridan and William L Verplank, 'Human and Computer Control of Undersea Teleoperators' (Massachusetts Institute of Technology, Man-Machine Systems Laboratory, 14 July 1978) 8–17 <https://apps.dtic.mil/dtic/tr/fulltext/u2/a057655.pdf>; NIST Engineering Laboratory, 'Autonomy Levels for Unmanned Systems' (National Institute of Standards and Technology, 6 June 2010) <http://www.nist.gov/el/isd/ks/autonomy_levels.cfm>.

45 See, eg, Jean Scholtz, 'Theory and Evaluation of Human Robot Interactions' in *HICSS'03 — Proceedings of the 36th Hawaii International Conference on System Sciences* (IEEE Computer Society 2003); Marti A Hearst, 'Trends & Controversies: Mixed-Initiative Interaction' (1999) 14(5) IEEE Intelligent Systems 14.

46 This phenomenon is emerging in some current weapon systems: Rain Liivoja, Kobi Leins and Tim McCormack, 'Emerging Technologies of Warfare' in Rain Liivoja and Tim McCormack (eds), *Routledge Handbook of the Law of Armed Conflict* (Routledge 2016) ch 35 s 2.1.

Finally, the level of autonomy exhibited by a system might vary with circumstances that arise during an operation. A system that can operate entirely unassisted in normal circumstances might refer a decision to an operator if an unexpected problem arises, or if an unanticipated opportunity presents itself.

The US Department of Defense summarises the variability inherent in autonomous software operation: 'The key point is that humans and computer agents will interchange initiative and roles across mission phases and echelons to adapt to new events, disruptions and opportunities as situations evolve.'[47]

V
CONCLUSION

Autonomous software entities, whatever advanced capabilities they possess, remain exactly that: software. They are tools wielded by human operators, sequences of human-written instructions executed by human-constructed computers for human-defined purposes, qualitatively identical to any other software. They will generate the rules by which they operate if that is what they are programmed to do, but in doing so, their ties to their developers are not severed; the process of generating those rules in pursuit of their designed purpose is the behaviour that their developers intended. Despite that the core meaning of autonomy is 'self-governance', software cannot be regarded as a root cause of its own behaviour, at least for legal purposes, in the same sense that a human can (leaving aside the question of whether and to what extent human beings truly determine their own behaviour).

The most important point to take away from this chapter is that 'autonomy', as the concept applies to software, does not mean freedom from of human control; it is, rather, a form of control. For obvious practical reasons, autonomous software entities must be directed toward fulfilling the purpose for which they were designed. Autonomous control is the means by which that is achieved. If control can be conceived of generally as the set of measures taken to determine the behaviour of a software entity, then specific control actions can be applied by a human

47 Defense Science Board (n 36) 27.

operator either in advance of an operation, in the form of programmed behaviours (whether proactive behaviours or responses to environmental stimuli), or during an operation, in response to some indication that a control action is required (such as how an RPA pilot guides the aircraft in response to imagery captured by the RPA's camera). Autonomous control is a control paradigm which relies on control inputs applied to a system in advance, to the partial or complete exclusion of human interaction applied during an operation, so as to realise some practical benefit of relevance to a mission (speed, accuracy, persistence, stealth, and so on).

Autonomy is inherently a matter of degree. Practical limitations on a system's ability to define its own behaviour will always exist, whether as 'hard' limits in the form of programmed behaviours or 'soft' limits in the form of environmental constraints which the system is unable to overcome. The objective to be achieved in a specific mission and the need to interoperate with other entities, whether human or artificial, in pursuit of that objective, must necessarily constrain 'self-governance' to some extent. Nor will a system's degree of autonomy necessarily be constant; in practice it may be expected to vary with respect to the specific function in question and the circumstances in which the system is operating.

For the purposes of a legal analysis, that means that attempts to classify software systems as autonomous or not in a binary sense are highly error prone. If autonomous capability *per se* is to be used as the basis of a legal argument, significant care must be taken to select a representation which is technically accurate as well as legally relevant. Alternatively, perhaps the challenge for lawyers studying software autonomy is to relate the human and software behaviours that are the signature of autonomous operations directly to those that are subject to legal regulation.

Chapter 3

Autonomous Cyber Defence Capabilities

Tanel Tammet

I
INTRODUCTION

Cyber attacks and cyber defence are a cat-and-mouse game where the adversaries are continuously on the lookout for a new edge to improve their capabilities over the opponent. Since both offence and defence are conducted on computers, automation is always at hand, ranging from simple attack scripts used by 'script kiddies' to extremely complex attack systems like Stuxnet,[1] employed against the Iranian nuclear enrichment facility. However, as Bruce Schneider has said, 'if you think technology can solve your security problems then you don't understand the problems and you don't understand the technology'.[2]

1 David Kushner, 'The Real Story of Stuxnet' (*IEEE Spectrum*, 26 February 2013) <https://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet>.

2 Bruce Schneier, *Secrets & Lies: Digital Security in a Networked World* (Wiley 2000) preface <https://www.schneier.com/books/secrets-and-lies-pref/> accessed 13 January 2021.

As we will explain in the following, most of the cyber defence has always been semi-automated: it relies on the highly qualified work of human specialists using a wide range of specialized software for performing repeated mundane tasks. Actual mitigation and reaction to attacks is mostly not automated and is thus slow. The needs for better information exchange and quicker reaction to attacks appear to be the main driving forces for the ongoing deepening of non-AI automation. Complex AI-based tools exhibiting higher levels of autonomous behaviour are slowly emerging from the early experimental stage, but it does not look like they have reached the quality and maturity necessary for wide use yet.

Let us consider the meaning of 'autonomy'. It is generally agreed that autonomy is a vague term existing on a continuum. As McFarland writes[3], 'a self-governing system is more likely to be described as 'autonomous' where human observers lack the ability to precisely foresee the exact sequence of steps that the system must take in order to complete its assigned task (or, equivalently, cannot foresee all events that will transpire when the system is activated)'. This statement holds for most nontrivial automated systems. For example, almost all such systems contain bugs and this alone makes it impossible to predict with certainty what they will do next. Similarly, the behaviour of a system depends on the data it is given: again, it is impossible to predict what data the system will be given in the future and to prepare or predict actions for all the possible data combinations. In particular, it is very hard — or impossible — to predict the exact behaviour of systems employing machine learning, yet such systems are typically not self-governing and do not have intents or autonomy in a meaningful sense.

Moreover, when we think about 'autonomy' in the stronger, AI sense — as in being able to fully replace a human specialist or level 5 autonomy[4] in the context of self-driving cars — we must acknowledge that no such cars exist so far and it is likely that fully autonomous cyber defence or attack systems in that sense may be harder to achieve than fully self-driving cars. Even for a lower level of autonomy on the spectrum we observe that while complex AI-based tools are slowly emerging from the early experimental stage, it does not look like they have reached quality and maturity for wide use yet.

3 Tim McFarland, 'The Concept of Autonomy', this volume, ch 2, 17; Defense Science Board, 'The Role of Autonomy in DoD Systems' (US Department of Defense Task Force Report, July 2012) 4.

4 See Synopsys, 'The 6 Levels of Vehicle Autonomy Explained' (2021) <https://www.synopsys.com/automotive/autonomous-driving-levels.html> accessed 1 May 2021.

II
CYBER SECURITY, DEFENCE AND OFFENCE

Since 'cyber security' and 'cyber defence' cover a wide range of goals and activities, their meanings are — inevitably — somewhat vague and mostly overlapping. However, by 'cyber defence' people typically mean a more pro-active stance than is conveyed by 'cyber security'. For example, cyber intelligence and reconnaissance are often described as 'cyber defence' activities.

The absolute majority of practical cyber defence activities are, as the name says, defensive, with focus on prevention, detection and response to attacks. Since the spectrum of potential attackers is very wide, it is unrealistic to pre-emptively attack the potential attackers or even just 'hack back':[5] we just do not know whom to attack, not to speak of the high cost of doing so. Still, there is a gray zone for specific cases where offensive cyber attack may turn out to be the best defence. The most common element of the gray zone is a so-called honeypot: useless data and systems seeming important, set up specially to attract potential attackers and thus detect their actions before the real assets are targeted.

In contrast to the gray zone activities, performing real pre-emptive cyber offence first requires that we know whom to attack, ie the list of our opponents must be severely limited. This assumption normally holds true for nation states. Several countries, notably US, have regulated and legalized such offensive cyber operations and created the capability to conduct real operations[6]. Probably the most famous state-sponsored cyber operation is STUXNET, an extremely complex automated attack which paralysed the Iranian capacity of uranium enrichment.

In the cyber defence field it is commonly understood that it is easier to automate attacks than defence: the defender has to be on the lookout for a very wide range of attack methods employed by a huge number of potential attackers, from the basic employee risk to phishing, malware, DDOS, spoofing, GUI intrusion, and so on, all the way to the advanced persistent

5 Martin Giles, 'Five Reasons "Hacking Back" is a Recipe for Cybersecurity Chaos' (*MIT Technology Review*, 2019) <https://www.technologyreview.com/2019/06/21/134840/cybersecurity-hackers-hacking-back-us-congress/> accessed 1 May 2021.

6 See US Department of Defense, 'Cyber Strategy Summary' (September 2018) <https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF> accessed 1 May 2021.

threat actors. All of these major attack types have numerous subcategories, continuously evolving technical details and rising levels of automation. Last but not least, the defender must have an intimate understanding of the people, systems and business assets under their protection.

The direction of automation and use of AI for attacks is quite different from the automation and AI use for defence. We see two tendencies for automating attacks. First, the level of 'basic automation' is always growing, but mostly this means either (a) increasing the scale of attack (more systems targeted, more break-in attempts tried etc) or (b) combining numerous existing basic tools into the automatic process of conducting a complex multi-stage attack. Second, the AI-based complex automation appears to be developing in the direction of using natural language tools for automating and improving social engineering and spear-phishing. For the latter it is useful to automatically collect information about the targeted organization and individuals and then use this information for automatic fraudulent, personalized email or social media message exchange, where an AI bot impersonates a well-meaning human. The first widely reported case of a malicious chatbot comes already from 2007.[7]

III
CONVENTIONAL SEMI-AUTOMATED CYBER DEFENCE

The spectrum of cyber defence activities is very wide. Practices vary a lot between different organizations and IT setups. Cyber security specialists must understand both the types of attacks, and the ways to prevent, detect, analyze and mitigate them. They must also understand the structure and dependencies of IT assets and networks of the organisation, as well as the business value of data and software kept and running on these systems. Last but not least, they need to have a good overview of the structure of the organisation, business processes and the people actually working in the organisation.

An obvious example to illustrate the last point: cybersecurity education, monitoring of best practices and personal consultations are a part of

7 Sandra Rossi, 'Beware the CyberLover that Steals Personal Data' (PCWorld, 15 December 2007) <https://www.pcworld.com/article/140507/article.html>.

cyber defence activities. A less obvious example: during the 2007 Bronze Night cyberattacks on Estonia,[8] the cyber defence of attacked Estonian systems was conducted in close cooperation with the major telecommunications operators of Estonia, backed by existing informal networks of people. It is an interesting question whether similar actions could have been taken in case the systems under attack had been located in cloud servers of large cloud providers outside Estonia.

From the previous discussion it is clear that big parts of the competence and activities are impossible or unfeasible to automate with existing technology. Nevertheless, several other big parts of the cyber defence competence and activities can be — and routinely are — automatized. The conventional automation methods of cyber defence focus on the employment of known and trusted technologies like firewalls and virus defence systems along with the methods requiring significant amounts of regular work by a cyber security expert.

Security operations centers ('SOCs') usually employ a variety of specialised systems. First, virus defence systems and firewalls: these systems are typically also run on personal computers and thus most people have some experience with them. A more specialized set of tools are intrusion detection systems like Suricata: these monitor the selected parts of the networks of the organization.[9] They look for patterns of traffic matching large sets of concrete pre-defined rules for detecting suspicious or malicious activity on the network. Both free and commercial rulesets are actively developed and distributed, encoding the knowledge of multiple experts who build and update the rulesets.

Next, the IT assets communicating on the networks of the organization can be automatically detected using network mappers like nmap.[10] These mappers may serve a dual role of detecting known vulnerabilities. Some scanners like the F-Secure Radar focus mostly on vulnerability detection.[11] The vulnerability scanners typically produce reports about unpatched software, open unidentified network ports and such. A crucial part of the SOC arsenal is collecting and analysing logs continuously produced by most of the running software. These logs are typically text files to which an operating system or an application software regularly appends basic information about the most important actions it takes,

8 NATO StratCom Centre of Excellence, '2007 Cyber Attacks on Estonia' <https://www.stratcomcoe.org/download/file/fid/80772> accessed 13 January 2021.
9 Suricata, 'Suricata' <https://suricata-ids.org/> accessed 1 May 2021.
10 Nmap.org, 'News' <https://nmap.org/> accessed 1 May 2021.
11 F-Secure, 'F-Secure Radar: Vulnerability Management Platform' (2021) <https://www.f-secure.com/en/business/solutions/vulnerability-management/radar> accessed 1 May 2021.

or the current status of the system. There exists a large set of tools for collecting the logs from different computers, analyzing their contents and looking for suspicious patterns of activity.

In addition to specific information collected from the logs and networks of the organization, a subset of cyber security information is useful for a large number of organizations: for example, vulnerabilities detected in specific versions of widely used software, new malware identifiers and malicious IP-s and domains detected. Such information ('cybersecurity intelligence') is shared on multiple free and commercial information feeds. Organizations often exchange such knowledge via specialized cyber intelligence sharing systems set up for a limited number of cooperating organizations: the most prominent is MISP (Malware Information Sharing Platform).[12] For organizing the cybersecurity work on incident analysis and resolving, SOCs sometimes use specialized incident management software. Finally, security information and event management systems ('SIEMs') are software tools for collecting the data from the previously described multiple information sources, analysing and visualizing the results.

The installation, maintenance and monitoring the information produced by these specialized systems can quickly become overwhelming for cyber security specialists. For example, it is useful to have an eye on the logs of various systems to detect uncommon patterns potentially pointing to an incident. The amount of logs continuously generated by larger systems is staggering and the contents vary wildly. Hence it is utterly hopeless to regularly investigate the logs without a help of fairly complex log collection and analysis software: the task typically performed by SIEMs and machine learning outlier detectors. Similarly, once an organisation sets up an intrusion detection system like Suricata, it will start producing a large amount of alerts, generated by the rulesets chosen. Again, keeping an eye on these alerts requires support from software along with the regular reconfiguration of the subsystem filtering out the majority of alerts as harmless noise.

Obviously, it would be good to also keep an eye on threat intelligence feeds indicating new vulnerabilities and attack patterns to detect the snippets relevant to the protected systems. Next, the whole network of the protected IT systems should be monitored by nmap-type tools to detect new assets and various modifications, as well as scan for

12 MISP, 'MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing' <https://www.misp-project.org/> accessed 1 May 2021.

potential weaknesses with a vulnerability scanner. Incidentally, automatically scanning the vulnerabilities of a potential target is one of the most important methods used by attackers. Work on setting up defence software and hardening the systems is combined with analysing breaches and sources of new problems similarly to the police detective, along with the efforts going into proactive hardening of the systems, educating the personnel and recuperating from the harmful effects of successful attacks The list of tasks is therefore seemingly infinite.

Hence the need for a better automation of defence. But the complexity of actually achieving it is widely acknowledged. So far, the developments in the field are mostly of defensive nature. While all software performs automation by definition, we will not focus on the conventional cybersecurity tools automating specific complex tasks like efficient collection of logs from different systems, search, statistics and visualisation. Due to the wide spectrum of complex activities performed by cyber defenders, there is a tradition of mistrust of fully automated systems for cyber defence. On one hand, considering the current state of the art, the work of a cyber defender is far too complex for full automation by a hypothetical AI system. On the other hand, the cyber defence workload required for maintaining adequate defence against motivated attackers is too high for ordinary organisations. The following chapter will give a brief overview of the currently used technologies that are closer to the AI spectrum and hypothetical autonomous systems to be built in the future.

IV
AUTOMATING CYBER DEFENCE

Perhaps the easiest and hence most common AI technology used in cyber defence is outlier or anomaly detection:[13] detecting new uncommon patterns in various logs. By saying ‘easy’ we do not mean it literally: typical outlier detection systems employ different types of statistics along with machine learning and common knowledge, like the split of a week to workdays and holidays, the office hours rhythm etc. The first immediate problem with outlier detection is the large number of false positives:

13 Pierre Parend and others, ‘Foundations and Applications of Artificial Intelligence for Zero-Day and Multi-Step Attack Detection’ (2018) 4 EURASIP Journal on Information Security.

outliers which do not actually point to a weakness, attack or breach. In a way this is inevitable: the cost of getting fewer false positives is not noticing some of the true positives. In most cases, the false positives simply waste the time of the analyst. In worse cases — say, when connected to a system automatically blocking an external system — this may create problems exchanging information with external systems. The second problem which often occurs is that it is hard to understand what to do when an outlier is detected: analysing the causes and potential seriousness takes a lot of effort.

In recent years there has been significant interest in using machine learning to simulate the behaviour of a human analyst looking at alerts. For example, intrusion detection systems ('IDS') like Suricata commonly employ a large set of continuously updated expert-crafted rulesets for creating alerts for suspicious patterns of behaviour in the network traffic. The detection system monitors all the traffic in the network it is connected to, detecting matches with the installed attack detection rules, ie pluggable intelligence tidbits. The rulesets are either obtained from known free sources or bought from companies specialising in the creation and regular updates of rules that detect known threat signatures.

One such IDS installation may create millions of alerts per day. These alerts are normally filtered by throwing out known uninteresting alert patterns, and the human analyst will only investigate a small number of alerts they deem potentially interesting. It is in principle possible to use the methods of supervised machine learning to learn the patterns of 'interestingness' for a human analyst, and then propose only such potentially interesting alerts to the analyst. Current research indicates that a system learning interestingness from a human input needs regular retraining:[14] the performance of the system starts falling significantly in a few months due to the changes in the patterns of network traffic and new types of attacks. Thus, the learning system may potentially lighten the workload of the analyst, but cannot remove it completely. Due to the complexities involved such learning systems are still mostly the target of research and are not widely used in practical cyber defence systems. One technical observation from this research is that for this particular task the neural network systems perform worse than decision tree-based learning systems.[15]

14 Giovanni Apruzzese and others, 'On the Effectiveness of Machine and Deep Learning for Cyber Security' in Tomáš Minárik, Raik Jakschis and Lauri Lindström (eds), 10th International Conference on Cyber Conflict, *CyCon X: Maximising Effects* (NATO CCDCOE 2018).

15 ibid.

There are several companies focusing on developing supervised machine learning systems for detecting ransomware attacks,[16] and for merging input from multiple sources to create and maintain a universal threat recognition system which is not optimized for any specific organization.

At the high end of the autonomous defence spectrum are systems able to automatically isolate attacked or breached systems and recuperate after the incidents, for example, by automatically switching over to a backup system or switching off noncritical subsystems. Such systems are currently in the research and early deployment phase in armed forces of several countries: organizations with a huge amount of critical assets and short on the specialist manpower to defend all of them. One significant risk posed by such autonomous systems is the potential for erroneously shutting down or isolating critical systems on the basis of mistaken perception or incorrectly learned patterns: such cases will inevitably happen and the military organisations need to plan for overriding and verifying the decisions made by autonomous cyber defence systems. On the active-passive defence scale these systems can be considered to be either passive or active, depending on our interpretation of the meaning of the scale. They are passive in the sense that typically they do not launch counterattacks. They are active in the sense of a high degree of automation and their ability to directly influence the behaviour of the critical operational systems they are protecting.

The main current practice of using autonomous agents in civilian organizations like banks appears to be automatically isolating (blacklisting) dangerous or suspicious external agents from accessing the network of the organization. Even this action is not without risks: inevitably it sometimes happens that well-meaning external agents are erroneously blacklisted. The threshold of automatic countermeasures is regularly tuned by such organizations.

In 2016 NATO created the research group 'Intelligent Autonomous Agents for Cyber Defense and Resilience',[17] but as said before, R&D in the same direction is performed independently by a number of countries[18].

16 Li Chen and others, 'Towards Resilient Machine Learning for Ransomware Detection' (16 May 2019) <https://arxiv.org/pdf/1812.09400.pdf>.

17 See Michal Pechoucek and Alexander Kott (eds), *Proceedings of the NATO IST-152 Workshop on Intelligent Autonomous Agents for Cyber Defence and Resilience, Prague, Czech Republic, October 18–20, 2017* (CEUR WS 2018) <http://ceur-ws.org/Vol-2057/>.

18 US Department of Defense (n 6); see also United Kingdom, National Cyber Security Centre, 'Active Cyber Defense' <https://www.ncsc.gov.uk/section/products-services/active-cyber-defence> accessed 1 May 2021.

There are several areas of cyber defence where better automated analysis of natural language would significantly help. One of these is automatic scanning of the cybercrime marketplaces and information exchange forums. Another is spam and phishing detection: it is to be expected that due to the fast progress of AI-based text generation the amount of intelligent phishing attacks will rise significantly, thus requiring adequate countermeasures.

Yet another important and realistic usage is the development of autonomous cyber defence intelligence exchange systems. Continuous communication between the cyber defence professionals of different organisations exchanging fresh information on new threats and attacks is one of the critical and practically important parts of the cyber defence process.

Since information about cyber incidents and vulnerabilities is sensitive, organizations are not easily willing to distribute this. It is not uncommon that organizations of the similar type — for example, banks within a particular country, militaries of tightly collaborating countries etc — share such information among their closely guarded group. One of the methods used for decreasing the risk of sharing sensitive information is anonymizing it: recipients do not know who from the trusted group has sent the information. Special anonymizing servers set up by a trusted group is one of the helpful tools: for example, Airbus is reported to have special servers in Iceland for this purpose.[19] Another set of tools focuses on using proper, guaranteed privacy-preserving multi-party algorithms or trusted computation components of a microprocessor.

So far the exchanged threat intelligence data has mostly relied on short descriptions in natural language, augmented by structured data, and the interpretation of exchanged data has thus been almost exclusively done by human analysts. Since cyber defence systems and practices employed by different organisations vary significantly, converting this data to machine-processable structured knowledge has not been realistic. This can be incrementally changed by AI systems helping to both convert cyber intelligence data and to take action.

On the opposing side is the potential to use AI for cyber attacks. The first AI-supported cyberattack, recorded in 2007, came from a natural language AI chatbot CyberLover,[20] conducting flirting chats with high level of social engineering designed to steal passwords and send the victims to web sites infecting them with malware.

19 Private communication with an AIRBUS cyber security specialist.
20 Rossi (n 7).

V
SOAR, ACD AND THE DARPA CYBER GRAND CHALLENGE

The current catchphrases for automating cyber defence are SOAR (Security Orchestration, Automation and Response) and MMAR (Manage, Monitor, Automate and Respond). The SOAR[21] term was originally coined by Gartner, a global research and advisory firm. They defined the three capabilities: threat and vulnerability management, security incident response and security operations automation. Threat and vulnerability management (Orchestration) covers technologies that help amend cyber threats, while security operations automation (Automation) relates to the technologies that enable automation and orchestration within operations. According to Gartner, a SOAR platform uses 'machine-readable and stateful security data to provide reporting, analysis and management capabilities to support operational security teams'.[22]

The main goal of the SOAR technologies is not to replace, but help, the cyber security teams performing their daily tasks. This involves intelligent outlier detection in logs, machine learning, data and knowledge integration, intelligent filtering of intelligence feeds and similar supportive tasks. It also includes automating relatively mundane tasks like backups and restore, configuration migration, vulnerability scans and sometimes even basic threat response. Another facet of the increased automation is the potential for faster detection of attacks and breaches. A noticeable percentage of the initial compromise stage of data intrusions are very fast, while detection may take months. All this time is available for the attacker to deepen and widen their control.

Correspondingly, the US Department of Defense has defined a concept of Active Cyber Defense ('ACD') as 'DoD's synchronized, real-time capability to discover, detect, analyze, and mitigate threats and vulnerabilities'.[23] ACD is designed to be applicable across the US Government as well as critical infrastructure. ACD capability builds up situational awareness, which typically requires the orchestration of data collection, integration

21 FireEye, 'What is SOAR? Definition and Benefits' <https://www.fireeye.com/products/helix/what-is-soar.html> accessed 12 January 2021.

22 Paul Proctor and Oliver Rochford, 'Innovation Tech Insight for Security Operations, Analytics and Reporting' (Gartner Research, 11 November 2015) <https://www.gartner.com/en/documents/3166239/innovation-tech-insight-for-security-operations-analytic>.

23 National Security Agency, 'Active Cyber Defense (ACD)' (4 August 2015) <https://apps.nsa.gov/iad/programs/iad-initiatives/active-cyber-defense.cfm>.

and actions between different systems, different organizations and geographical locations. Thus, the question of the automatic interpretability of exchanged data becomes critical, which is borderline to the classic area of symbolic AI. ACD's six functional areas are defined as sensing, sense-making, decision-making, acting, messaging and control, and ACD mission management.[24]

A step in the same direction is the design of the STIX (Structured Threat Information eXpression)[25] and TAXII (Trusted Automated eXchange of Intelligence Information)[26] protocols for exchanging detailed structured data on cyber threat intelligence. STIX enables organizations to share cyber threat intelligence with one another in a consistent and machine-readable manner, allowing security communities to better understand what computer-based attacks they are most likely to see and to anticipate and/or respond to those attacks faster and more effectively. The importance of this direction can be exemplified by the ongoing project between the US Air Force and the Estonian Ministry of Defence to design a system for secure, interpretable and actionable exchange of cyber threat intelligence, with the main work on the Estonian side conducted by Cybernetica AS.[27]

NSA describes ACD as characteristics as follows:

> A comprehensive ACD solution would have characteristics that include the ability to operate with *dialable* levels of automated decision-making that enable the detection and mitigation of threats at cyber-relevant speed; it must be scalable to operate in any size enterprise, and work in an integrated manner with other network defense and hardening capabilities while creating and consuming shared situational awareness. Finally, these capabilities must be available soon and be designed in a manner that allows them to be built and operated by both the private sector and [the US Government]. …
>
> The ACD Framework, depicted here, describes the set of five high-level conceptual capabilities necessary to perform ACD anywhere in cyberspace. A foundational messaging fabric must exist to enable real-time communications using standard protocols,

24 ibid.
25 OASIS Open, 'Introduction to STIX' (29 November 2020) <https://oasis-open.github.io/cti-documentation/stix/intro>.
26 OASIS Open, 'Introduction to TAXII' (29 November 2020) <https://oasis-open.github.io/cti-documentation/taxii/intro>.
27 Cybernetica, 'Estonia and the United States to Build a Joint Cyber Threat Intelligence Platform' (14 January 2020) <https://cyber.ee/news/2020/01-14/>.

> interfaces and schema among the other four components. Then there must be sensors that report data on the current state of the network, sense-making analytics to understand current state, automated decision-making to decide how to react to current state information, and capabilities to act on those decisions to defend the network. Although not a unique part of the ACD framework, Shared Situational Awareness is a critical provider and consumer of actionable ACD information.[28]

An early example of the ACD in action is the NSA Sharkseer program,[29] which started around 2014 with the primary purpose to protect the US Department of Defense's networks. Sharkseer monitors emails, documents and incoming traffic that could infect the Department's networks. NSA describes the functions of Sharkseer as follows:

> IAP ('Internet Access Provider') protection: Provide highly available and reliable automated sensing and mitigation capabilities to all 10 DOD IAPs. Commercial behavioral and heuristic analytics and threat data enriched with NSA unique knowledge, through automated data analysis processes, form the basis for discovery and mitigation.
>
> Cyber Situational Awareness and Data Sharing: Consume public malware threat data, enrich with NSA unique knowledge and processes. Share with partners through automation systems, for example the SHARKSEER Global Threat Intelligence ('GTI') and SPLUNK systems. The data will be shared in real time with stakeholders and network defenders on UNCLASSIFIED, U//FOUO, SECRET, and TOP SECRET networks.[30]

In 2016, DARPA launched the Cyber Grand Challenge,[31] a competition to create automatic defensive systems capable of reasoning about flaws, formulating patches and deploying them on a network in real time. Citing their information about the event:

28 National Security Agency (n 23) (original italics).
29 Ronald Nielson, 'SHARKSEER Zero Day Net Defense' (National Institute of Standards and Technology, 10 September 2015) <https://csrc.nist.gov/Presentations/2015/SHARKSEER-Zero-Day-Net-Defense>.
30 ibid.
31 Dustin Fraze, 'Cyber Grand Challenge (CGC)' (Defense Advanced Research Projects Agency) <https://www.darpa.mil/program/cyber-grand-challenge> accessed 12 January 2021.

> DARPA hosted the Cyber Grand Challenge Final Event — the world's first all-machine cyber hacking tournament — on August 4, 2016 in Las Vegas. Starting with over 100 teams consisting of some of the top security researchers and hackers in the world, DARPA pit seven teams against each other during the final event. During the competition, each team's Cyber Reasoning System 'CRS' automatically identified software flaws, and scanned a purpose-built, air-gapped network to identify affected hosts. For nearly twelve hours, teams were scored based on how capably their systems protected hosts, scanned the network for vulnerabilities, and maintained the correct function of software. Prizes of $2 million, $1 million, and $750 thousand were awarded to the top three finishers.

CGC was the first head-to-head competition between some of the most sophisticated automated bug-hunting systems ever developed. These machines played the classic cybersecurity exercise of Capture the Flag in a specially created computer testbed laden with an array of bugs hidden inside custom, never-before-analyzed software. The machines were challenged to find and patch within seconds — not the usual months — flawed code that was vulnerable to being hacked, and find their opponents' weaknesses before they could defend against them.[32]A participating team from the University of Idaho reports that over a hundred teams registered to compete in the CGC.[33] Of these, twenty-eight entered the qualifying event and the top seven teams participated in the final event.

We must note that the performance of the winner of the competition — Carnegie Mellon University's ForAllSecure 'Mayhem' system[34] — was significantly weaker than the performance of human specialists. TechCrunch reports[35] that the team was invited to enter the similar 'Capture The Flag' tournament at the neighbouring DEF CON, where it was the worst performer among the fifteen participants. Still, we have to take into consideration that the development of fully automated cyber defence systems has just started. It is quite possible that similarly to the DARPA Grand Challenges for autonomous cars, the Cyber Grand Challenge was a landmark starting point for major developments in the field.

32 ibid.
33 Jia Song and Jim Alves-Foss, 'The DARPA Cyber Grand Challenge: A Competitor's Perspective' (2015) 13(6) IEEE Security & Privacy Magazine 72.
34 Thanassis Avgerinos and others, 'The Mayhem Cyber Reasoning System' (2018) 16(2) IEEE Security & Privacy Magazine 52.
35 Devin Coldewey, 'Carnegie Mellon's Mayhem AI takes home $2 million from DARPA's Cyber Grand Challenge' (*TechCrunch*, 5 August 2016) <https://techcrunch.com/2016/08/05/carnegie-mellons-mayhem-ai-takes-home-2-million-from-darpas-cyber-grand-challenge>.

VI
CONCLUSION

Since both cyber attacks and cyber defence are — by definition — conducted with the use of software, they are always, in some sense, automated. Thus, the question of what is meant by 'automated' cyber defence or attack is not always clear. Similarly, we cannot unambiguously say what does it mean to be 'autonomous': to some degree, any automated system is autonomous, while no truly 'autonomous' systems in the stronger AI sense currently exist.

This said, the drive to increase the automation level of cyber attacks and defence is obvious. We can observe that automation of cyber attacks is, in some sense, simpler than automating defence, since for the former it can mean clear quantitative increase: more systems attacked faster. Yet, the field of AI is still far from the level where complex high-level automated attacks could be made 'autonomously' in the sense of replacing a human specialist. Rather, these automations are, for the main part, complex human-developed scripts utilizing existing technological components and in-depth human knowledge of the systems to be attacked.

Automating cyber defence has, so far, turned out to be harder than automating attacks. The amount and structure of knowledge a cyber defence specialist must have is built upon a large amount of both general knowledge and specific knowledge about the defended systems. The work itself is highly complex and requires a lot of creativity, psychology and teamwork. All of these aspects are very hard to formalize and do not lend themselves well to machine learning techniques. Thus, when we speak about automating cyber defence, we are speaking about automating relatively mundane parts of the work. This said, we can be sure that further automation will be a major force in the development of cyber defence.

As for the practically useful autonomous cyber defence in the AI sense of replacing a human specialist, it is really impossible to predict when and how this will become a reality. We can, however, speculate that it will happen later than the large-scale deployment of fully autonomous cars: after all, the environment and tasks the driver has are less varied than what the cyber defence specialist has to tackle. Thus, the questions about issues specific to autonomous cyber attack or defence systems are still highly speculative.

Chapter 4

Ethical Artificial Intelligence: An Approach to Evaluating Disembodied Autonomous Systems

Daniel Trusilo and Thomas Burri

I
INTRODUCTION

'What our societies all over the world need is a shared and applicable ethical framework, to develop AI policies, regulations, technical standards, and business best practices.'[1] Addressing this call to action, our current project tackles the following question: How can an assessment tool designed to identify ethical issues of embodied autonomous systems be modified to apply to disembodied autonomous systems? The goal of such an undertaking is to inform a discussion about international norms and ethical principles that should apply to disembodied autonomous systems.

By applying an assessment tool we previously developed, henceforth referred to as the Schema, we are able to empirically identify ethical

1 Luciano Floridi and Lord Tim Clement-Jones, 'The Five Principles Key to Any Ethical Framework for AI' (*New Statesman*, 20 March 2019) <https://tech.newstatesman.com/policy/ai-ethics-framework>.

issues raised by autonomous disaster relief and weapon systems.[2] Such systems necessarily have a physical manifestation. They are robots with a 'body' which is why we say that they are 'embodied'.[3] The scope of the Schema has so far been limited to such embodied systems. The first step in applying the Schema is to determine if an embodied system is autonomous. We make the determination according to a composite of: (1) autarchy, which in this context refers to a system's capacity to function independently from external energy sources, (2) independence of human control, (3) interaction with the environment, (4) learning, and (5) mobility.[4] This composite picture is how we define autonomy and determine if the Schema is in fact applicable to a given system. For the purpose of this chapter, robotic systems that meet the threshold of this composite definition of autonomy are referred to as *embodied autonomous systems* or simply embodied systems. We then evaluate a system according to thirty-seven aspects to determine potential areas of ethical concern. Our practical review of embodied autonomous systems using the Schema allows us to supplement the widely agreed upon framework of international humanitarian law, human rights law, and regulatory norms.

With the following discussion, we are advancing this research by extending the Schema to cover autonomous cyber operations, or software systems. Though software systems must be integrated with physical hardware to function, we are interested in exploring the idea of autonomy as it relates to algorithms that are created to function in their own right, not as code that controls a robotic system. We label such autonomous programs used in cyber operations as *disembodied autonomous systems*. A disembodied autonomous system, for the purposes of this chapter, is therefore a software program that demonstrates properties on a spectrum of a modified composite definition of autonomy, which will be discussed in greater detail in section three.

We have chosen to use the specific terminology of embodied and disembodied systems as these terms clarify our approach to the discussion about autonomy in cyberspace. They distinguish between the physical systems that are a combination of hardware and software, which we have experience evaluating, and software or algorithms that exist to carry out their own function. This distinction is mainly drawn for didactical

2 Markus Christen and others, 'An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications' (University of Zurich Digital Society Initiative White Paper no 1, 2017) <https://ssrn.com/abstract=3063617>.

3 For similar terminology, see Curtis EA Karnow, 'The Application of Traditional Tort Theory to Embodied Machine Intelligence' in Ryan Calo, Michael A Froomkin and Ian Kerr (eds), *Robot Law* (Edward Elgar 2016).

4 Christen and others (n 2).

purposes. The aim is to improve the Schema and extend its scope while furthering the discussion of autonomous systems and how ethically problematic aspects of such systems can be practically identified. Since embodied systems may incorporate elements of disembodied systems, and vice versa, it may not always be straightforward to distinguish the two. However, a neat and clean distinction may be unnecessary. If we manage to extend the scope of the Schema, making it comprehensive and inclusive of all autonomous systems, regardless of whether they are embodied or disembodied, then it will not matter whether the lines between the types of systems are blurred. There would simply be one Schema applicable to all autonomous systems.

The discussion is complicated by the fact that Artificial Intelligence ('AI') lies at the heart of the capabilities and capacities of the autonomous systems we are investigating but does not necessarily equate to autonomy in and of itself. While the relationship between autonomy and AI may have to be researched further,[5] it is our hope that the experience of researching and evaluating autonomy in embodied systems is transferrable to research concerning autonomy in cyberspace and therefore can add value to discussions surrounding what we have chosen to call disembodied autonomous systems.

We will first describe the urgent need to develop a method of identifying ethical issues related to the design and operation of disembodied autonomous systems.[6] Next, in order to determine how the Schema can be applied to systems that only exist in cyberspace, or disembodied autonomous systems, we highlight the factors that distinguish disembodied systems from embodied systems. We then explore and highlight those selected criteria of the Schema which will need to be modified in order to be applied to disembodied systems. The next section pushes the boundary further by discussing 'systems of systems', that is systems in which a collection of autonomous or semi-autonomous systems compose a larger system. This is particularly relevant in the discussion of cyber systems as the notion of a 'system' with a specific 'beginning' and 'end' becomes further blurred. We conclude with a brief overview of key takeaways from this chapter.

5 See the discussion below, section II.
6 For further elaboration on the notion of autonomous cyber system see Tim McFarland, 'The Concept of Autonomy', this volume, ch 2.

II
THE CRITICALITY OF EVALUATING ETHICAL ISSUES RAISED BY DISEMBODIED AUTONOMOUS SYSTEMS

Developing a method to identify ethical issues concerning disembodied autonomous systems is practically relevant. Pure software programs with autonomous characteristics already exist. For example, in 2018 IBM Research demonstrated DeepLocker, an AI-powered malware that is able to evade detection until reaching a specific target. Using a deep neural network AI model, DeepLocker seems benign, only deploying malicious code when it is triggered by its intended target, which it identifies through facial recognition, geolocation, and voice recognition.[7]

At the State level, US and Russian cyber operations have actively targeted each other's critical infrastructure, namely power grids.[8] On 13 March 2020, the cyberthreat to critical infrastructure was made palpable with an attack on Brno University Hospital in the Czech Republic, which led to the postponement of surgeries, the turning away of new patients, and the shutting down of all the hospital's computers.[9] The attack, coinciding with the global COVID-19 pandemic, demonstrates the life-threatening potential of cyberattacks and lends support to calls for an emergency regime for cyberspace.[10] It is not far-fetched to surmise that cyber weapons[11] being deployed by the US, Russia, and other actors may have autonomous capabilities, at least according to the composite

7 Marc Ph Stoecklin and others, 'DeepLocker: How AI Can Power a Stealthy New Breed of Malware' (*Security Intelligence*, 8 August 2018) <https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/>.

8 A June 2019 article in the New York Times publicized the years-long cyber operations by both Russian and US entities to implant malicious code in their adversary's critical infrastructure. David E Sanger and Nicole Perlroth, 'US Escalates Online Attacks on Russia's Power Grid' (*New York Times*, 17 June 2019) <https://www.nytimes.com/2019/06/15/us/politics/trump-cyber-russia-grid.html>.

9 Matt Burgess, 'Hackers are Targeting Hospitals Crippled by Coronavirus' (*Wired*, 22 March 2020) <https://www.wired.co.uk/article/coronavirus-hackers-cybercrime-phishing>.

10 See Henning Lahmann's blog post calling for an emergency regime related to cyberattacks on hospital infrastructure. Henning Lahmann, 'Cyberattacks against Hospitals during a Pandemic and the Case for an Emergency Regime for Cyberspace' (*Fifteen Eightyfour*, 20 April 2020) <http://www.cambridgeblog.org/2020/04/cyberattacks-against-hospitals-during-a-pandemic-and-the-case-for-an-emergency-regime-for-cyberspace/>.

11 A broad definition of a cyber weapon includes software and IT systems that, through ICT networks, manipulate, deny, disrupt, degrade or destroy targeted information systems or networks. The pros and cons of this definition is discussed in Tom Uren, Bart Hogeveen and Fergus Hanson, 'Defining Offensive Cyber Capabilities' (Australian Strategic Policy Institute, 4 July 2018) <https://www.aspi.org.au/report/defining-offensive-cyber-capabilities>.

definition of autonomy we apply. For example, the NotPetya cyber-attack of 2018 relied on self-propagating malware to become one of the most destructive and costly cyberattacks ever carried out.[12] Therefore, the concept of autonomy and what it means for disembodied systems must be discussed if any regime is to be relevant to current capabilities and trends.

Though there is an active debate about moral and legal issues related to autonomy in embodied weapon systems, or autonomous weapons systems, via the Convention on Certain Conventional Weapons, discussions concerning cyber systems have so far failed to address many of the similarly applicable implications of autonomy.[13] This situation may be partly due to the tendency to silo discussions of legal ramifications of various new technologies in warfare through a technology-specific approach.[14] However, this tendency is alarming considering the likelihood that decision authorities will be delegated to both embodied and disembodied autonomous systems and that the various systems are conflated when the discussion centers on autonomy.

In a 2016 interview with the *Washington Post*, the then US Deputy Secretary of Defense, Robert Work, stated that the use of unmanned systems by the US Department of Defense (DoD) is inexorable. In clarifying DoD's position, Work explained that autonomy is a matter of delegating authorities to unmanned systems in a battle network and that delegation of authority can be expected in situations in which machines have faster than human reaction times. Work then specifically identified electronic and cyber warfare as examples of situations in which machines have faster than human reaction times, warranting the delegation of decision-making authorities to unmanned systems.[15] Despite this recognition, the US DoD Directive 3000.09 on Autonomy in Weapons Systems, which addresses authorities related to autonomous systems, explicitly

12 Andy Greenberg, 'The Untold Story of NotPetya, the Most Devastating Cyberattack in History' (*Wired*, 22 August 2018) <https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/>.

13 For a discussion of the state of international law and autonomous cyber operations as well as the importance of addressing autonomous cyber capabilities, see Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>.

14 Rain Liivoja, 'Technological Change and the Evolution of the Law of War' (2015) 97(900) International Review of the Red Cross 1157.

15 See interview with US Deputy Secretary of Defense Robert Work in 'David Ignatius and Pentagon's Robert Work Talk about New Technologies to Deter War' (*The Washington Post*, 31 March 2016) <https://www.washingtonpost.com/video/postlive/david-ignatius-and-pentagons-robert-work-on-efforts-to-defeat-isis-latest-tools-in-defense/2016/03/30/0fd7679e-f68f-11e5-958d-d038dac6e718_video.html>. The referenced discussion concerning autonomy and the delegation of authorities begins at 27:18.

states that it does not apply to autonomous or semi-autonomous systems for cyberspace operations.[16] A United Nations Institute for Disarmament Research (UNIDIR) paper on Autonomous Weapon Systems (AWSs) and Cyber Operations notes that the DoD directive excluded cyber considerations for pragmatic reasons — the directive was urgently needed and addressing autonomy in cyber operations would have delayed publication of the directive.[17] The DoD Directive 3000.09 was published in 2012 and updated in 2017, yet autonomy in cyber operations remains unaddressed.

The UNIDIR Report highlights the fact that international discussions related to what we call embodied and disembodied systems are completely divorced from each other 'with virtually no overlap between the participating experts and policy practitioners', despite the relevance of autonomy for both.[18] The Group of Governmental Experts ('GGE') discussions related to cyber security have addressed neither the concept of meaningful human control nor Article 36 obligations on the testing of the means and methods of cyber warfare,[19] both of which are topics that are heavily featured in GGE discussions of embodied AWSs.

In a similar vein, there is a need to bridge the discussions of AI and autonomy. The 2019 Defense Innovation Board's (DIB) *Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* were adopted as principles by the DoD on 24 February 2020.[20] Bounding the applicability of the DIB's recommendations to AI, the report explicitly states: 'AI is not the same thing as autonomy.'[21] The report goes on to highlight that DoD Directive 3000.09 'neither addresses AI as such nor AI capabilities not pertaining to weapon systems.'[22] Though it is clear that AI is not the same thing as malware, the fact remains that AI may

16 US Department of Defense, Directive 3000.09: Autonomy in Weapons Systems (21 November 2012, incorporating change 1, 8 May 2017) <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>.

17 United Nations Institute for Disarmament Research, 'The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations' (16 November 2017) <https://www.unidir.org/publication/weaponization-increasingly-autonomous-technologies-autonomous-weapon-systems-and-cyber>.

18 ibid.

19 James Lewis and Kerstin Vignard, 'Report of the International Security Cyber Issues Workshop Series' (United Nations Institute for Disarmament Research, 2016) <https://www.unidir.org/files/publications/pdfs/report-of-the-international-security-cyber-issues-workshop-series-en-656.pdf>.

20 US Department of Defense, 'DOD Adopts Ethical Principles for Artificial Intelligence' (24 February 2020) <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

21 Defense Innovation Board, 'AI Principles: Recommendations on the Ethical Use of Artifiical Intelligence by the Department of Defense' (US Department of Defense, 31 October 2019) <https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF>.

22 ibid.

be used as part of malware and cyber operations in general.[23] Taken as a whole, this information signals the gap in the framing of ethical and legal discussions surrounding the subjects of AI and autonomy despite their convergence in disembodied autonomous systems.[24]

Article 36 of the 1977 Additional Protocol to the 1949 Geneva Conventions requires States to conduct a weapon review prior to the acquisition or adoption of 'a new weapon, means or method of warfare.'[25] There is an existing body of literature on how the design and testing process applies to non-autonomous cyber weapons.[26] But there are further challenges to applying an Article 36 review to a system that has autonomous capabilities.[27] These challenges have led to an active debate about how to apply weapons reviews to embodied AWSs. However, autonomy does not figure in the review of cyber weapons, meaning cyber weapons that are currently under development are being designed and tested without any specific institutionalized rules or international norms. This is problematic as autonomous cyber weapons that incorporate learning, even if such learning is frozen at the moment of operationalization, may behave unpredictably.[28] The complications are obvious when one looks at the commentary to Rule 110 of the *Tallinn Manual 2.0*, in which the consensus of the group of experts states: 'Any significant changes to means or methods necessitate a new legal review.'[29] Based on this language, a State that deploys an autonomous cyber weapon may no longer be in compliance with the law of armed conflict once the autonomous cyber weapon goes beyond predicted behavior or learns and modifies it. Therefore, addressing autonomy when ethically evaluating such systems is vitally important.

23 Stoecklin (n 7).

24 Compare Heather M Roff, 'Artificial Intelligence: Power to the People' (2019) 33 Ethics & International Affairs (2) 127, 140: '[W]e need to be careful of conflating AI with automation or autonomy, for doing so risks aggregating benefits and harms in different ways, when we would do better to keep them separate.' On autonomy and AI, see also Alan L Schuller, 'At the Crossroads of Control: The Intersection of Artificial Intelligence and Autonomous Weapons Systems with International Humanitarian Law' (2017) 8 Harvard National Security Journal (2) 379, 390 ff.

25 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 ('AP I') art 36.

26 The *Tallinn Manual 2.0*, a study on the application of international law to cyber-warfare, includes part IV on Cyber Armed Conflict with extensive rules concerning the means and methods of warfare and specific guidance on the applicability of the Article 36 weapons review process to cyber weapons. Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017).

27 Vincent Boulanin and Maaike Verbruggen, 'Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies' (Stockholm International Peace Research Institute, 2017) <https://www.sipri.org/sites/default/files/2017-12/article_36_report_1712.pdf>.

28 United Nations Institute for Disarmament Research (n 17).

29 *Tallinn Manual 2.0* (n 26) commentary to Rule 110, [9].

III
DISTINGUISHING DISEMBODIED AUTONOMOUS SYSTEMS FROM EMBODIED AUTONOMOUS SYSTEMS

In order to identify how the regulatory framework applicable to embodied systems can be applied to disembodied systems and to extend our tool, the emphasis must be placed on what *distinguishes* disembodied autonomous systems from embodied autonomous systems.

When focusing on disembodied autonomous systems the task of determining the constitutive parts of the 'system' to be assessed changes. With embodied systems, the existence of some kind of robotic manifestation imparts an intuition of where and how the system is bounded. This intuition is less clear with regard to disembodied systems because of their lack of a physical manifestation. Disembodied systems may also propagate without incurring additional cost; they can be passed on like fire.[30] Such an analogy allows one to envision a disembodied autonomous system spreading widely. Such a possibility may warrant even more vigilance in the development of disembodied autonomous systems based on the precautionary principle.[31]

The autonomy of an embodied system may be viewed as *composite*. In this way, a system's autonomy may be assessed according to the five axes of: (1) autarchy, (2) independence of human control, (3) interaction with the environment, (4) learning, and (5) mobility. A system can then be positioned on each axis resulting in an overall picture of its autonomy.[32] For disembodied systems, however, 'autarchy' becomes meaningless. Electricity is a pre-condition for software to function so if the environment that a disembodied system inhabits is functioning, no additional battery or fuel source is required for the disembodied system to also function. The concept of 'mobility' also changes when applied to a disembodied system as software is incapable of physically moving on its own, though it may migrate through a network. Therefore, the concept

30 We draw here on a statement made by a legal scholar with regard to legal personhood: 'legal personhood is like fire: it can be granted by anyone who already has it'. Shawn Bayern, 'The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems' (2016) 2 European Journal of Risk Regulation 297, 304.

31 See AP I art 57; Jonathan David Herbach, 'Into the Caves of Steel: Precaution, Cognition and Robotic Weapons Systems Under the International Law of Armed Conflict' (2012) 4(3) Amsterdam Law Forum 3, 6 ff.

32 Christen and others (n 2) 10.

of mobility is not applicable if meant in the physical sense and must be adapted to relate to a disembodied system's characteristics. Applying the above considerations, an autonomous disembodied system must be one that, to a certain degree, can: (1) operate independent of human control once deployed, (2) interact with its environment based on characteristics that define the environment, and (3) learn.

A system may vary along the described axes, that is to say, not every system needs to be capable of learning. To a certain extent, we accept Tim McFarland's statement that independence from operator control is not an ideal determinant of autonomy.[33] However, we construe the term in a similar way that McFarland interprets autonomy. This means that typically a programmer/operator defines the high-level goal to be achieved by the autonomous system, while the low-level steps are subject to the system's 'discretion'[34] — it being understood that low-level steps also need to be programmed or at least learned at one point. This construction of 'independence of operator control' also has the advantage of focusing the Schema on systems exhibiting a certain degree of complexity, while excluding simple software. Image processing software such as Adobe Photoshop, for instance, is not programmed to attain high-level goals and hence cannot be considered to be 'independent from control', even though it can remove red eyes at the click of a button.

The requirement that a system interacts with the environment mirrors McFarland's emphasis on the environmental uncertainty that autonomous systems typically have to cope with. However, the Schema does not insist that the environment be uncertain. This would initially have been conceivable when the focus of the Schema had been on embodied systems. For embodied systems, uncertainty of physical environmental factors are typically a hard to overcome challenge. The pathway ahead of a robot may, for instance, become icy, there may be debris, or gusts of

33 McFarland (n 6) [26]: 'Regardless of its complexity, autonomous software amounts to a set of instructions guiding a system toward such a goal. Those instructions may endow a system with a capacity for complex actions and responses, including the ability to operate effectively in response to new information encountered during operations which may not be precisely foreseeable to a human operator. However, that does not constitute independence from human control in any sense. Rather, it is best seen as control applied in a different way, in advance rather than in real time.' And ibid 34: '"autonomy", as the concept applies to software, does not mean freedom from of human control; it is, rather, a form of control.'

34 cf ibid 21. Unlike McFarland, we refrain from using terms like 'intent' and 'awareness' to avoid the risk of anthropmorphizing the system. Cf Neil M Richards and William D Smart, 'How Should the law think about robots?' in Ryan Calo, Michael A Froomkin, and Ian Kerr (eds), *Robot Law* (Edward Elgar 2016) 13: '[W]hen it comes to new technologies, applying the right metaphor for the new technology is especially important. How we regulate robots will depend on the metaphors we use to think about them. There are multiple competing metaphors for different kinds of robots, and getting the metaphors right will have tremendously important consequences for the success or failure of the inevitable law (or laws) of robotics.'

wind may unexpectedly impact it. In cyberspace in contrast, to require a system to cope with environmental uncertainty seems to go beyond what is necessary. The environment in cyberspace is more structured and therefore the types of interactions a system can possibly face are more limited. Ice, rubble, and wind cannot occur in cyberspace (except in metaphors). So, if the Schema should also cover disembodied systems, requiring a capacity to cope with environmental uncertainty would be unnecessary. Indeed, if coherently applied across embodied and disembodied systems, such a requirement may prove overly exclusionary.

There is a clear distinction between embodied systems that are *intended to cause harm* ('weapons') and systems that are not intended to cause harm (for example, search and rescue systems). In the case of the former, the Schema evaluates an additional set of criteria. While the distinction may also make sense for disembodied systems, the notion of 'harm' may have to be construed more broadly. Observational systems, such as systems that exclusively serve to gather data, may be considered not to cause harm. On the other hand, not only systems that cause physical damage (by kinetic means, for instance breaking infrastructure), but also systems that actively cause malfunctions, delay services, and so on, may be considered harmful when such malfunctions or delays can lead to actual physical harm.

The *Tallinn Manual* provides useful orientation on the notion of harm. Regarding the definition of the use of force, the majority of the international group of experts agreed, 'acts that injure or kill persons or damage or destroy objects are unambiguously uses of force'.[35] The 2010 Stuxnet cyber-attack on the Iranian nuclear program that led to the destruction of centrifuges is an oft-cited example of a real-world cyber operation that resulted in physical damage. Such an attack could be considered a use of force. Furthermore, the consensus view of the experts, in commentary to Rule 13 of the *Tallinn Manual*, was that the aggregate sum of a series of cyber-attacks can be treated as a composite armed attack thus allowing a State to exercise the right of self-defense.[36] A disembodied autonomous system, not explicitly designed to cause physical damage, may spread through a network where it was not intended to operate, causing physical damage or delays in service to multiple systems that then lead to physical damage and/or the loss of life. Therefore, to apply our assessment tool to disembodied systems, we must revisit our method of determining if a system is intended to cause harm.[37] For disembodied

35 *Tallinn Manual 2.0* (n 26) commentary to rule 11 [8].
36 ibid commentary to rule 13 [8].
37 The notion of 'harm', which we have chosen to employ in order to be inclusive of all potentially

systems, both intention and harm should notably be understood in a less direct sense. When the operation of a disembodied system may indirectly lead to harm, we will have to apply the set of criteria that are normally reserved for weapon systems in order to ensure the ethical implications of operating the evaluated system are fully considered.

IV
THE EVALUATION OF DISEMBODIED AUTONOMOUS SYSTEMS

Once a disembodied system has been determined to have aspects of autonomy and its potential to cause harm is known, we can apply the Schema's criteria to identify ethical issues raised by a particular system. The majority of criteria that are applicable to evaluating an embodied system will directly cross-over to an evaluation of a disembodied system. An example of a directly relatable criteria is the concept of 'emergent properties'. The question of whether a system to system interaction can yield unexpected or emergent properties is relevant, but it needs no modification to apply to a disembodied system. For the purposes of this chapter we will not highlight criteria that are directly transferrable but rather the criteria that must be modified or interpreted differently to account for differences between embodied and disembodied autonomous systems.

One criterion, classified in our tool as an aspect of how the system interacts with the operator, which requires review and re-interpretation is 'responsibility attribution'. Whereas an embodied system is a physical entity that an operator must deploy from a specific location, disembodied systems are less tied to physicality and location. They migrate through the network of fiber-optic cables that connect the globe and 'lend themselves to plausible deniability'.[38] Furthermore, embodied systems are physically

ethically problematic systems, is distinctly different than the notion of an 'attack'. As pointed out by Rain Liivoja and Tim McCormack, the question of what kind of cyber operations could trigger armed conflict while falling below the threshold of an attack is not thoroughly addressed in the *Tallinn Manual*. Rain Liivoja and Tim McCormack, 'Law in the Virtual Battlespace: The Tallinn Manual and the Jus in Bello' (2012) 15 Yearbook of International Humanitarian Law 45.

38 In a July 2019 *New Yorker* article, Sue Halpern describes the June 2019 use of cyber weapons by the US against Iran in retaliation for the downing of a US surveillance drone. The article frames the challenge of attribution as a question, asking, 'How do you levy a threat when it's not clear where an attack is coming from or who is responsible?' *See* Sue Halpern, 'How Cyber Weapons are Changing the Landscape of Modern Warfare' (*The New Yorker*, 18 July 2019) <https://www.

constituted of manufactured components, which can be serial-numbered and traced. A disembodied system is a sequence of code and may be hidden within a completely innocuous program that comes from another source. For these reasons, tracing an autonomous cyber weapon for attribution purposes, even with an array of digital forensic tools, may prove time and resource intensive, if not nearly impossible.

Further complicating the ability to attribute a system to a responsible party and raising questions about proliferation, is the possibility of a disembodied system multiplying ('self-replicating') without any immediate command to do so by the human that initially developed and programmed the system. This possibility raises questions of not only how international actors can identify the human party that is responsible for the actions of a disembodied system but also if the distribution and proliferation of such a system could be monitored even if international regulations were agreed upon. Lastly, in a chapter exploring the human element of cyber operations, David Danks and Joseph H Danks emphasize the challenge of clear responsibility attribution even if it is technically known who initially programmed a system as the speed and velocity of cyber-actions means that humans will inevitably be out-of-the-loop when events occur.[39] These questions echo the notion of a responsibility gap, a well-known concern with embodied autonomous systems.[40]

Considering the deployment conditions of a system, we also need to modify the criterion that assesses a system's 'effects on [the] general population'. When evaluating an embodied system one can determine if the system is likely to come into contact with a civilian population such as crowds and other neutral populations. Disembodied systems, on the other hand, may come into contact and influence a population without the individuals ever knowing they were interacting with the systems. For instance, though humans controlled the operations, Cambridge Analytica was able to use AI to aggregate vast amounts of data and deploy targeted disinformation campaigns to influence unwitting voters via social media and affect democratic elections.[41] We will therefore explore, in the

newyorker.com/tech/annals-of-technology/how-cyber-weapons-are-changing-the-landscape-of-modern-warfare>.

39 David Danks and Joseph H Danks, 'Beyond Machines: Humans in Cyber Operations, Espionage, and Conflict' in Fritz Allhoff, Adam Henschke, and Bradley Jay Strawser (eds), *Binary Bullets: The Ethics of Cyberwarfare* (Oxford University Press 2016).

40 Robert Sparrow, 'Killer Robots' (2007) 24(1) Journal of Applied Philosophy 62.

41 See US Senate Select Committee on Intelligence, 'Report of the US Senate Select Committee on Intelligence: Russian Active Measures Campaigns and Interferences in the 2016 US Election, Volume 2: Russia's Use of Social Media with Additional Views' (116th Congress, Report 116-XX, 2019) <https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf>;

following section, how to assess the potential impact of a disembodied autonomous system that is designed to be deployed in an interconnected network of both military and civilian infrastructure.

A physical characteristic of an embodied system that can be assessed fairly easily is the 'degree of lethality'. One can determine the properties of an embodied system's physical armaments — are they lethal or not? By definition, however, a disembodied system will have no physical weapons, though it may be designed in such a way as to make it lethal. For example, a cyber weapon that targets a self-driving automobile's operating system and then causes the vehicle to accelerate into pedestrians, is lethal. What core questions must be asked then to determine the disembodied system's intended use and its degree of lethality beyond the notion of physical armaments?

Another criterion, classified in the Schema as a behavioral characteristic, relates to 'constraining the system in time and space'. An embodied system may be temporally and geographically bound through a variety of methods. However, a disembodied system does not operate in a physical space. That is not to say that constraints cannot be applied to disembodied systems or that such constraints cannot be assessed, but the concept of boundaries will need to be modified to account for the non-physical environment of cyber space.

When evaluating the behavioral characteristics of a system, one must also be able to guarantee the reliability of the system's behavior. Relating this to the assessment of a system's 'targeting' capability, one must be able to say with confidence that a system will reliably target what it has been deployed to target. When applied to an embodied system, one can determine if the system is able to reliably distinguish between lawful and unlawful targets via extensive testing including an Article 36 weapons review. Even if Article 36 reviews of physical weapons are carried out (they are not always), testing complex autonomous embodied systems that have limited autarchy, mobility, and cannot self-replicate, is already difficult.

Applying the requirement of reliability to a disembodied system presents a challenge of a different order of magnitude.[42] The *Tallinn Manual* explicitly requires certainty that both offensive and defensive cyber-attacks

European Commission, 'Code of Practice on Disinformation' (26 September 2018) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>.

42 Robert Work stated in a 2016 interview with author Paul Scharre: 'When you delegate authority to a machine, it's got to be repeatable… So, what is going to be our test and evaluation regime for these smarter and smarter weapons to make sure that the weapon stays within the parameters of what we expect it to do?' Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (WW Norton 2018) 149.

are directed at lawful targets.[43] However, the *Tallinn Manual* does not take into account autonomy. Autonomous cyber weapons may be deployed in a lawful manner, but if a disembodied system has the ability to choose targets by means of AI, how can one ensure the system's targets will remain lawful? This question is especially difficult to address given the '(in)ability to predict rapid sequences of events that can result from the use of automated responses (the chain reaction challenge).'[44] This chain reaction challenge means that the behavior of autonomous cyber systems could result in an unpredictable escalation of consequences through feedback loops that are too fast for a human to stop.[45]

Other assessment criteria in the Schema may be irrelevant to non-physical entities. In the modified Schema that applies to disembodied autonomous systems these criteria can simply be ignored. They include: the appearance of the system, physical safeguards, and environmental effects. A detailed discussion of each of these criteria is not warranted here.

V
THE NEAR FUTURE CHALLENGE: EVALUATING A SYSTEM OF SYSTEMS

Near-future applications of autonomous systems may incorporate networks of interconnected disembodied and embodied systems. From an operational perspective, intelligent collective behavior, or swarm strategies, offer incredible promise of new and powerful capabilities.[46] This concept of networked autonomous systems, or *systems of systems*, complicates attempts at classification and evaluation.[47] Evaluating a system of

43 Jeffrey S Caso, 'The Rules of Engagement for Cyber-Warfare and the Tallinn Manual: A Case Study' in *The 4th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems (CYBER)* (IEEE 2014).

44 David Danks and Joseph H Danks, 'The Moral Permissibility of Automated Responses During Cyberwarfare' (3013) 12(1) Journal of Military Ethics 18, 19.

45 For a discussion of human control over (embodied) autonomous weapon systems by means of veto power, see Noel Sharkey, 'Staying in the Loop: Human Supervisory Control of Weapons' in Nehal Bhuta, Susanne Beck, Robin Geiß, Hin-Yan Liu, and Claus Kreß (eds), *Autonomous Weapons Systems — Law, Ethics, Policy* (Cambridge University Press 2016) 35–6.

46 Joe Burton and Simona R Soare, 'Understanding the Strategic Implications of the Weaponization of Artifical Intelligence' in Tomáš Minárik and others (eds), *2019 11th International Conference on Cyber Conflict (CyCon)* (NATO CCDCOE 2019).

47 According to Airbus, '[t]he cornerstone of FCAS is the next-generation weapon system where next-generation fighters team up with remote carriers as force multipliers. Additionally, manned and unmanned platforms also will provide their uniqueness to the collective capabilities while being fully interoperable with allied forces across domains from land to cyber. The air combat cloud will enable the leveraging of networked capabilities of all pooled platforms.'

systems according to norms, values, and regulations may not be the same as individually evaluating its constituent parts. While we are discussing the broadening of the Schema to disembodied systems, let us therefore briefly contemplate the implications of an aggregate of systems in which both embodied and disembodied systems may play a role.

As Paul Scharre highlighted in his 2018 book, *Army of None*, networked systems will have the capability to perform some tasks independently with human oversight, 'particularly when speed is an advantage… Future weapons will be more intelligent and cooperative, swarming adversaries.'[48] With increasingly advanced defensive weapons, the use of swarms of low-cost unmanned systems is likely. Such systems are considered attritable, meaning a force can plan on losing any number of the individual systems without detrimental consequences to strategic outcomes, budgets, or overall capabilities. Such swarms are being developed for ground and sea operations and have already been operationally deployed in air operations.

On 1 March 2020, the Turkish military announced it had deployed swarms of drones to attack Syrian government forces.[49] Though the extent of the autonomy of the systems deployed is not evident, the Turkish military does have weaponized drones that are capable of automated functions.[50] The March 2020 operation is the first instance of a government explicitly stating it had used a swarm of weaponized drones in a coordinated offensive.

Turkey is not the only State racing to develop low-cost, AI-piloted and networked weapon platforms. Manned fifth-generation aircraft like the F-35 Joint Strike Fighter have production costs close to USD 100 million per aircraft. To augment expensive, low-volume platforms, lower-cost, autonomous aircraft such as the XQ-58 Valkyrie are being prototyped. For example, Assistant Secretary of the US Air Force Will Roper stated that the US is developing a program known as Skyborg in order to prototype an AI-piloted wingman capability. The publicly stated goal of the Skyborg program is to have autonomous and attritable systems ready by 2023.[51]

Airbus, 'Future Combat Air System (FCAS)' (2020) <https://www.airbus.com/defence/fcas.html>.

48 Scharre (n 42) 93.

49 Selcan Hacaoglu, 'Turkey's Killer Drone Swarm Poses Syria Air Challenge to Putin' (*Bloomberg*, 1 March 2020) <https://www.bloomberg.com/news/articles/2020-03-01/turkey-s-killer-drone-swarm-poses-syria-air-challenge-to-putin>.

50 See Baykar Defence technical description of the Bayraktar TB2 unmanned aerial vehicle and its capabilities including fully autonomous taxiing, take-off, landing, and cruise. Baykar, 'Bayraktar TB2' (2019) <https://baykardefence.com/uav-15.html>.

51 Valerie Insinna, 'Under Skyborg Program, F-35 and F-15EX Jets could Control Drone Sidekicks' (*Defense News*, 22 May 2019) <https://www.defensenews.com/air/2019/05/22/

These facts reinforce the urgent need for a method of ethically evaluating not just individual, embodied and disembodied autonomous systems but rather the combined effect of a network of autonomous systems coordinated and controlled by AI as a system of systems. Though one could classify a swarm of drones as one whole embodied autonomous system and apply the Schema as it is, the lines begin to blur when autonomous cyber systems play a role in coordination with embodied autonomous systems. For example, it is conceivable that an autonomous software platform that only exists in cyberspace could be used to command and control an interconnected fleet of systems including drone swarms; associated logistical support systems; and intelligence, surveillance, and reconnaissance assets.[52] The notion of a system of systems based entirely in cyberspace also raises questions. For example, how could one determine where in a cyber system of systems the 'system' to be evaluated begins or ends?

VI
CONCLUSION

Certain norms purportedly govern cyberspace, but the kind of consensus supporting traditional law has so far proven elusive. Tools, such as a modified version of the Schema, which can be used to identify ethical issues raised by disembodied autonomous systems, must be further developed. Such tools can inform pragmatic discussions of ethical and regulatory norms in a proactive way. There is a clear link between the issues raised by autonomous embodied systems and those raised by disembodied autonomous systems. Discussions about the norms and laws governing these two distinct yet related manifestations of autonomy should inform one another. Already much of our research focuses on how the underlying software used in autonomous robotics manifests in the physical world. The extension of our assessment tool to software systems is an attempt at linking the two discussions and brings us into the legal and ethical discussions of cyberspace.

under-skyborg-program-f-35-and-f-15ex-jets-could-control-drone-sidekicks/>.

52 For an example of a currently operational system of systems that uses a software platform to command and control multiple interlinked hardware systems see the US Navy's Aegis Combat System. The Aegis system features fully-autonomous and semi-autonomous functions. Lockheed Martin, 'Aegis: The Shield of the Fleet' (2 May 2020) <https://www.lockheedmartin.com/en-us/products/aegis-combat-system.html>.

Chapter 5

Will Cyber Autonomy Undercut Democratic Accountability?

Ashley Deeks[1]

I
INTRODUCTION

In recent years, democratic legislatures have struggled to maintain a role for themselves in government decisions to conduct extraterritorial military operations, including those that involve the use of force. The US Congress offers a prime example of this phenomenon, but other legislatures such as the British Parliament and the French National Assembly face similar challenges.[2] Some of these challenges are due to constitutional provisions, institutional structures and historical practice. Even

1 Thanks to Kristen Eichensehr, Duncan Hollis, John Hursh, Chris Spirito, Paul Stephan, and participants in the NATO CCDCOE group that is examining the legal implications of cyber autonomy for very helpful comments and conversations, and to Ben Doherty and Christopher Kent for outstanding research.

2 See, eg, *United Kingdom, House of Commons, Public Administration and Constitutional Affairs Committee, 'The Role of Parliament in the UK Constitution: Authorizing the Use of Military Force'* (6 August 2019) <https://publications.parliament.uk/pa/cm201719/cmselect/cmpubadm/1891/189102.htm>; Delphine Deschaux-Dutard, 'Parliamentary Scrutiny of Military Operations in France and Germany' (European Consortium for Political Research) <https://ecpr.eu/filestore/paperproposal/ca1d8496-d41c-47d7-96c7-d35ef4532c90.pdf> accessed 14 October 2020. Although legislatures in non-democratic systems also face challenges in regulating and overseeing their executives, that problem extends far beyond the cyber issues that I discuss here.

constitutions that give legislatures a role in authorizing military force *ex ante* often empower executives to respond to sudden attacks without legislative blessing. Further, executive branches are necessarily better structured than legislatures to collect classified information, respond quickly to urgent security threats and direct military operations.[3]

Not all legislative limitations are linked to constitutional rules or structures, however. These legislatures are also struggling to preserve their roles because of the changing nature of conflict: a shift away from large-scale, kinetic operations toward smaller-scale operations, including operations in cyberspace, that are harder to detect publicly and do not require the type of robust legislative support that large-scale conflicts do.[4] These modern operations leave legislatures struggling to learn the facts and engaging in *ex post* and sometimes ineffective efforts to hold their executive branches accountable for offensive cyber operations that could lead to hostilities with other States.

The introduction of increased autonomy into this setting has the potential to further alter the existing relationships between executives and legislatures in making decisions that implicate the use of force. Because the use of autonomous cyber tools may lead States into serious tensions — if not armed conflict — with other States without advance notice, these capabilities pose particular hurdles for legislatures that already struggle to stay relevant on use of force and cyber issues. Additionally, a State's ability to employ autonomous cyber tools may alter the dynamics among different actors *within* executive branches themselves — by, for instance, diverting deliberative input and oversight abilities away from foreign, intelligence and justice ministries and toward defense ministries in the lead-up to conflict.

This article explores how the use of increasingly autonomous cyber tools may alter the current state of legislative oversight and internal executive decision-making about the resort to force. It also illustrates how these changes may impact international peace and security; and it identifies ways in which States may prevent a further erosion of democratic accountability for cyber-related *jus ad bellum* decisions. Unless legislatures take steps now to preserve a role for themselves, and unless executive

3 Overclassification by executive branches, or an excessive unwillingness to share classified information with legislative overseers, are persistent problems in checking executive national security activities. This article assumes that legislatures will continue to face hurdles on this front, and intends to highlight how cyber autonomy will create additional hurdles.

4 Jack Goldsmith and Matthew Waxman, 'The Legal Legacy of Light-Footprint Warfare' (2016) 37 The Washington Quarterly 7, 10 (noting that cyberattacks are low-visibility and arguing that they 'attract[] less public, congressional, and diplomatic scrutiny than the operations [they] replaced').

branches ensure that an appropriate diversity of officials remains involved in use of force decisions, key vestiges of democratic accountability for those decisions may fall away. Executives will not wait long for their legislatures to act, given the urgency of cyber threats.

Part II examines the likely trajectory of national security-related cyber autonomy within various States. Part III briefly sets out the powers that various States have allocated to their legislatures on use of force issues. Part IV synthesizes those analyses to contemplate the additional challenges that growing levels of cyber autonomy will pose to legislatures — and civilian actors within executive branches — that seek to retain input into governmental decisions that may lead to interstate conflict. Part V sets out some normative proposals for ways in which legislatures and executive branches can meet these challenges. This Part argues that legislatures should bolster their own technological expertise and consider enacting laws that place appropriate parameters on the executive branches' development and use of cyber autonomy. Within executive branches, civilian policymakers and lawyers from a range of agencies should insist on a role for themselves in developing the rules of the road for using autonomous cyber tools.

II
THE PROSPECTS FOR CYBER AUTONOMY

In national security settings, States are increasingly likely today to deploy cyber tools that use heightened levels of autonomy. This Part describes generally the prospects for burgeoning cyber autonomy within State military and intelligence systems, and then details the ways in which cyber autonomy may lead to situations of serious interstate tensions or even armed conflict.

A DEFINING CYBER AUTONOMY

Before discussing why States have incentives to increase the levels of autonomy that they build into their cyber tools, it is necessary to explain what this article means by 'autonomy'. Autonomy exists on a continuum:

systems may be more or less autonomous, or not autonomous at all.[5] As Tim McFarland writes:

> While there is no precise threshold [beyond which a system becomes autonomous], the term is generally associated with self-governing machines whose task requires higher levels of 'algorithmic and hardware sophistication' and the ability to operate in the face of uncertainty .... [A] self-governing system is more likely to be described as 'autonomous' where human observers lack the ability to precisely foresee the exact sequence of steps that the system must take in order to complete its assigned task (or, equivalently, cannot foresee all events that will transpire when the system is activated).[6]

Others have noted, 'A system with a high level of autonomy is one that can be neglected for a long period of time without [human] interaction'.[7]

There is a modest level of autonomy in any system that achieves goals previously programmed by its operator without needing to receive instructions from the operator on an ongoing basis.[8] As the task or the environment in which the system is operating becomes more complex, autonomous systems will require more complex coding to achieve the operator's desired result.[9] This might be the case, for instance, when a State's military expects that its system will encounter a 'high degree of uncertainty in the environment in which it operates', perhaps because it may confront an adversary's autonomous system.[10] The more self-adaptive a cyber system is, the more likely it is that the system will be able to operate in those uncertain environments.[11] It is possible to design systems so that they do not need 'detailed foreknowledge of all combinations of circumstances which the software entity may encounter once it is in operation'; other systems may learn 'online' once deployed.[12] Such systems fall on the higher end of autonomy.

5 Tim McFarland, 'The Concept of Autonomy', this volume, ch 2, at 35 ('Autonomy is inherently a matter of degree'.); Defense Science Board, 'The Role of Autonomy in DoD Systems' (US Department of Defense 2012) 4 <https://fas.org/irp/agency/dod/dsb/autonomy.pdf> (noting that 'system autonomy is a continuum').
6 McFarland (n 5) 16–17.
7 Michael A Goodrich and Alan C Schultz, 'Human-Robot Interaction: A Survey' in Youn-kyung Lim (ed), *Foundations and Trends in Human-Computer Interaction* (Korea Advanced Institute of Science and Technology 2007) 203, 217.
8 McFarland (n 5) 21–22.
9 ibid 22.
10 ibid 23.
11 ibid 23–24 (discussing self-adaptive systems).
12 ibid 25.

B THE COMING OF INCREASED CYBER AUTONOMY

The trend toward increasing autonomy across weapons and weapons systems is pronounced. In his book Army of None, Paul Scharre predicts that this same trend will manifest itself in cyberweapons. He writes, 'Cyberweapons of the future — defensive and offensive — will incorporate greater autonomy, just the same way that more autonomy is being integrated into missiles, drones, and physical systems like Aegis'.[13] Indeed, another scholar notes that States already are widely deploying autonomous cyberweapons.[14] Stuxnet is an example of a cyber operation that entailed considerable autonomy: the cyber worm that the United States and Israel reportedly directed against Iran's nuclear centrifuges was 'an autonomous goal-oriented intelligent piece of software capable of spreading, communicating, targeting and self-updating'.[15]

There are at least two reasons why States increasingly will rely on autonomy in their cyber operations. First, and most obviously, the speed of adversaries' offensive cyber operations requires States to *defend* their systems at the same battle speed — which may be faster than a human can react. States will need to rely on some level of autonomy to have a chance at successfully defending their systems.[16] In the United States, a 2016 Defense Science Board (DSB) report described existing autonomous systems that 'carry out real-time cyber defense' while 'also extract[ing] useful information about the attacks and generat[ing] signatures that help predict and defeat future attacks across the entire network'.[17] It also cited a tool called Tutelage, which autonomously inspects and analyzes three million packets per second on an unclassified Defense Department computer system to prevent attacks.[18] The DSB report further imagined the existence of autonomous systems 'to control rapid-fire exchanges of cyber weapons and defenses', which would seem to require greater

13 Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (WW Norton 2018) 222.

14 Rebecca Crootof, 'Autonomous Weapons and the Limits of Analogy' (2018) 9 Harvard National Security Journal 51, 81; see also Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities Under International Law' (NATO CCDCOE 2019) 11–12 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/> (discussing existing defensive and offensive cyber capabilities).

15 Stamatis Karnouskos, 'Stuxnet Worm Impact on Industrial Cyber-Physical System Security' (Paper presented at IECON 2011 — 37th Annual Conference of the IEEE Industrial Electronics Society, Melbourne, 7–10 November 2011) <https://ieeexplore.ieee.org/document/6120048>.

16 Crootof (n 14) 81 (noting that 'the speed of cyber will nearly always require that countermeasures be automated or autonomous to be effective').

17 Defense Science Board, 'Summer Study on Autonomy' (US Department of Defense 2016) 92 <https://www.hsdl.org/?view&did=794641>.

18 ibid 58.

elements of autonomy than packet inspection systems.[19] The US government seems to have pursued those systems. In 2017, the Defense Innovation Unit Experimental contracted for the Voltron project, which uses artificial intelligence to 'automatically detect, patch and exploit existing software vulnerabilities'.[20] The contract outlined defense use cases, but the system also 'has the potential to be used for offensive hacking purposes'.[21]

Second, deploying *offensive* cyber systems that are increasingly autonomous will make it easier for States to identify and then exploit adversaries' cyber vulnerabilities[22] because the systems can take advantage of machine-learning tools. These tools can identify patterns or abnormalities among vast quantities of data, which is helpful when trying to detect flaws in and infiltrate adversaries' cyber defenses. As James Johnson and Eleanor Krabill note, 'The machine speed of AI-augmented cyber tools could enable even a low-skilled attacker to penetrate an adversary's cyber defenses. It could also use advanced persistent threat tools to find new vulnerabilities'.[23]

Of course, defensive and offensive uses of autonomous cyber systems are interconnected. Even if States would prefer to use autonomous cyber systems solely in a defensive posture, Eric Messinger argues that the development of cyber defenses means that 'the development and deployment of offensive [autonomous cyber weapons] may well be unavoidable'.[24] Messinger notes,

> Powerful trends will exist toward optimizing offensive operations in cyber, and the paths of development for offensive malware could increasingly involve autonomous agents. Consider, for instance, a Washington Post report on the NSA's proposed use of a system, 'code-named TURBINE, that is capable of managing "potentially millions of implants"' — e.g., sophisticated malware — 'for intelligence gathering "and active attack"'. Though the details would matter for classifying such a system

19 ibid 4.
20 Chris Bing, 'The Tech Behind the DARPA Grand Challenge Winner Will Now Be Used by the Pentagon' (*Cyberscoop*, 11 August 2017) <https://www.cyberscoop.com/mayhem-darpa-cyber-grand-challenge-dod-voltron/>.
21 ibid.
22 United Nations Institute for Disarmament Research ('UNIDIR'), 'The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapons and Cyber Operations' (2017) 4 <https://unidir.org/files/publications/pdfs/autonomous-weapon-systems-and-cyber-operations-en-690.pdf>; Eric Messinger, 'Is It Possible to Ban Autonomous Weapons in Cyberwar?' (*Just Security*, 15 January 2015) <https://www.justsecurity.org/19119/ban-autonomous-weapons-cyberwar/>.
23 James Johnson and Eleanor Krabill, 'AI, Cyberspace, and Nuclear Weapons' (*War on the Rocks*, 31 January 2020) <https://warontherocks.com/2020/01/ai-cyberspace-and-nuclear-weapons/>.
24 Messinger (n 22).

> as autonomous, as opposed to 'semi-autonomous' or automated, it is easy to envision capabilities in the medium-term for which no other description is possible.[25]

Scharre contemplates a world in which offensive cyber operations go a step further. Instead of simply developing tools that actively manage implants or seek out enemy vulnerabilities, Scharre speculates that States might develop cyber tools that, once deployed, can fix themselves in the field and resist attack. He notes, 'Adaptive malware that could rewrite itself to hide and avoid scrutiny at superhuman speeds could be incredibly virulent'.[26] In the Defense Advanced Research Projects Agency's 2016 Grand Cyber Challenge, ForAllSecure's system was 'capable of automatically healing a friendly system while simultaneously scanning and attacking vulnerabilities in adversary systems'.[27] The US National Security Agency reportedly developed, or at least sought to develop, a system that would employ algorithms that constantly analyze metadata to detect malicious patterns, stop those attacks and autonomously initiate retaliatory counterattacks.[28] Others have envisioned decentralized swarms of autonomous agents that could attack systems without the need for centralized command and control.[29]

The United States is not the only State interested in bolstering the autonomy of its cyber operations. The United Kingdom has expressed an interest in pursuing autonomous cyber weapons as well.[30] Russian officials have stated that they view artificial intelligence as 'a key to dominating cyberspace and information operations', which suggests they intend to rely on certain levels of autonomy to achieve that goal.[31] China, too, appears committed to developing autonomous cyber capabilities.[32]

25 ibid.
26 Scharre (n 13) 226; see also Alessandro Guarino, 'Autonomous Intelligent Agents in Cyber Offense' in Karlis Podins and others (eds), *2013 5th International Conference on Cyber Conflict* (NATO CCDCOE 2013) (envisioning autonomous agents that are able to identify 'possible threats from defenders' and 'prevent and react to countermeasures').
27 Bing (n 20).
28 Nicholas Sambaluk (ed), *Conflict in the 21st Century: The Impact of Cyber Warfare, Social Media, and Technology* (ABC-CLIO 2019) 55.
29 Guarino (n 26).
30 United Kingdom, *National Cyber Security Strategy 2016–2021* (2016) [7.3.6] <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf>.
31 Peter Apps, 'Are China, Russia Winning the AI Arms Race?' (*Reuters*, 15 January 2019) <https://www.reuters.com/article/us-apps-ai-commentary/commentary-are-china-russia-winning-the-ai-arms-race-idUSKCN1P91NM>.
32 Bill Gertz, 'US and China Racing to Weaponize AI' (*Asia Times*, 7 November 2019) <https://asiatimes.com/2019/11/us-and-china-racing-to-weaponize-ai/> (stating that 'Chinese multi-domain AI warfare will expand the battlespace from traditional air, sea, and land, to … cyberspace' and discussing military operations to include 'electronic countermeasures' and 'cybertakeover').

Although fully autonomous offensive cyber systems may remain speculative today, they lie within the realm of possibility. It is therefore worth considering how these tools — or even systems with moderate levels of autonomy — might escalate low-level cyber exchanges into uses of force that implicate international and domestic laws, or at least leave States poised on the brink of armed conflict.

C HOW CYBER AUTONOMY COULD LEAD TO HOSTILITIES

Cyber operations have the capacity to cause physical damage and, potentially, human harm. To date, very few of the known cyber operations have caused levels of damage that constitute uses of force or armed attacks under the UN Charter.[33] Yet States clearly have contemplated that cyber operations could produce such a result. Former US State Department Legal Adviser Harold Koh noted, for instance, 'Commonly cited examples of cyber activity that would constitute a use of force include, for example: (1) operations that trigger a nuclear plant meltdown; (2) operations that open a dam above a populated area causing destruction; or (3) operations that disable air traffic control resulting in airplane crashes'.[34] These types of operations, though still unrealized, are well within the realm of the possible, whether States or non-state actors commit them using cyber attacks with low or high levels of autonomy.

Even if an initial offensive cyber operation does not rise to the level of a use of force, some scholars have argued that the cyber domain is one in which escalation is likely.[35] Because it is harder to predict the impact of a given cyber operation than to predict the impact of a missile, there is greater room for miscalculation, even if the victim State intends to respond in a proportionate manner. As Scharre notes, 'You can have an accident that spirals out of control very badly that has a widespread effect

33 Gary Corn and Eric Jensen, 'The Use of Force and Cyber Countermeasures' (2018) 32 Temple International & Comparative Law Journal 127 (noting that 'most unfriendly acts between nations fall below the use of force').

34 Harold Hongju Koh, 'International Law in Cyberspace' (US Department of State, Remarks at the USCYBERCOM Inter-Agency Legal Conference, 18 September 2012) <https://2009-2017.state.gov/s/l/releases/remarks/197924.htm>.

35 See, eg, Herbert Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace' (2012) 6 Strategic Studies Quarterly 46; Michèle Flournoy, Avril Haines and Gabrielle Chefitz, 'Building Trust through Testing' (*WestExec*, October 2020) 8 <https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf> ('The potential for unintended engagement or escalation is even greater when US and/or adversary systems have the sorts of advanced autonomy features that deep learning can enable, and their interaction cannot be studied or fully tested in advance of deployment'.).

in ways that are not possible with people' because humans cannot make the same number of errors as fast.[36] It also can be hard for States to signal their intentions in cyberspace, and those signals are an important way to avoid inadvertent escalation.[37]

Other scholars have suggested that concerns about cyber escalation are overblown. One pair of scholars, for instance, notes the world has seen little such escalation to date, perhaps because the tools and knowledge about vulnerabilities that a State needs to retaliate in cyberspace may not exist at the time the responding State needs them.[38] Further, decision-makers may be hesitant to respond to hostile cyber operations in some circumstances.[39]

Some of these constraints on escalation may weaken, however, when a State employs highly autonomous cyber systems. First, highly autonomous systems might by their nature be able to penetrate adversary systems more quickly and deftly than human-in-the-loop systems, requiring fewer advanced manual efforts to develop targets. Second, assuming that clear signaling is a good way to avoid unintended escalation, it may be harder for State operators to signal their intent to adversaries in advance of or during an autonomous cyber operation when those specific operations may happen without human pre-planning and possibly without knowledge of the opponent's identity. Third, highly autonomous cyber tools may act less predictably than human-in-the-loop systems, especially when confronting other autonomous systems. A UN report noted,

> As with the occasional stock market 'flash crashes', different algorithms — and even systems with very little autonomy — may interact in unforeseen ways before a human has time to intervene. … Emergent effects (unplanned and unintended) arise from interaction between the systems, and these effects are by definition

36 Johanna Costigan, 'Four Specialists Describe Their Diverse Approaches to China's AI Development' (*New America*, 30 January 2020) <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/four-specialists-describe-their-diverse-approaches-chinas-ai-development/>.

37 Brandon Valeriano, 'Managing Escalation Under Layered Cyber Deterrence' (*Lawfare*, 1 April 2020) <https://www.lawfareblog.com/managing-escalation-under-layered-cyber-deterrence>.

38 See Erica Borghard and Shawn Lonergan, 'Cyber Operations as Imperfect Tools of Escalation' (2019) 13 Strategic Studies Quarterly 122; see also Sarah Kreps and Jacquelyn Schneider, 'Escalation Firebreaks in the Cyber, Conventional, and Nuclear Domains: Moving Beyond Effects-Based Logics' (2019) 5 Journal of Cybersecurity 1; Valeriano (n 37) (arguing that the cyber domain is not 'escalation dominant' but noting that there is 'no uniform view of how escalation should work in cyberspace').

39 See Borghard and Lonergan (n 38); Thomas Rid, *Cyber War Will Not Take Place* (Oxford University Press 2013) (arguing that the real threats are espionage, sabotage, and subversion, not armed conflict initiated in cyberspace); Jon Randall Lindsay, 'Restrained by Design: The Political Economy of Cybersecurity' (2017) 19 Digital Policy, Regulation & Governance 493.

> unpredictable, so our ability to plan for how to mitigate their consequences is poor.[40]

Fourth, even if a State itself takes steps to avoid a 'flash conflict' between its own cyber algorithm and another actor's algorithm, a third State could deliberately design a cyber operation to trigger this type of event between two of its adversaries.[41] Particularly for autonomous systems driven by artificial intelligence, 'autonomy itself will likely increase a military's vulnerability to cyberattacks' because artificial intelligence can increase the anonymity of attacks in cyberspace and thus facilitate an adversary's efforts to 'use malware to take control, manipulate, or fool the behavior and pattern-recognition systems of autonomous systems'.[42] These factors, taken together, suggest that autonomous systems may be susceptible to escalating cyber hostilities, even if States do not engineer them to be so.

None of this is to suggest that the developers of highly autonomous systems lack control over the parameters of their systems; after all, the 'behaviour of an autonomous system ultimately depends upon actions of people in relevant positions, notably its designer and operator, due to the nature of computers and software'.[43] What it does suggest is that a highly autonomous system may not act entirely predictably on its own, especially if it relies on machine learning, and it may act especially unpredictably when it confronts another actor's autonomous system. It is this situation — when the system deviates in problematic ways from decisions that a human would have made had the human undertaken the task — that gives rise to new types of democratic and strategic concerns.

D AUTONOMY AND INTERNATIONAL LAW

Notwithstanding these looming problems with increased autonomy, international law does not expressly prohibit States from using autonomous cyber tools. Although many States have agreed that existing bodies of international law — including the UN Charter and the laws of armed conflict — apply in cyberspace, those laws do not specifically preclude the

40 UNIDIR (n 22) 9.
41 ibid 10.
42 Johnson and Krabill (n 23).
43 McFarland (n 5) 20; see also Defense Science Board (n 5) 1–2 ('[A]ll autonomous systems are supervised by human operators at some level, and autonomous systems' software embodies the designed limits on the actions and decisions delegated to the computer'.).

use of autonomous systems or weapons. Instead, States are governed by the traditional *jus ad bellum* rules that regulate their resort to force and jus in bello rules that regulate the conduct of armed conflict, whether they use autonomous cyber tools or not. This means that States have a legal obligation to ensure that they deploy autonomous cyber systems in a way that comports with those rules. It would be lawful, for instance, for a State to 'produce and rely on machine-learning algorithms that allow them to defend' against cyber armed attacks 'at the speed of light, in what may come to look like automatic self-defense',[44] as long as those algorithms act consistent with the customary international law rules of necessity and proportionality.[45] States that deploy autonomous cyber tools during armed conflict will need to ensure that those tools can comply with the *jus in bello* requirements of distinction, proportionality, and precautions. Finally, concepts of state responsibility may help deter States from engaging in internationally wrongful acts while using autonomous cyber tools.

That said, building autonomous cyber systems that are able to detect when an incoming operation rises to the level of an armed attack, determine whether a cyber use of force is necessary in response, and initiate a cyber self-defense operation that is proportional to the incoming attack is easier said than done, as both a legal and practical matter. Former US Deputy Secretary of Defense Robert Work was willing to accept the possibility that the United States might need to deploy automated cyber counterattacks but recognized that delegating authority to autonomous or automated systems comes with risks. He noted that a 'machine might launch a counter cyber attack' and inadvertently cause an airplane to crash, for example, something that might violate the rules of the *jus ad bellum* and *jus in bello*.[46] Further, because cyberattacks are likely to be 'disguised by being routed through third-party machines, such as an unwittingly infected botnet or third-party private or public servers', autonomous responses risk targeting entities other than the State that initiated the attack.[47] An unwitting third-party State that suddenly faces

44 Ashley Deeks, Noam Lubell and Daragh Murray, 'Machine Learning, Artificial Intelligence, and the Use of Force by States' (2019) 10 Journal of National Security Law & Policy 1, 7.

45 Although most states have accepted that the UN Charter and the right to self-defense attach in the cyber setting, a few States have resisted this idea, including Cuba. See Michael Schmitt and Liis Vihul, 'International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms' (*Just Security*, 30 June 2017) <https://www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms/>.

46 See Liivoja, Naagel and Väljataga (n 14) 23 (discussing autonomous responses that could violate the *jus ad bellum* and *jus in bello*).

47 Thomas Remington and others, 'Toward US–Russian Bilateral Cooperation in the Sphere of Cybersecurity' (Working Group on Future of US–Russia Relations 2016) 14 <https://futureofusrussiarelations.files.wordpress.com/2016/06/wg_working_paper7_cybersecurity_final.pdf>. This is not to suggest that such mistakes could never happen in human-in-the-loop cyber responses.

hostile cyber operations from the original victim may well respond in kind, setting the stage for unintended conflict.

Autonomous activities in cyberspace thus risk escalating cyber interactions to levels that violate international law, and possibly even to levels that constitute armed attacks that would trigger the adversary's right to self-defense. Delegating the authority to an autonomous system to decide when to respond to incoming attacks and effectively go on the counter-offensive 'could be very dangerous'.[48] This is especially true when States have asserted that they will only decide that something constitutes an armed attack based on a range of factors, including the apparent intent of the attacker and its identity.[49] It would be virtually impossible for an autonomous cyber system today to ascertain and evaluate factors such as intent before taking a response in national self-defense.

This all assumes that States would launch offensive or counter-offensive autonomous systems into the ether without plans to maintain meaningful control over them. It is far from clear that States such as the United States would do so. For instance, to help avoid consequences such as unintentional airplane crashes as the result of autonomous cyber operations, then-Deputy Secretary Work envisioned a role for scientists, lawyers and ethicists; automated safeties; and human oversight of the autonomous systems.[50] Others have noted that 'command and control of a true autonomous agent, especially a purely computational one … would have to translate chiefly in precise specifications of the agent's target and objectives — the goals — or, in military terms, in precise briefings before any mission'.[51] In short, there are strategic measures that States should take to avoid unintended escalation and conflict when deploying highly autonomous cyber systems.[52] The fact remains, however, that unless carefully managed, autonomous cyber exchanges risk escalating offensive and counteroffensive operations to a point that could trigger one State's right of self-defense and bring two States into hostilities without considered governmental decisions to do so.

48 Scharre (n 13) 223.
49 See Koh (n 34) ('In assessing whether an event constituted a use of force in or through cyberspace, we must evaluate factors: including the context of the event, the actor perpetrating the action (recognizing challenging issues of attribution in cyberspace), the target and location, effects and intent, among other possible issues.').
50 Scharre (n 13) 228.
51 Guarino (n 26).
52 See Part V.

III
LEGISLATIVE ROLES IN USES OF FORCE AND OTHER MILITARY OPERATIONS

Part II illustrated that a range of States are likely to pursue high levels of cyber autonomy in an effort to protect their military systems and that such autonomy, unless carefully managed, raises the prospect of deliberate or unplanned escalation into hostilities. In light of this, how can States ensure that their governments deploy cyber autonomy in a manner consistent with their constitutions and laws?[53] In particular, how should legislatures regulate autonomous cyber tools to ensure that their executive branches remain faithful to domestic and international law regulating the resort to interstate force or other military operations?[54] This Part considers the several roles that legislatures play today in authorizing or overseeing their executives' military operations, to set the stage for Part IV's analysis of how cyber autonomy may alter those dynamics.

A DEMOCRACIES AND MILITARY OPERATIONS

Several scholars have examined the extent to which legislatures play a role in States' decisions to use interstate force and therefore provide democratic accountability for those choices. In 1996, Lori Damrosch, for instance, identified a trend toward a greater legislative role in State decisions to resort to force.[55] In 2003, she asserted that 'democratic parliaments [] play active roles in determining the scope and terms of national commitments to multilateral peace operations' such as the

53 We should also care about the extent to which the use of autonomous cyber tools comports with *international* law — and in particular the *jus ad bellum* and *jus in bello*. See, eg, Liivoja, Naagel and Väljataga (n 14); Guarino (n 26) (discussing the applicability of those bodies of law to autonomous cyber agents).

54 Some scholars argue that remote warfare technologies are intended to subvert democratic control of war. See, eg, Peter Singer, 'Do Drones Undermine Democracy?' (*Brookings Institution*, 22 January 2012) <https://www.brookings.edu/opinions/do-drones-undermine-democracy/> (arguing that 'new technology is short-circuiting the decision-making process for what used to be the most important choice a democracy could make'). This article assumes, however, that democratic states such as those in NATO wish to retain democratic accountability for their use of autonomous military systems.

55 Lori Damrosch, 'Is There a General Trend in Constitutional Democracies Toward Parliamentary Control over War-and-Peace Decisions?' (1996) 90 Proceedings of the ASIL Annual Meeting 36.

operations in the First Gulf War and Kosovo.[56] Other scholars have argued that since 1990, legislatures, at least in Europe, have sought to expand their involvement in decisions to use force abroad.[57]

One reason why it matters whether legislatures play a role in a State's decisions to deploy forces abroad or resort to force outside its territory is that mature democracies usually do not go to war with each other; they also are more likely to win the wars that they fight against autocratic states.[58] This suggests that there are virtues to retaining a healthy role for democratic legislatures in war-making decisions because they may help their States avoid 'bad' wars and fight only 'good' wars.[59] Tom Ginsburg notes,

> The democratic advantage in war, some theorize, results from the need to mobilize support among the public before going to war. Legislatures can play a role here, most obviously … by requiring evidence to justify wars …. Another source of democratic advantage is signaling: when the debate about going to war takes place in public and results in a decision to fight, the counterparty can more reliably assume that the state in question is really committed. This might lead the counterparty to back down ….[60]

In other words, the legislative role in making decisions to use force may play an important role in determining whether and when States go to war and whether they win those wars.

B SPECIFIC LEGISLATIVE ROLES IN WAR-MAKING

Even if many State constitutions and laws assign legislatures some role in making decisions about initiating and conducting war, not all systems

56 Lori Damrosch, 'The Interface of National Constitutional Systems with International Law and Institutions on Using Military Forces: Changing Trends in Executive and Legislative Powers' in Charlotte Ku and H Jacobsen (eds), *Democratic Accountability and the Use of Force in International Law* (Cambridhe University Press 2003) 39, 58 (noting that '[o]nly when military policies are fully debated and understood through the constitutional processes of democratic societies will there be sufficient assurance of public support for them').

57 Anne Peters, 'The (Non-)Judicialisation of War: German Constitutional Court Judgment on Rescue Operation Pegasus in Libya of 23 September 2015 (Part 1)' (*EJIL Talk!*, 21 October 2015) <https://www.mpil.de/files/pdf4/Peters_EJILTalk-The_Non-Judicialisation_of_War_Pegasus1.pdf>.

58 Tom Ginsburg, 'Chaining the Dog of War: Comparative Data' (2014) 15 Chicago Journal of International Law 138, 139 (discussing the democratic peace literature).

59 This is particularly true for multi-party systems, where legislatures are more likely to serve as a veto point. Legislatures in single-party systems or parliamentary systems in which the executive comes from a strong majority party may play a weaker role in checking the executive's resort to force.

60 Ginsburg (n 58) 146.

work identically. Some constitutions envision a role for legislatures to approve the use of force or troop deployments *ex ante*, while others authorize legislatures to approve or condemn executive decisions *ex post*. Legislatures also may oversee the executive's military strategy, hold votes of 'no confidence' and approve conflict-related expenditures. This section briefly details these distinct roles to set the stage for understanding how cyber autonomy might affect these roles in the future.[61]

1 *Authorizing Force ex ante*

Some constitutional systems envision a role for legislatures in authorizing force *ex ante.* The Czech Republic, Denmark, Germany, Hungary, Italy, Norway, Netherlands, Sweden and Mexico all ostensibly require prior parliamentary approval before the executive may send troops abroad.[62] In Sweden, for instance, the government can only send armed forces abroad in accordance with a specific law that sets forth the grounds for such action.[63] The German Constitutional Court has held that German armed forces can only be deployed abroad for non-defensive purposes with prior legislative approval.[64] In contrast, the legislatures of Belgium, Canada, France, Spain, the United Kingdom and the United States lack the right of prior authorization in most cases.[65]

In the United States, for instance, the executive currently interprets the Constitution to allow it to use force abroad without advance congressional authorization except in a limited set of cases in which the number of troops and the circumstances in which they would be deployed rise to the level of 'war in a constitutional sense'.[66] In the United Kingdom, the British government possesses prerogative powers to deploy the UK armed forces, and therefore historically did not seek legislative permission in advance to do so. In 2011, however, the government acknowledged that a new expectation had emerged that the House

61 Hans Born and Heiner Hänggi, 'The Use of Force under International Auspices: Strengthening Parliamentary Accountability' (Geneva Centre of the Democratic Control of Armed Forces 2005) <https://www.dcaf.ch/sites/default/files/publications/documents/pp07_use-of-force.pdf>.

62 ibid 8 (including citations to relevant provisions). For Mexico, see *Constitución Política de los Estados Unidos Mexicanos* [Constitution] art 89, § VIII (giving the President the power to declare war, 'having the previous authorization of the Congress') art 73, § XII (giving Congress the power to declare war). Of course, the start of a cyber conflict would not entail sending troops abroad, but could quickly transition to that.

63 Born and Hänggi (n 61) 7; Government of Sweden, Sveriges Riksdag, *The Constitution of Sweden: The Fundamental Laws and the Riksdag Act* (2016) 50 <https://www.riksdagen.se/globalassets/07.-dokument--lagar/the-constitution-of-sweden-160628.pdf>.

64 Russ Miller, 'Germany's Basic Law and the Use of Force' (2010) 17 Indiana Journal of Global Legal Studies 197, 202.

65 Born and Hänggi (n 61) 6 and 7.

66 See, eg, Memorandum from Assistant Attorney General Steven A Engel to Counsel to the President, April 2018 Airstrikes Against Syrian Chemical-Weapons Facilities (31 May 2018) <https://www.justice.gov/olc/opinion/file/1067551/download>.

of Commons would have the chance to debate the deployment of military forces in advance, except in an emergency.[67] That new convention was put to the test when the UK government sought legislative approval in 2013 for military action in Syria and Parliament voted it down. However, the UK undertook limited airstrikes against Syrian chemical weapons capabilities in 2018 without consulting Parliament first, suggesting that the government will only follow the convention where possible military action is premeditated and will entail the deployment of military forces in an offensive capacity.[68]

One obvious benefit to legislative participation in decisions to resort to force in the first instance is that legislatures can constrain 'overzealous executives by requiring evidence to justify wars'.[69] As Ginsburg notes, the Framers of the US Constitution believed that congressional involvement in decisions related to force would slow down war-making except in true emergencies. For democracies today, such deliberation may help '"screen" wars: ensuring that the conflicts that the nation enters into are "good" wars, while eschewing "bad" wars'.[70]

A constitutional requirement of *ex ante* authorization is a powerful tool for legislatures compared to *ex post* powers because the introduction of troops often operates as a one-way ratchet. Once a State has committed troops to a conflict, legislatures have a hard time voting to withdraw those troops because doing so may be seen by the public as unpatriotic or a sign of weakness.[71] Therefore, legislatures that have a role in authorizing force *ex ante* have far more leverage in the decision-making process than do those whose only authorizing role arises after the fact.

Nevertheless, most systems that give their legislature *ex ante* powers include an exception that allows the executive to respond to imminent attacks or emergencies without advance legislative approval.[72] Even the

67 United Kingdom, House of Commons, *Parliamentary Approval for Military Action* (17 April 2018) <https://commonslibrary.parliament.uk/research-briefings/cbp-7166/>.

68 ibid. Most uses of highly autonomous cyber operations would not meet that test.

69 Ginsburg (n 58) 146.

70 ibid 142, 145; Yasuo Hasebe, 'War Powers' in Michel Rosenfeld and Andras Sajo (eds), *Oxford Handbook of Comparative Constitutional Law* (Oxford University Press 2012) 465 (noting that legislative approval for armed force provides more legitimacy and popular support for the operations).

71 See, eg, *Mitchell v Laird*, 488 F2d 611 (DC Cir 1973) (discussing why members of Congress who opposed the continuation of the Vietnam War might nevertheless vote to appropriate money, to avoid abandoning the forces already fighting).

72 See, eg, *Regeringsformen* [Constitution] 15:13 (Sweden) (giving the government the right to deploy Swedish armed forces to meet an armed attack on Sweden or prevent a violation of Sweden's territory); *The Prize Cases*, 67 US (2 Black) 635 (1863) (implying a presidential 'repel attacks' power); *Grondwet voor het Koninkrijk der Nederlanden* [Constitution] art 96, sub 2 (Netherlands) ('approval [for a declaration of a state of war] shall not be required in cases where consultation with Parliament proves to be impossible as a consequence of the actual existence of a state of war'), *Glasilo Uradni List Republike Slovenije* [Constitution] art 92 (Slovenia); *Eesti Vabariigi põhiseadus* [Constitution] arts 65, sub 15, 128 (Estonia); *Türkíye Cumhuríyetí Anayasasi* [Constitution] art 92 (Turkey).

laws of a State such as Germany, in which both the legislature and the judiciary play significant roles in decisions about the resort to force, contemplate that there will be situations of 'imminent danger' in which the executive must act on its own without pre-approval by the legislature.[73] In such cases, however, the executive must promptly seek approval from the German parliament afterwards.[74]

One way that legislatures can implement their *ex ante* authority is to enact laws that stipulate the settings in which and adversaries against whom the executive is authorized to use force. In the United States, these often take the form of Authorizations to Use Military Force (AUMFs). In a little-noticed statute in 2018, Congress accorded the President authority akin to an AUMF for certain cyber operations. Section 1642 of the John McCain National Defense Authorization Act (NDAA) for FY 2019 states,

> In the event that the National Command Authority [i.e., the President and the Secretary of Defense] determines that the Russian Federation, People's Republic of China, Democratic People's Republic of Korea, or Islamic Republic of Iran is conducting an active, systematic, and ongoing campaign of attacks against the Government or people of the United States in cyberspace, … the National Command Authority may authorize the Secretary of Defense, acting through the Commander of the United States Cyber Command, to take appropriate and proportional action in foreign cyberspace to disrupt, defeat, and deter such attacks ….[75]

When the Defense Department employs this authority, the Secretary of Defense must report to the congressional defense committees no later than forty-eight hours after the operation; must include the actions in a quarterly report to the defense committees; and must report annually to the congressional defense, intelligence and foreign affairs committees about the 'scope and intensity' of the cyber attacks on the United States.[76] Although the provision does not resemble most of Congress's *ex ante* force

73 Peters (n 57).

74 *Parlamentsbeteiligungsgesetz* [Parliamentary Participation Act] § 5 (Germany); see also Hasebe (n 70) 478 (noting that the Japanese Self-Defense Forces Act provides that the Diet (national legislature) must authorize force in advance, except when there is no time to obtain such authorization, and that the Prime Minister 'may order the engagement of the [Self-Defence Forces] when an attack is clearly imminent and the necessity of the engagement is recognized').

75 John S McCain National Defense Authorization Act for Fiscal Year 2019, Pub L No 115-232, § 1642(a)(2), 1642(c), 132 Stat 1636 (2018) ('2019 NDAA').

76 ibid.

authorizations, 'it is an AUMF of a very narrow and specific variety'.[77] Part IV considers the effect of cyber autonomy on authorizations like this one.

2 *Ratifying Force ex post*

Another role for legislatures is to ratify or shape the executive's use of force *ex post*. Ginsburg, who reviewed 745 constitutions that entered into force since 1789, noted that since the early 1800s, constitutions have tended to assign the executive the power to resort to force. However, 'legislatures retain a major role in war policy' because they retain the power after the fact to approve or strike down the executive's decision to resort to force or to deploy troops.[78] France's current constitution, for instance, anticipates that its National Assembly must authorize declarations of war but 'includes no requirement that parliamentary authorization be prior to the declaration of war'.[79] For uses of force short of war, which include many forcible acts, the French executive must notify the Assembly of its decision to forcibly intervene abroad no later than three days after the intervention. The Assembly can debate the question, but does not actually vote on it, though if the intervention exceeds four months, the executive must ask the Assembly to authorize that extension.[80] Some States envision greater legislative control *ex post*. The laws of Denmark, Germany and the Netherlands, for example, contemplate not only that those legislatures will have powers of prior authorization but also that they will have the opportunity to subsequently approve the mission's mandate, operational guidelines and duration.[81]

Under a model of *ex post* legislative approval, it is possible that the executive will reject or ignore subsequent legislative condemnation of its troop deployments or other military operations. As noted above, though, the more likely scenario is that legislatures will find it hard not to support executive decisions, at least where the executive is responding to an actual attack on the country or where it has committed troops already. There is more political room for a legislature to condemn after the fact the executive's decision to use force or deploy troops where the forcible episode is completed quickly or there are few troops on the ground overseas.

77 Robert Chesney, 'The Law of Military Cyber Operations and the New NDAA' (*Lawfare*, 26 July 2018) <https://www.lawfareblog.com/law-military-cyber-operations-and-new-ndaa> (noting that the US Congress has enacted at least two other provisions that bolster the Defense Department's ability to undertake cyber operations when appropriately authorized to do so); see 2019 NDAA (n 75) § 1632; 10 USC § 394 (2019); National Defense Authorization Act for Fiscal Year 2012, Pub L 112–81, § 954 (2011), 125 Stat 1551.

78 Ginsburg (n 58) 149–50.

79 Hasebe (n 70) 473 (discussing Article 35 of the French Constitution).

80 Hasebe (n 70) 474–5.

81 Born and Hänggi (n 61) 8.

3 *Funding and oversight*

In addition to helping to regulate the initiation, conduct and cessation of military operations, legislatures play at least two other significant force-related roles. First, legislatures fund the military operations. This power of the purse can provide significant leverage over how and where the executive conducts those operations and the length of time for which the executive can fight. Like *ex post* ratifications, however, legislators may feel pressure to continue to fund conflicts they do not support because withholding funds from the troops risks seeming unpatriotic.[82]

Second, legislatures can conduct oversight for the duration of the conflict, to examine how the executive is conducting the conflict, whether it is exceeding its mandate, whether it is using resources wisely and whether the armed forces are complying with international and domestic laws.[83] Depending on the capacity of the legislative committees tasked with oversight responsibilities, these legislators can play an important role in holding the executive accountable for illegal, incompetent or unwise military and policy decisions.[84]

Even though most States authorize their executives to act without legislative approval in the face of imminent attacks, legislatures have a range of roles to play in authorizing their executives to use force, demanding justifications from the executives about the decision to enter into a conflict and generally enhancing democratic accountability for warfighting. A legislature's ability to enhance its executive's compliance with public law values — including international law — depends on a reliable flow of information between the executive and the legislature; on the legislature's competence to understand the strategy, tactics and tools that the executive is using; and on adequate time to make informed decisions. The introduction of significant levels of cyber autonomy into the mix is likely to complicate these already-challenging tasks.

82 See *Mitchell v Laird*, 488 F2d 611 (DC Cir 1973).

83 One salient example here is the US Congress's decision to try to terminate President Reagan's funding of the Contras in Nicaragua. See Boland Amendment, Pub L No 98-473, § 8066(a), 98 Stat 1837 (1984).

84 Ashley Deeks, 'Secrecy Surrogates' (2020) 106 Virginia Law Review 1395 (discussing these qualities as public law values).

IV
THE EFFECT OF CYBER AUTONOMY ON DEMOCRATIC ACCOUNTABILITY

Burgeoning cyber autonomy may affect democratic accountability for the use of force — as well as domestic checks and balances — in at least three ways. First, it may alter the balance of power between legislatures and executives, further empowering executives at the expense of legislative input about the timing, scope and legality of particular uses of force or offensive cyber operations. Second, it may alter the balance among a state's executive agencies. Third, it may alter power dynamics among different types of officials within those agencies. If obtaining the input of a diversity of executive officials and securing a legislative role in decisions about the use of force helps improve the quality of decision-making, the overall effect of robust uses of cyber autonomy may be to increase the potential for 'bad' conflicts between States.[85]

A ALTERING THE BALANCE BETWEEN LEGISLATURES AND EXECUTIVES

There are several ways in which autonomous cyber capabilities might further empower executives at the expense of the legislative role in force decisions, an imbalance that seems to dominate most governmental regimes today.[86] First, legislatures may suffer from information deficits about the existence and capabilities of the cyber systems. Second, there may be fewer opportunities temporally for legislators to weigh in about the wisdom of forcible responses. Third, executive reliance on highly autonomous systems may make it very hard for legislators to provide meaningful oversight *ex post*.

85 See Ginsburg (n 58) 145.

86 I do not mean to suggest that the growing autonomy of cyber operations is the only aspect of these operations that poses a threat to legislative capacity and oversight. For instance, the increased precision of cyber tools means that they can produce a more potent effect on the intended victim, which could increase the risks of escalation. Further, the growth of the Internet of Things and the interconnectedness of many publicly- and privately-owned systems means that there are more ways for a State's cyber operations to go wrong and have cascading, unintended effects. As with the growing autonomy of cyber systems, both of these developments make it critical for Congress to retain a role in oversight.

1 *Information Deficits*

Assume that a State's military develops autonomous cyber systems that can operate offensively or counter-offensively. An initial concern might be that legislators are unaware that the autonomous systems exist. Although legislatures sometimes appropriate money for specific programs, appropriations laws may not necessarily articulate in detail the types and nature of weapons that militaries are and are not authorized to develop. Legislators may also have difficulty obtaining information about executive cyber doctrines that will guide how the executives will utilize their cyber tools — including autonomous tools. In the United States, even though Congress has well-staffed committees that oversee the defense and intelligence agencies, and recently has legislated with particularity in the cyber area, Congress had difficulty gaining access to a classified US executive policy that sets out the approval process for conducting offensive cyber operations.[87] It stands to reason that Congress — let alone the legislatures of other States — might also have problems obtaining information about the extent of the human role in those cyber operations.

As a related matter, even if militaries share information with legislators about their cyber capabilities or doctrines, legislators may have difficulty understanding particular cyber capabilities, including autonomous capabilities and the risks attendant thereto. There are many reasons to think that the average legislator is not particularly savvy about technology.[88] In one salient example, several US senators proposed legislation in 2016 that would have required companies to provide the government with access to encrypted data when a court had so ordered. Critics savaged

87 Mark Pomerleau, 'After Tug-of-War, White House Shows Cyber Memo to Congress' (*Fifth Domain*, 13 March 2020) <https://www.fifthdomain.com/congress/2020/03/13/after-tug-of-war-white-house-shows-cyber-memo-to-congress/> (describing a multi-month struggle to obtain access to National Security Presidential Memorandum 13).

88 See Ashley Deeks, 'Facebook Unbound?' (2019) 105 Virginia Law Review Online 1, 6–7 (noting that members of Congress lack sophisticated understandings of how new technologies work); Matthew Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 Harvard Journal of Law & Technology 353, 380 (noting that 'only the small subset of the legislature that sits of the relevant committee will hear the experts' testimony, and even those legislators cannot afford to spend an inordinate amount of time conducting hearings on any on particular issue'); Karen Hao, 'Congress Wants to Protect You from Biased Algorithms, Deep Fakes, and Other Bad AI' (*MIT Technology Review*, 15 April 2019) <https://www.technologyreview.com/2019/04/15/1136/congress-wants-to-protect-you-from-biased-algorithms-deepfakes-and-other-bad-ai/> (noting that 'only a handful of members of Congress have a deep enough technical grasp of data and machine learning to approach regulation in an appropriately nuanced manner'); Julia Black and Andrew Murray, 'Regulating AI and Machine Learning: Setting the Regulatory Agenda' (2019) 10 European Journal of Law and Technology 3, s 5 <https://ejlt.org/index.php/ejlt/article/view/722> ('[T]here is little evidence that regulators have the necessary capacity properly to evaluate all the actual and potential uses of AI in their regulatory domains. Asymmetries of knowledge and skills are amplified in the highly technical area of AI'.).

the bill, not only because they objected to the policy but also because the bill seemed to reflect a flawed understanding of encryption technology.[89]

To counter this deficit, the US Government Accountability Office — an agency within the legislative branch — has proposed setting up a new office to help Congress understand the impacts of technology-related policies that it pursues,[90] and others have suggested reviving the now-defunct Office of Technology Assessment, which provided Congress with scientific expertise to match that of the Executive Branch.[91] In the UK, a joint parliamentary committee has recommended that the Government Office for Artificial Intelligence and the Centre for Data Ethics and Innovation — which will consist of technical and ethics experts — should identify for Parliament any gaps in existing regulations, suggesting that Parliament itself must rely on outside experts for artificial intelligence-related analysis.[92] Legislatures with small defense committees may face particular challenges in overseeing cyber operations generally — to say nothing of highly autonomous cyber operations — because their legislators presumably are spread more thinly across issue areas. Further, if they have small budgets, they will be able to employ fewer staffers and can convene fewer hearings in which outside experts could help them understand the issues and technologies they confront.[93]

Even legislators with a basic understanding of cyber operations may not have a full appreciation for the risks of autonomous operations and may not be positioned to ask the right questions of the executive branch.

89 Julian Sanchez, 'Feinstein-Burr: The Bill that Bans Your Browser' (*Just Security*, 29 April 2016) <https://www.justsecurity.org/30740/feinstein-burr-bill-bans-browser/>.

90 Jack Corrigan, 'Inside GAO's Plan to Make Congress More Tech-Savvy' (*NextGov*, 20 March 2019) <https://www.nextgov.com/cio-briefing/2019/03/inside-gaos-plan-make-congress-more-tech-savvy/155689>; Cat Zakrzewski, 'These Scientists Are Trying to Help Congress Get Smarter About Tech' (*Washington Post*, 27 January 2020) <https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/01/27/the-technology-202-these-scientists-are-trying-to-help-congress-get-smarter-about-tech/5e2b1fcc602ff14e6605928f/>.

91 US Government Accountability Office, 'Office of Technology Assessment' (13 October 1977), <https://www.gao.gov/products/103962>. See also US House of Representatives, Congressional Artificial Intelligence Caucus <https://artificialintelligencecaucus-olson.house.gov> accessed 14 October 2020 (describing the 'AI Caucus' in Congress, created to 'inform policymakers of the technological, economic and social impacts of advances in AI' by bringing together academics, private sector officials, and government officials); Mike Miesen and others, 'Building a 21st Century Congress: Improving Congress's Science and Technology Expertise' (Belfer Center for Science and International Affairs, September 2019) <https://www.belfercenter.org/publication/building-21st-century-congress-improving-congresss-science-and-technology-expertise> (discussing Congress's demand for science and technology expertise and the root causes of its lack of technological capacity); Caroline Kenny and others, 'Legislative Science Advice in Europe: The Case for International Comparative Research' (2017) 3 Palgrave Communications 17030 (discussing the role for scientific advice in legislatures in the UK and Europe).

92 United Kingdom, House of Lords, Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?' (2018) [386] <https://ec.europa.eu/jrc/communities/sites/jrccties/files/ai_in_the_uk.pdf>; see also United Kingdom, Office for Artificial Intelligence, <https://www.gov.uk/government/organisations/office-for-artificial-intelligence>.

93 For example, Hungary's Defense Committee had a budget of €4,000 ($4,800) in 2004: Born and Hänggi (n 61) 10.

Indeed, not all of the executive branch officials involved in decision-making may understand the capabilities and risks of complex, highly autonomous cyber systems. In the context of electronic surveillance systems, for example, in 2013 the US Director of National Intelligence (DNI) declassified a set of documents that revealed a lack of compliance with judicial mandates. The DNI explained that the compliance problems

> stemmed in large part from the complexity of the technology employed in connection with the bulk telephony metadata collection program, interaction of that technology with other NSA systems, and a lack of a shared understanding among various NSA components about how certain aspects of the complex architecture supporting the program functioned. These gaps in understanding led, in turn, to unintentional misrepresentations in the way the collection was described to the FISC.[94]

If some intelligence officials within a single agency were unclear about how the technology supporting an electronic surveillance program worked, it is easy to imagine how legislators would have had trouble understanding that program and — likewise — how they might struggle to understand very technical cyber tools that include significant levels of autonomy.

To some extent, this lack of understanding reflects a broader societal challenge posed by systems that rely on machine-learning tools. Those systems are often described as 'black boxes' because the weight that they give to factors within the data to reach predictions or recommendations is generally opaque. As a result, not only legislators but humans generally find it difficult to interpret or explain the outputs of systems that operate with high levels of autonomy. Computer scientists and militaries are keenly aware of this problem and are working to produce 'explainable' or 'interpretable' artificial intelligence, sometimes referred to as 'white box' models. As discussed below, legislatures have an opportunity to shape the level of explainability of the executives' cyber algorithms. Requiring executives to produce algorithms that are more transparent might also make it easier for legislators to hold executive actors accountable because

94 Office of the Director of National Intelligence, 'DNI Clapper Declassifies Intelligence Community Documents Regarding Collection Under Section 501 of the Foreign Intelligence Surveillance Act (FISA)' (Press Release, 10 September 2013) <https://www.dni.gov/index.php/newsroom/press-releases/press-releases-2013/item/927-dni-clapper-declassifies-intelligence-community-documents-regarding-collection-under-section-501-of-the-foreign-intelligence-surveillance-act-fisa>.

transparent algorithms might be easier to audit after the fact than human decisions are.

2 *Limited Opportunity for Legal and Policy Input*

In some States, legislatures can constrain 'overzealous executives by requiring evidence to justify wars'.[95] This is primarily true when the State's system contemplates legislative approval for the use of force *ex ante*. It also assumes that there is time for legislative input before the executive makes a decision to resort to force. But the US executive branch, for one, has taken the view that very few uses of force require congressional pre-authorization. If the only time *ex ante* congressional authorization for military operations is legally necessary is when the United States plans to deploy hundreds of thousands of troops abroad, cyber operations — whether human-in-the-loop or out-of-the-loop — will almost never reach the threshold of 'war in a constitutional sense'.[96] Hostile cyber exchanges, at least when the salvos remain within the cyber realm, are unlikely to pose an immediate and significant threat to US troops and will not trigger the need for congressional authorization under the 'Declare War' clause. Yet autonomous cyber systems may pose a reasonable chance of escalation — whether intended or unintended — such that legislative input might be normatively desirable *ex ante*. Even for States whose legal systems contain a clear *ex ante* requirement for legislative authorization, that authorization may be limited to troop deployments, which will not cover cyber exchanges, or may contain an emergency carveout, which would cover responses to sudden cyber attacks.[97]

As noted above, the US Congress has already provided limited *ex ante* authorization for the executive to 'take appropriate and proportional action in foreign cyberspace to disrupt, defeat, and deter such attacks' when those systematic attack campaigns come from Iran, North Korea, Russia or China.[98] This provision may actually serve as a limitation on

95 Ginsburg (n 58) 146.

96 Matthew Waxman, 'Cyber-Attacks and the Constitution' (Aegis Series Paper No 2007, Hoover Institution 2020) 4–5 <https://www.hoover.org/sites/default/files/research/docs/waxman_webready.pdf > ('If war powers are a special constitutional category demanding formal congressional approval because of the risks to American blood, most cyber-attacks barely if at all implicate this concern, because the risks are so tiny and remote.'); Eric Jensen, 'Future War and the War Powers Resolution' (2015) 29 Emory International Law Review 499, 541 (noting that the War Powers Resolution's reporting threshold fails to encompass cyber operations).

97 That said, in the US, the President often complies with statutory restrictions on his use of the military, even as he asserts constitutional objections to those statutes. See David Barron and Martin Lederman, 'The Commander in Chief at the Lowest Ebb: A Constitutional History' (2008) 121 Harvard Law Review 941. Thus, it would be worthwhile for Congress — and possibly other legislatures — to carefully consider how to set boundaries on the use of autonomous cyber tools.

98 See 2019 NDAA (n 75) § 1642.

the use of autonomous cyber systems, as it requires the executive to identify the source of the hostile cyber campaign. Unless the executive's autonomous cyber system is crafted to respond only to hostile operations that bear attack signatures from the named States, the executive would have difficulty relying on this authorization to support the use of such a system.[99] As discussed in Part V, legislatures should consider providing this kind of advance authorization, which can both serve as permission for and constraint on the use of cyber autonomy.

3 *Time Constraints*

As a related matter, highly autonomous cyber systems narrow significantly whatever consultative role legislatures may retain for themselves, at least in the window before a specific forcible cyber exchange takes place. The most significant reason to deploy autonomous cyber tools is to allow the system to operate at lightning speeds. Yet it is already the case today — before the widespread use of highly autonomous cyber tools — that executives, acting in response to perceived imminent threats of armed attacks on their States, employ force without legislative approval or even consultation. These threats may mostly come from terrorists today, but it is increasingly possible to conceive of cyber attacks as creating situations in which executive officials will need to respond in a very short time frame.

Purely defensive autonomous cyber operations — those that use autonomy only to identify and fend off hostile cyber operations within one's own system — are unlikely to implicate congressional prerogatives, as these settings will fall within the executives' 'repel attacks' powers found in many States' constitutions. But 'offensive' cyber capabilities that leave one's own system,[100] even in an act of self-defense, are more likely to implicate those prerogatives because they increase the chance of escalation and error. Further, autonomous systems 'may operate at speeds that make it impossible for the operator to meaningfully intervene'.[101] Thus, once a State deploys an autonomous cyber tool that has the capacity to reach outside that State's own system and inflict substantial harm, there will be no opportunity for congressional consultation on particular operations.

99 However, the US executive might conclude that it could rely on its broad Article II powers, including the commander-in-chief power, under the Constitution, even if it lacked specific statutory authority to act. It is also possible that providing legislative authorization for the executive to use autonomous responses to cyber operations only when they come from certain States will stimulate other States to engage in false-flag attacks from one of the named States in an effort to escalate cyber hostilities between the victim State and the named State.

100 See Liivoja, Naagel and Väljataga (n 14) 12–13.

101 ibid 15.

4 *Challenges to ex post Oversight*

One of the more reliable roles for legislatures during a conflict is the provision of oversight. A legislative body can help unearth how conflicts started, whether the State is achieving its military and strategic goals and whether it is complying with domestic and international laws during the fight. Legislatures often rely on executive actors to provide information about the conflict, but legislators can also convene hearings of outside experts and collect open-source intelligence about the situation from journalists on the ground.

Cyber hostilities, particularly those conducted by highly autonomous systems, will be far harder to understand and oversee. Conducting forensic audits that recreate what happened during a cyber exchange and translate them into language that congressional overseers can understand will be more challenging than reviewing radar patterns or identifying the source of limpet mines found on oil tankers.[102] The use of artificial intelligence to facilitate autonomy will pose 'black box' problems for legislators who seek to audit how the cyber operations played out. Further, there will be no 'war zone' to which journalists or outside analysts can travel to talk to troops on the ground about what they are seeing. As a result, there will be far fewer open-source reports about what has transpired during these 'invisible' cyber operations, unless and until they morph into kinetic conflicts.

In the United States, Congress has begun to address this potential lack of visibility by mandating that the executive report to it after conducting certain types of cyber operations. As Matthew Waxman notes,

> Congress has mandated special reporting requirements for offensive and 'sensitive' cyber-operations to the armed services committees.[103] Cyber-attacks conducted as covert action by the CIA would be reported separately to the intelligence committees, as would other intelligence activities that might fit within the definition here of cyber-attacks. Such reporting is foundational to other congressional roles, because it keeps Congress — or at least

102 'Iran News: US Says Mines Used in Tanker Attacks Bear "Striking Resemblance" to Weapons Touted by Tehran' (*CBS News*, 19 June 2019) <https://www.cbsnews.com/news/iran-news-us-shows-limpet-mine-parts-case-against-iran-in-tanker-attacks-today-2019-06-19/>.

103 The 2013 NDAA required the Department of Defense to 'provide to the Committees on Armed Services of the House of Representatives and the Senate quarterly briefings on all offensive and significant defensive military operations in cyberspace carried out by the Department of Defense during the immediately preceding quarter'. National Defense Authorization Act for Fiscal Year 2013, Pub L No 112-239, § 939, 126 Stat 1632 (2012); 10 USC § 484 (2011). Congress updated and expanded this provision in the 2017 and 2019 NDAAs.

> certain committees — informed of executive branch actions that would otherwise be largely invisible.[104]

Existing statutes require the US military to report to the congressional defense committees within forty-eight hours when it conducts a cyber operation determined to have a medium or high probability of political retaliation, detection or collateral effects and is intended to cause effects in an area in which the United States is not already involved in hostilities.[105] This kind of requirement is helpful — at least on its face — because it puts some members of Congress on notice of situations that might lead to conflict. But a situation between two States could escalate significantly within forty-eight hours, particularly if the States involved are using autonomous systems that are not adequately engineered to avoid escalation and to minimize risks of misdirecting responses. Further, it is not yet clear how these reporting rules are functioning and whether Congress is receiving the information that it believes it needs to provide adequate oversight.[106]

B ALTERING THE BALANCE AMONG EXECUTIVE AGENCIES

The growth of autonomous cyber systems is likely to further alter the current balance between executives and legislatures in use of force decisions. But the use of autonomous cyber tools also has the potential to affect the balance of power within executive branches themselves. One interesting question is whether the use of high levels of cyber autonomy will continue to push power out to the militaries as the creators and operators of these autonomous tools, or whether it offers an unexpected opportunity to readjust and centralize the locus of some of the decision-making associated with these tools.

On its face, it might appear that highly autonomous cyber tools will empower militaries at the expense of other executive agencies that have important equities in foreign policy decision-making, such as foreign and justice ministries. Even if these other agencies are involved in discussions about cyber strategy, they likely lack the technological sophistication that

104 Waxman (n 96).
105 10 USC § 395 (2019).
106 Robert Chesney, 'The Domestic Legal Framework for US Military Cyber Operations' (Aegis Series Paper No 2003, Hoover Institution 2020) 15 <https://www.hoover.org/sites/default/files/chesney_webreadypdf.pdf>.

military coders and cyber operators possess and so may have difficulty understanding whether highly autonomous cyber tools advance or hinder certain policy objectives and what level of risk these systems pose. Further, as with any military operation, those who sit closest to the point of execution have the greatest power to make last-minute decisions and adjustments. Although autonomous systems will take some of that control from those cyber operators, those operators nevertheless have more direct 'eyes on' the operations and their effects. In the United States, Congress's recent legislative acts seem to have enabled this. As Waxman notes, 'Congress has clarified the Defense Department's authority to conduct offensive cyber-operations, thereby strengthening its position within the executive branch and facilitating action by alleviating legal doubts about its mandate'.[107]

However, there is a possibility that increased autonomy could reverse this flow of power to militaries. Increased autonomy in warfighting tasks may — perhaps ironically — offer the opportunity to centralize decision-making, as the process of building machine-learning algorithms for warfighting systems, including cyber systems, seeks to incorporate the commander's intent and remain sensitive to legal constraints. These centripetal forces may even mean that other national security agencies begin to play a role in developing the policies undergirding those algorithms.[108] In the United States, the National Security Council and the State Department, for instance, may seek to inform the algorithms' contents and structure to ensure that they comply with the laws of armed conflict and the UN Charter.

Today, the US military has a well-established weapons review process; non-military lawyers are not involved. Likewise, judge advocates provide legal advice to commanders during armed conflict without consulting the Defense Department's Office of the General Counsel, let alone the National Security Council or other executive agencies. And yet there may be pressure to adjust the traditional process when the government builds machine-learning systems that can undertake autonomous action during conflict. If the use of the system will have significant foreign relations implications and if the system's recommendations implicate legal questions that already have been the subject of significant interagency

107 ibid 10–11 (referring to 2012 NDAA, Pub L No 112-81, § 954 (2011)); 10 USC § 111 (2011); National Defense Authorization Act for Fiscal Year 2018, Pub L No 115-91, §1633(a), §1633(b)(5)(B), 131 Stat 1283 (2017).

108 Some of the discussion in this section is drawn from Ashley Deeks, 'Will Autonomy in US Military Operations Centralize Legal Decision-Making?' (*Articles of War*, 5 August 2020) <https://lieber.westpoint.edu/autonomy_military_operations_decision-making/>.

interest, other agencies' policymakers and lawyers may demand a role. The lawyers might want to craft guidance in advance about what types of autonomous cyber tools would or would not meet underlying international law standards, for instance. And because the coding process will involve decisions about the nuances of that law and will happen before the system is deployed, there may be greater opportunities for a broader set of US government actors to claim a stake in those decisions than there is in kinetic lethal operations downrange.

There would be both benefits and costs to such a development. Militaries likely would perceive this potential centralization of decision-making as unattractive and might resist sharing the authority to make algorithmic choices about autonomous cyber tools. Interagency lawyers might also struggle to reach consensus about what features to incorporate into those tools. On the other hand, obtaining interagency understanding and acceptance of autonomous cyber tools would bolster the military's confidence about their use and would also allow that State's diplomats and foreign ministry lawyers to engage more deeply with allies on what may be controversial uses of machine learning and cyber tools.

Whether the growth in cyber autonomy ends up diminishing or increasing the role of non-military executive agencies will depend on decisions made by legislatures, choices by executive branch leadership, and the efforts (or lack thereof) of civilian national security agencies to help define the parameters of autonomous cyber tools as they are developed.

C ALTERING THE BALANCE WITHIN EXECUTIVE AGENCIES

Finally, within individual executive agencies, autonomous cyber tools, like other high-technology tools, will almost inevitably empower operators and computer scientists over lawyers. As I have noted elsewhere, in contexts driven by high-technology problems, data scientists will become relatively more important to policymakers than they have been in the past, and senior officials may start to treat their input as just as important to an international law or foreign policy decision as that of their international lawyers. [109] In my view, 'It will be the data scientists who can suggest new text-as-data tools and interpret the results of existing models. This means that the data scientists who embrace and understand the problems

109 Ashley Deeks, 'High-Tech International Law' (2020) 88 George Washington Law Review 574, 647.

that international lawyers and diplomats face will be most effective in this setting'.[110] Among officials who are not cyber experts, military and civilian actors who are technologically literate will be empowered relative to those who disdain technology or are unable to grasp its basic capabilities, limitations, and risks.[111] Thus, lawyers and policymakers who seek to work with data scientists and programmers to understand autonomous cyber tools will gain power relative to their counterparts who cannot or will not do so.[112]

V
PRESERVING ACCOUNTABILITY

In light of the range of challenges to democratic accountability and oversight that high levels of cyber autonomy will pose, this Part considers steps that States might take to meet some of those challenges. A State's legislature, its executive branch and its allies all can take actions to ensure that the State's use of autonomous cyber tools remains responsive to democratic systems of governance.

A PRESERVING LEGISLATIVE PARTICIPATION

Legislatures could take at least two steps to help preserve a role for themselves in a world of autonomous cyber tools. First, they could bolster their own technological expertise and access to high-tech experts. Second, they could embrace the possibilities for legislation that sets appropriate parameters on the executive branch's development and use of highly autonomous cyber systems.

1 *Developing Expertise*

A range of scholars have suggested ways in which legislatures could improve their understanding of technology and thus enhance their ability to legislate intelligently about such issues. One underlying issue is a lack of resources: if legislatures want to be able to hire and retain

110 ibid.
111 See Linell Letendre, 'Lethal Autonomous Weapons Systems: Translating Geek Speak for Lawyers' (2020) 96 International Law Studies 274.
112 Deeks (n 109) 647.

technologically savvy staff, and conduct hearings that bring in a range of expert views on issues such as autonomous cyber tools, they need the funds to do so. In the United States, one think tank notes, 'Congress has simply not given itself the resources needed to efficiently and effectively absorb new information — particularly on complex [science and technology] topics'.[113] Others have advocated that the US Congress establish an internal body that is nimble, bipartisan and focused on providing options rather than recommendations.[114] Various European States have already established bodies that provide science and technology advice to legislatures; the United States could draw ideas from some of the different models represented there.[115] The European bodies should also ensure that they have experts at hand who understand machine learning and autonomous cyber systems, which will facilitate the legislators' ability to regulate such systems as they come online. Outside experts can be very useful here, both to educate legislatures and to surface and articulate competing views about the benefits and costs of this technology.

Legislatures should also consider setting up 'machine learning boot camps' for staffers who work on national security-related committees, to expose them to the basics of machine learning and cyber tools. Sessions run by outside tech experts who can present the information in clear, non-partisan, policy-relevant ways would be a helpful tool to ensure basic competence among policy and legal staff. In the United States, for example, Stanford University runs a 'Cyber and Artificial Intelligence Boot Camp' for congressional staffers. The boot camp draws on the experience of cybersecurity professionals, scholars, business leaders and lawyers to provide staffers with basic technical instruction, threat perspectives and exposure to simulated attacks.[116] Legislatures might also ask to observe actual testing and verification processes that take place inside the militaries, to understand how militaries decide that they have confidence in a particular autonomous system before deploying it.

2 *Updating Legislative Structures and Authorities*

In addition to raising their level of technological fluency, legislators should resist further erosion of their roles in overseeing the use of force and offensive cyber operations by updating their own ability to oversee

113 Miesen and others (n 91) 9.

114 Chris Tyler, 'Legislative Science Advice in Europe and the United Kingdom: Lessons for the United States' (*Lincoln Policy*) 8–9 <https://lincolnpolicy.org/wp-content/uploads/2020/02/TYLER.pdf>; Miesen and others (n 91).

115 Tyler (n 114).

116 Hoover Institution, 'Cyber and Artificial Intelligence Boot Camp' (August 2019) <https://www.hoover.org/events/cyber-and-artificial-intelligence-boot-camp-2019> accessed 14 October 2020.

cyber operations. One way to do this is to establish oversight committees dedicated specifically to cyber issues, as the recent Cyber Solarium project in the United States recommended. The Solarium report proposes that the US Congress create House and Senate committees on cybersecurity 'to provide integrated oversight of the cybersecurity efforts dispersed across the federal government'.[117] The committees, which presumably would draw their membership from existing armed services, intelligence and homeland security committees, could develop a deeper expertise on cyber issues — including the functions of autonomy in cyber settings — while building on their members' past experiences with war powers, use of force and technological questions.

Legislatures could also direct new regulatory efforts at autonomous cyber systems. For States in which existing statutes (rather than the constitution) allocate powers between the executive and legislatures, those legislatures should evaluate whether the statutes adequately reach cyber operations that either constitute or could quickly lead to international uses of force. In the United States, for example, the War Powers Resolution (WPR) creates a structure for executive consultation with and reporting to Congress before deploying armed forces into hostilities, but it quite clearly would not apply to the bulk of cyber operations, whether autonomous or not. One scholar has suggested amending the WPR to trigger the executive's notice requirement not only upon the introduction of troops but also upon the effectuation of military capabilities (such as cyber tools) in a situation that violates the sovereignty of another State.[118] This proposal might capture too many operations, however, especially if Congress's real interest lies in retaining some input into cyber operations that have the potential for escalation.

In any event, amending the WPR will be difficult, because the President would likely veto such changes. Thus, Congress would need to assemble a veto-proof majority that favors the bill.[119] But there may be more modest fixes that could achieve similar goals: in the United States, one adjustment might be to expand the list of committees that receive the forty-eight hour reports from the Defense Department under section 1642 of the 2019 NDAA.[120] That is, when the military has undertaken a 'sensitive military cyber operation' against Russia, China, North Korea

117 William Ford, 'The Cyberspace Solarium Commission Makes Its Case to Congress' (*Lawfare*, 18 May 2020) <https://www.lawfareblog.com/cyberspace-solarium-commission-makes-its-case-congress>.

118 Jensen (n 96) 553–54.

119 ibid (discussing legislative proposals to amend the War Powers Resolution).

120 2019 NDAA (n 75) § 1642; see also 10 USC § 395 (2019) (cross-referenced within § 1642).

or Iran, Congress should amend section 1642 to require that the military provide its written report not just to the armed services committees, but also to the intelligence and foreign affairs committees. Congress should also expand this notice requirement to cover sensitive military cyber operations against any State, not just these four States. Other legislatures should ensure that they are receiving adequate notice of significant cyber operations that implicate their regulatory and oversight powers.

Legislatures might also turn their attention specifically to the growing use of autonomous cyber tools, erecting guard rails around their use. Even if, as argued above, legislatures are not particularly well-suited to legislate in high-tech areas, legislatures should be able to navigate core legal and policy questions associated with autonomy.[121] First, legislatures should evaluate whether they are willing to accept their militaries' use of highly autonomous cyber tools generally. Some legislatures may accept the potential risks of such tools because they believe that the benefits are considerable. Others may not.

Second, those legislatures that accept in theory the use of autonomous cyber tools should define the basic contexts in which those tools are permissible, identify the adversaries against which the military may use the tools, define what kinds of foreseeable effects they are willing to tolerate, require the tools to be deployed in a way that is consistent with international legal requirements and require the executive to build in hard stops on escalation. Tim McFarland suggests, for instance, that a 'cyber weapon might be trusted to locate and identify potential targets autonomously, but be required to seek human confirmation before attacking them'.[122] The US Defense Department's Defense Innovation Board suggested that the department consider setting 'limitations on the types or amounts of force particular systems are authorized to use, the decoupling of various AI cyber systems from one another, or layered authorizations for various operations'.[123] Legislatures might fix in statute rules that require militaries to avoid uncontrolled escalation or impose the need for the effects of autonomous cyber operations to be reversible. They also could require that their executive branches only employ software in their cyber systems that is explainable or interpretable.

121 Liivoja, Naagel and Väljataga (n 14) 24.
122 McFarland (n 5) 33.
123 Defense Innovation Board, 'AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Supporting Document' (US Department of Defense, 31 October 2019) 30 <https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF> .

At a more granular level, legislatures might take advantage of the fact that many cyber tools, even those that are increasingly autonomous, still require humans to carefully identify the tools' targets in advance and tailor those tools specifically to that threat. Even if legislatures have significant difficulties weighing in on hostile cyber operations close to the time at which the executive initiates those operations (including by unleashing a largely autonomous system), the legislatures could seek information from executive cyber operators about the pre-positioning efforts that the operators have undertaken to be able to launch operations in the future. Even if those pre-positioning efforts may primarily be to gather intelligence rather than to conduct an offensive operation, their dual-use nature means that legislatures would be within their rights to understand where and how their militaries or intelligence services are poised to initiate future cyber operations.

If a legislature is worried about its own abilities to substantively understand autonomous cyber tools and the risks that they pose, it could establish a commission of independent experts — with appropriate security clearances — to review, analyze and report on executive branch conduct involving relevant technologies. Such a commission might examine compliance with both international and domestic law, and could report regularly to legislatures and, in an unclassified form, to the public. Precedent for these types of bodies include the Privacy and Civil Liberties Oversight Board in the United States and the Investigatory Powers Commissioner in the United Kingdom.[124]

Finally, legislatures should impose reporting requirements on executives so that legislators are aware of the types of autonomous cyber systems their militaries are using and what effects the systems are producing or have the capacity to produce. They might even require reports from foreign ministries on the foreign policy implications of any autonomous cyber operations that occur, thus ensuring that those ministries retain visibility into those operations.[125] These steps will help preserve a level of democratic accountability for uses of force or other escalatory cyber actions.

124 Joanna Dawson and Samantha Godec, 'Oversight of the Intelligence Agencies: A Comparison of the "Five Eyes" Nations' (House of Commons Library Briefing Paper No 7921, 15 December 2017) <https://commonslibrary.parliament.uk/research-briefings/cbp-7921/>.

125 In the US statute creating the Global Security Contingency Fund (the FY 2012 National Defense Authorization Act), Congress required a form of 'dual-key' authorization and reporting, whereby decisions about funding are made jointly by the Secretaries of State and Defense, and those agencies send reports jointly to multiple committees. Nina Serafino, 'Global Security Contingency Fund: Summary and Issue Overview' (Congressional Research Service Report No R42641, 4 April 2014) <https://fas.org/sgp/crs/row/R42641.pdf>.

B SECURING EXECUTIVE BALANCING AMONG AGENCIES

Legislators are not the only actors whose input may be threatened by increasingly autonomous military tools. As Part IV discussed, the operation of highly autonomous cyber tools might diminish the opportunities for civilian officials within the executive branch to provide input into activities that could produce major foreign policy consequences. Because the cyber tools that will perform these autonomous operations will be constructed in advance, however, there is an opportunity for a range of relevant agencies to provide input into the parameters of those systems. One way to do this is to establish standing rules of engagement to guide how the military deploys the systems, and to craft those rules of engagement through an interagency process.[126] This would give civilian officials insight into and influence on the ways that the military uses advanced autonomous cyber systems.

Even if interagency officials such as diplomats, career analysts and civilian national security lawyers are not directly engaged in crafting military rules of engagement, there is still room for interagency participation in developing the rules of the road for use of autonomous cyber tools. Two scholars recently noted, 'Insights from the literature on civil-military relations and planning suggest not leaving cyber strategy to soldiers alone'.[127] They add, 'There are major questions regarding how to craft a policy framework for cyber strategy that does not create dangerous escalation pathways or jeopardize civil liberties and the free flow of information. These questions should not be reduced to expediting authorities at the expense of interagency coordination or civilian oversight'.[128] These scholars propose developing 'flexible response options precleared to balance equities and assess risks', which would ensure 'time-sensitive responses without sacrificing interagency coordination'.[129]

126 Erica Borghard and Shawn Lonergin, 'What Do the Trump Administration's Changes to PPD-20 Mean for US Offensive Cyber Operations?' (*Council on Foreign Relations*, 10 September 2018) <https://www.cfr.org/blog/what-do-trump-administrations-changes-ppd-20-mean-us-offensive-cyber-operations> (noting that 'some risks [that attach to loosening interagency control over cyber operations on the tail end] can be mitigated through developing standing rules of engagement' that could 'mitigate some concerns about escalation' and that the process of establishing the rules of engagement could codify and address those concerns).

127 Benjamin Jensen and JD Work, 'Cyber Civil-Military Relations: Balancing Interests on the Digital Frontier' (*War on the Rocks*, 4 September 2018) <https://warontherocks.com/2018/09/cyber-civil-military-relations-balancing-interests-on-the-digital-frontier/> (arguing that letting soldiers plan in isolation produces 'narrow plans prone to escalation risks', leads to 'false optimism [and] overconfidence', and 'diminishes the probability of successful, coercive diplomacy').

128 ibid.

129 ibid.

The need for militaries to respond in a timely way is a real one; disorganized interagency processes can hinder that. Under the Obama Administration, the United States used an interagency cyber process that often got bogged down in infighting.[130] Its 'interagency de-confliction process suffered from delays, bureaucratic inertia, ill-defined decision pathways, and the lack of a clear "referee" to resolve competing positions at the working level'.[131] For example, there was a 'fierce debate' among different executive agencies about whether to notify States hosting computer services used by ISIS that the United States planned to sabotage those services, a dispute that took weeks to resolve.[132] The Trump Administration modified the interagency process, apparently delegating far more decisions about offensive cyber operations to military commanders and decreasing interagency input. Further, the Trump Administration seems to have authorized the CIA to undertake covert offensive cyber operations against several adversaries, and to do so with a new level of independence from the White House.[133]

Because the Trump policies and the subsequent operations under them remain classified, it is unclear whether these policies have produced better or worse results from a US foreign policy perspective.[134] In any event, developing an executive process that adequately balances the need for effective military cyber responses against harm to diplomatic, law enforcement and intelligence cooperation may take time and multiple iterations to get it right. In the United States, there is a debate, for example, about whether to create a 'National Cyber Director' to coordinate those responses or whether to rely on the National Security Council to do so.[135] Regardless of the specific mechanisms they use, States must

130 ibid.
131 ibid.
132 Ellen Nakashima, 'US Military Cyber Operation to Attack ISIS Last Year Sparked Heated Debate over Alerting Allies' (*Washington Post*, 9 May 2017) <https://www.washingtonpost.com/world/national-security/us-military-cyber-operation-to-attack-isis-last-year-sparked-heated-debate-over-alerting-allies/2017/05/08/93a120a2-30d5-11e7-9dec-764dc781686f_story.html>.
133 Zach Dorfman and others, 'Exclusive: Secret Trump Order Gives CIA More Powers to Launch Cyberattacks' (*Yahoo News*, 15 July 2020) <https://news.yahoo.com/secret-trump-order-gives-cia-more-powers-to-launch-cyberattacks-090015219.html>.
134 David Kris, 'What Hard National Security Choices Would a Biden Administration Face?' (*Lawfare*, 27 May 2020) <https://www.lawfareblog.com/what-hard-national-security-choices-would-biden-administration-face> ('A Biden administration will need to understand … exactly how much power to act now resides in military commanders; and how much interagency coordination is required before action can be taken. It might want to adjust protocols in favor of less delegation, particularly at the beginning of the administration when new officials are less aware of and comfortable with the precedents and understandings developed in the Trump administration'.); Eric Geller, 'Trump Scraps Obama Rules on Cyberattacks, Giving Military Freer Hand' (*Politico*, 16 August 2018) <https://www.politico.com/story/2018/08/16/trump-cybersecurity-cyberattack-hacking-military-742095> (discussing pros and cons of the interagency process under the Obama administration).
135 Philip R Reitinger, 'Establishing a National Cyber Director Would Be a Mistake' (*Lawfare*, 17 July 2020) <https://www.lawfareblog.com/establishing-national-cyber-director-would-be-mistake>.

preserve important elements of civilian control and oversight over autonomous military cyber operations as they try to strike the proper balance among their various security and foreign policy equities.

Just as legislative staff should improve their cyber literacy, so too should executive officials who work on cyber issues. Governments could detail national security lawyers in foreign, justice and intelligence ministries to technology offices in their own or other agencies. They could also detail cyber experts to policy positions, such as to positions in NATO or in their foreign ministries. This would have to be done in a way that rewards these officials for taking these non-traditional postings, along the lines of the requirement in the US Goldwater-Nichols Act that requires joint-duty assignments for military officers seeking career advancement. Further, like legislative staffers, executive agencies should mandate that those civilian officials who work on cyber and technology policy issues attend machine learning and cyber bootcamps to establish basic familiarity with those tools and their future prospects.

These measures, which would provide a form of internal checks and balances among different executive agencies, should improve the quality of executive decision-making. As I noted elsewhere:

> Particularly in the national security area, where Congress and the courts face institutional and structural challenges to providing robust oversight, it has become commonplace to turn to checks within the executive branch itself as an alternative to inter-branch checking. The inter-agency policy-making process requires — and indeed benefits from — exchanges among different executive agencies with distinct mission statements. Each agency pursues its own goals and policies, while trying to avoid policies that undercut the agency's mission or unduly weaken its standing in relation to other agencies.[136]

It therefore seems healthy to ensure that a range of civilian agencies and officials retains a role in shaping the use of highly autonomous cyber tools. This is particularly true because it may be hard for legislatures to serve in their constitutional checking role in relation to these tools. Cyber autonomy may be critical at the moment of an attack, but there is ample room in advance to shape that autonomy's characteristics and uses.

136 Ashley Deeks, 'A (Qualified) Defense of Secret Agreements' (2017) 49 Arizona State Law Journal 713, 776.

C ROLES FOR ALLIES AND OTHER EXTERNAL ACTORS

This article suggests that a range of States face some shared challenges when it comes to democratic accountability for the use of cyber autonomy. As a result, there may be value in sharing experiences among executive and legislative branches of NATO member States. Understanding how allied counterparts approach regulatory issues, deficiencies in technological knowledge and legal questions raised by highly autonomous military operations could produce creative ideas about ways to preserve and even bolster democratic accountability. Close allies might even consider sharing detailed information about their own autonomous systems, to identify and troubleshoot international legal issues.

Another source of constraint on executive actors undertaking classified national security operations, such as cyber operations, is US technology and cybersecurity companies. In some settings, these companies have incentives to check poor executive decision-making that happens behind the veil of classification. These actors often have access to incoming cyber threats, have independent tools by which to attribute attacks and have the expertise to observe and critique certain US government cyber operations.[137] The US Congress might do well to harness these 'surrogates' as information-gatherers and a source of technological expertise about the growing autonomy of cyber operations by the United States and other States.

VI CONCLUSION

Highly autonomous cyber operations are near at hand. Even if States manage them very carefully, the potential exists for States to engage in unintended cyber hostile acts that might lead to armed conflict. At least in democracies, legislatures have historically had a role to play in checking executive branch military and foreign policy decisions, even if that role today is increasingly narrow. Both legislatures and executives have a responsibility and an opportunity to establish appropriate parameters for

137 Deeks (n 84) 145–46.

the use and oversight of autonomous cyber weapons. These parameters should preserve input from a range of knowledgeable actors and thus ensure that democratic accountability and other public law values, such as competence and legal compliance, are preserved in States' autonomous cyber operations.

Chapter 6

Preconditions for Applying International Law to Autonomous Cyber Capabilities

Dustin A Lewis[1]

I
INTRODUCTION

In this chapter, I seek to set out some of the preconditions arguably necessary to apply international law to employments — by a State, an international organization (IO) or a natural person[2] — of autonomous cyber capabilities. Through this thought experiment, I aim in part to help detect preconditions arguably necessary to facilitate compliance with international law or incurrence of responsibility for violations of international law that may arise in respect of such employments.

I proceed as follows. In section II, I frame some basic aspects of the inquiry. In section III, I seek to elaborate on some of the preconditions

1 I am grateful for comments from the workshop participants, the editors, and Naz K Modirzadeh.

2 I do not address the responsibility of other entities for which international legal responsibility may arise. For example, for a recent scholarly analysis concerning the (potential) responsibility of non-state parties to armed conflict, see Laura Íñigo Álvarez, *Towards a Regime of Responsibility of Armed Groups in International Law* (Intersentia 2020).

arguably necessary concerning humans *involved* in an employment of autonomous cyber capabilities by or on behalf of a State or an IO. In section IV, I aim to set out some of the preconditions arguably necessary concerning the application of international law by humans and entities *not involved* in relevant conduct attributable to a State or an IO. In section V, I outline some of the preconditions arguably necessary for the application of international law under the Rome Statute of the International Criminal Court ('ICC').[3] In section VI, I briefly conclude.

Two caveats are in order. First, the bulk of the research underlying this chapter drew primarily on English-language materials. The absence of a broader examination of legal materials, scholarship, and other resources in other languages narrows the study's scope. Second, this chapter seeks to set forth preconditions underpinning the application of international law in broad brush strokes.[4] The analysis and the identification of potential issues and concerns are, therefore, far from comprehensive. Analysis in respect of particular circumstances or fields of international law may uncover additional preconditions arguably necessary to apply international law, including as it relates to facilitating incurrence of responsibility for violations.

II
FRAMING

There is no agreed definition under international law of autonomous cyber capabilities. For this chapter, I adopt a definition rooted in the concept of autonomy elaborated in this volume by Tim McFarland.[5] Thus, the provisional description of autonomous cyber capabilities adopted here relates to bringing about desired effects through a system involving software

3 Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3 ('ICC Statute').

4 My analysis in this chapter draws on the work of a research project at the Harvard Law School Program on International Law and Armed Conflict entitled 'International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns'. That project seeks to strengthen international debate and inform policy-making on the ways that artificial intelligence and complex computer algorithms are transforming, and have the potential to reshape, war. See Harvard Law School Program on International Law and Armed Conflict, 'Project on International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns' (November 2019) <https://pilac.law.harvard.edu/international-legal-and-policy-dimensions-of-war-algorithms>.

5 See Tim McFarland, 'The Concept of Autonomy', this volume, ch 2.

subject to control inputs applied in advance that partially or entirely exclude human interaction with the system during an operation.

Due to the nature of autonomous cyber capabilities and the apparent complexities of the socio-technical arrangements through which they are configured,[6] these capabilities have been said to raise certain issues concerning the application of international law.[7] Those issues relate to an array of matters, including whether the performance and effects of particular autonomous cyber capabilities are sufficiently foreseeable *before* employment,[8] sufficiently administrable *during* employment,[9] and sufficiently assessable *after* employment.

At least two categories of actors may be involved in applying international law to an employment of autonomous cyber capabilities governed (at least in part) by international law.

The first set is made up, first and foremost, of the humans who *are involved* in relevant acts or omissions (or both) that form the employment. This first category of actors also includes the entity or entities — such as a State or an IO or some combination of State(s) and IO(s) — to which the employment is attributable, including software engineers, operators, and legal advisers engaging in conduct on behalf of the entity.

The second set of actors is made up, first and foremost, of humans *not involved* in an employment of autonomous cyber capabilities but who may nevertheless seek to apply international law in relation to the conduct that forms the employment. This second category of actors also includes entities (such as other States, other IOs, international courts, and the like) that may seek, through the humans who compose them, to apply international law in relation to the conduct.

International law sets out particular standard assumptions of responsibility for the acts and omissions — that is, the conduct — of States and IOs. It is on the basis of those assumptions that specific legal provisions exist and are applied.[10] It is in the interrelationships between the 'primary' substantive legal provisions (whatever their source or origin,

6 See generally Lucy Suchman, 'Configuration' in Celia Lury and Nina Wakeford (eds), *Inventive Methods* (Routledge, 2012). See also Tanel Tammet, 'Autonomous Cyber Defence Capabilities', this volume, ch 3.

7 See Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>.

8 See Alec Tattersall and Damian Copeland, 'Reviewing Autonomous Cyber Capabilities', this volume, ch 10.

9 That is, under the definition adopted for this chapter, the (in)ability for a human to interact with the system in the sense of exercising oversight, control, judgment, or some combination thereof in relation to the capabilities during an operation.

10 See James R Crawford, 'State Responsibility' in Rüdiger Wolfrum (ed), *Max Planck Encyclopedia of Public International Law* (Oxford University Press 2006).

including treaty law and customary international law) and the 'secondary' responsibility institutions that international law exists and is applied in relation to States and IOs. Regarding both State responsibility and IO responsibility, standard assumptions of responsibility are rooted in underlying concepts of attribution, breach, circumstances precluding wrongfulness, and consequences.[11] Those assumptions are general in character and are assumed and apply unless excluded, for example through an individual treaty or rule.[12]

An employment of autonomous cyber capabilities may give rise to individual criminal responsibility under international law, whether in addition to or separate from the responsibility of a State or an IO. Such individual criminal responsibility may arise where the conduct that forms such an employment constitutes, or otherwise sufficiently contributes to, the commission of an international crime.[13] For example, under the ICC Statute, the court has jurisdiction over the crime of genocide, crimes against humanity, war crimes, and the crime of aggression.[14] An employment of autonomous cyber capabilities may form part or all of the conduct underlying one or more of the crimes prohibited under the ICC Statute. Concerning imposition of individual criminal responsibility, it may be argued that standard assumptions of responsibility are based, at least under the ICC Statute, on certain underlying concepts.[15] Those concepts may arguably include jurisdiction,[16] ascription of responsibility,[17] material elements,[18] mental elements,[19] modes of responsibility,[20] grounds for excluding responsibility,[21] trial,[22] penalties,[23] and appeal and revision.[24] It is arguably on the basis of the

11 See Crawford (n 10); 'Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentary: Report of the Commission to the General Assembly on the Work of its Fifty-Third Session' (2001) 2(2) Yearbook of the International Law Commission, A/CN.4/SER.A/2001/Add.1 ('DARSIWA'); Draft Articles on Responsibility of International Organizations, with Commentary: Report of the Commission to the General Assembly on the Work of its Sixty-Third Session' (2011) 2(2) Yearbook of the International Law Commission, A/CN.4/SER.A/2011/Add.1 (Part 2) ('DARIO').
12 Crawford (n 10).
13 Regarding war crimes, see Abhimanyu George Jain, 'Autonomous Cyber Capabilities and Individual Criminal Responsibility for War Crimes', this volume, ch 12.
14 ICC Statute arts 5, 10–19.
15 See Dustin A Lewis, 'International Legal Regulation of the Employment of Artificial-Intelligence-related Technologies in Armed Conflict' [2020] Moscow Journal of International Law 53, 61–3.
16 See ICC Statute arts 5–19.
17 ibid arts 25–26.
18 ibid arts 6–8bis.
19 ibid art 30.
20 ibid arts 25, 28. Regarding command responsibility concerning autonomous cyber capabilities, see Russell Buchan and Nicholas Tsagourias, 'Command Responsibility and Autonomous Cyber Weapons', this volume, ch 13.
21 See ICC Statute arts 31–33.
22 ibid arts 62–76.
23 ibid art 77.
24 ibid arts 81–84.

assumptions related to those concepts that the provisions of the ICC Statute exist and are applied.

In sections III–V below, I outline some preconditions underlying elements that are arguably necessary for a satisfactory application of international law to an employment of autonomous cyber capabilities governed (at least in part) by international law. In this chapter, by satisfactory application of international law, I mean the bringing of a binding norm, principle, rule, or standard to bear on a particular employment of autonomous cyber capabilities[25] in a manner that accords with the object and purpose of the relevant provision, that facilitates observance of the provision, and that facilitates incurrence of responsibility in case of breach of the provision.

III
PRECONDITIONS CONCERNING THE APPLICATION OF INTERNATIONAL LAW TO THE CONDUCT OF A STATE OR AN IO BY HUMAN AGENTS ACTING ON BEHALF OF THAT ENTITY

In this section, I focus on employments of autonomous cyber capabilities attributable to one or more States, IOs or some combination thereof. In particular, I seek to outline some preconditions underlying elements that are arguably necessary for a satisfactory application of international law by a State or an IO to an employment of autonomous cyber capabilities.

1 HUMANS ARE LEGAL AGENTS OF STATES AND IOS

The first precondition is that humans are arguably the agents for the exercise and implementation of international law applicable to States and IOs.[26] This precondition is premised on the notion that existing

25 Derived in part from 'application, n.': OED Online, 'Definition 4.a' (Oxford University Press, September 2020).

26 See Switzerland, 'Towards a "Compliance-Based" Approach to LAWS (Lethal Autonomous

international law presupposes that the functional exercise and implementation of international law by a State or an IO in relation to the conduct of that State or IO is reserved solely to humans. In line with the formulated precondition, the exercise and implementation of international law may not be partly or fully reposed in non-human (artificial) agents.[27]

If the premise underlying the first precondition is valid, the absence of an exercise and implementation of international law by human agents of the State or the IO may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

2 HUMAN AGENTS OF THE STATE OR THE IO SUFFICIENTLY UNDERSTAND THE PERFORMANCE AND EFFECTS OF THE AUTONOMOUS CYBER CAPABILITIES

The second precondition is that one or more human agents of the State or the IO that engages in conduct that forms an employment of autonomous cyber capabilities arguably need to sufficiently understand the technical performance and effects of the employed capabilities in respect of the specific circumstances of the employment and in relation to the socio-technical system through which the capabilities are employed.[28] To instantiate this precondition, such an understanding arguably needs to encompass (among other things) comprehension of the dependencies underlying the socio-technical system, the specific circumstances and conditions of the employment, and the interactions between those dependencies, circumstances, and conditions.

Weapons Systems)' (30 March 2016), <https://www.unog.ch/80256EDD006B8954/(httpAssets)/D2D66A9C427958D6C1257F8700415473/$file/2016_LAWS+MX_CountryPaper+Switzerland.pdf> (expressing the position that '[t]he Geneva Conventions of 1949 and the Additional Protocols of 1977 were undoubtedly conceived with States and individual humans as agents for the exercise and implementation of the resulting rights and obligations in mind.'); see also US Department of Defense, *Department of Defense Law of War Manual* (December 2016) 354, [6.5.9.3] (expressing the position that law-of-war obligations apply to persons rather than to weapons, including that 'it is persons who must comply with the law of war').

27 For an exploration concerning non-human (artificial) agents, see Samuli Haataja, 'Autonomous Cyber Capabilities and Attribution in the Law of State Responsibility', this volume, ch 11, section V.

28 On certain issues related to predicting and understanding military applications of artificial intelligence, see Arthur Holland Michel, 'The Black Box, Unlocked: Predictability and Understandability in Military AI' (UN Institute for Disarmament Research 2020) <https://unidir.org/publication/black-box-unlocked>. With respect to machine-learning algorithms more broadly, see Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' [January–June 2016] Big Data & Society 1. For recent arguments concerning limits on autonomy in weapons systems in particular, see Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, 'Limits on Autonomy in Weapon Systems' (Stockholm International Peace Research Institute, June 2020) <https://www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy_0.pdf>.

Suppose the premise underlying the second precondition is valid. If that is the case, the absence of a sufficient understanding of the technical performance and effects of the employed autonomous cyber capabilities in relation to the circumstances of use and the socio-technical system through which the capabilities are employed may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

3 HUMAN AGENTS OF THE STATE OR THE IO DISCERN THE LAW APPLICABLE TO AN EMPLOYMENT

The third precondition is that one or more human agents of the State or the IO that engages in conduct that forms an employment of autonomous cyber capabilities arguably need to discern the law applicable to the State or the IO in relation to the employment.

The applicable law may vary based on the specific legal provisions applicable to the State or the IO through various sources, or origins, of international law, such as treaty law, customary international law, and general principles of international law. The applicable law may also vary depending on the specific legal situation in which the autonomous cyber capabilities are employed. For example, international humanitarian law/law of armed conflict ('IHL'/'LOAC') is applicable in relation to an employment of autonomous cyber capabilities sufficiently connected with a situation of armed conflict. As another example, international law governing the threat or use of force in international relations is applicable in relation to an employment of autonomous cyber capabilities that forms part or all of an 'armed attack' as defined in that body of law.

If the premise underlying the third precondition is valid, the absence of the discernment of the law applicable to the State or the IO in relation to the employment may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

4 HUMAN AGENTS OF THE STATE OR THE IO ASSESS THE LEGALITY OF THE ANTICIPATED EMPLOYMENT BEFORE EMPLOYMENT

The fourth precondition is that one or more human agents of the State or the IO that engages in conduct that forms an employment of

autonomous cyber capabilities assess — before the employment is undertaken — whether the anticipated employment conforms with applicable law in relation to the anticipated specific circumstances and conditions of the employment.[29] In line with this precondition, only those employments that pass this legality assessment may be initiated and only then under the circumstances and subject to the conditions necessary to pass that legality assessment.

Suppose the premise underlying the fourth precondition is valid. In that case, the absence of an assessment of whether the employment conforms with applicable law in relation to the anticipated specific circumstances and conditions of the employment may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

5 HUMAN AGENTS OF THE STATE OR IO IMPOSE LEGALLY MANDATED PARAMETERS BEFORE AND DURING EMPLOYMENT

The fifth precondition is that one or more human agents of the State or the IO that engages in conduct that forms an employment of autonomous cyber capabilities need to impose — before and during employment — limits or prohibitions (or both) as required by applicable law in respect of the employment.

Human agents of the State or the IO need to discern and configure the particular limits or prohibitions by interpreting and applying international law in relation to the employment. Factors that these human agents might need to consider could include (among many others) interactions

29 See Tattersall and Copeland (n 8); Netta Goussac, 'Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting' (*Humanitarian Law & Policy*, 18 April 2019) <https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/>; Dustin A Lewis, 'Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider' (*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/>; Argentina, 'Questionnaire on the Legal Review Mechanisms of New Weapons, Means and Methods of Warfare' (29 March 2019) UN Doc CCW/GGE.1/2019/WP.6; Australia, 'The Australian Article 36 Review Process' (30 August 2018) UN Doc CCW/GGE.2/2018/WP.6; Argentina, 'Strengthening of the Review Mechanisms of a New Weapon, Means or Methods of Warfare' (4 April 2018) UN Doc CCW/GGE.1/2018/WP.2; The Netherlands and Switzerland, 'Weapons Review Mechanisms' (7 November 2017) UN Doc CCW/GGE.1/2017/WP.5; Germany, 'Implementation of Weapons Reviews Under Article 36 Additional Protocol I' (statement delivered at the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, 11–15 April 2016) <https://www.unog.ch/80256EDD006B8954/(httpAssets)/56540402E64EC6BEC1257F9A00437856/$file/2016_LAWS+MX_ChallengestoIHL_Statements_Germany.pdf>; United States, 'Weapon Reviews' (statement delivered at the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, 13 April 2016) <https://www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2016/meeting-experts-laws/statements/13April_US.pdf>.

between the socio-technical system's dependencies and the specific circumstances and conditions of the employment.[30] If those dependencies, circumstances, or conditions — or some combination thereof — materially change after the employment commences, the human agents of the State or the IO arguably need to discern and configure the limits or prohibitions (or both) in light of those changes.

To the extent, if any, required by the law applicable in relation to a specific employment or generally, human agents of the State or the IO may need to facilitate at least partial interaction by one or more humans with the system during the employment. Such interactions may take the form (among others) of monitoring, suspension, or cancellation.[31]

If the premise underlying the fifth precondition is valid, an absence of imposition of limits or prohibitions (or both) as required by applicable law in respect of the employment may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

6 HUMAN AGENTS OF THE STATE OR THE IO ASSESS (IL)LEGALITY AFTER EMPLOYMENT

The sixth precondition is that one or more human agents of the State or the IO that engages in conduct that forms an employment of autonomous cyber capabilities arguably need to assess, after employment, whether or not the employment complied with applicable law. To instantiate this precondition, those human agents need to piece together (among other things) which humans engaged in which elements of relevant conduct, the circumstances and conditions pertaining to that conduct, and whether the anticipated and actual performance and effects of the socio-technical system underlying the employment conformed with the legally mandated parameters.

Suppose the premise underlying the sixth precondition is valid. In that case, the absence of an assessment after employment of whether

30 For broader critiques and concerns — including some informed by socio-technical perspectives — related to (over-)reliance on algorithmic systems, see, among others, Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity 2019); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Enforce Racism* (New York University Press 2018); Brent Daniel Mittelstadt and others, 'The Ethics of Algorithms: Mapping the Debate' [July–December 2016] Big Data & Society 1; Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2016).

31 See, eg, Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 ('AP I') art 57(2)(b). For an exploration of certain legal aspects concerning precautions related to autonomous capabilities in situations of armed conflict, see Eric Talbot Jensen, 'Precautions and Autonomy in the Law of Armed Conflict', this volume, ch 9.

the employment complied with applicable law may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

7 HUMAN AGENTS OF THE STATE OR THE IO ASSESS POTENTIAL RESPONSIBILITY FOR VIOLATIONS

The seventh precondition concerns suspected violations that may arise in relation to an employment of autonomous cyber capabilities by or on behalf of a State or an IO. The precondition is that one or more human agents of the State or the IO that engaged in the conduct assess whether or not the conduct constitutes a violation and, if so, evaluate whether the international legal responsibility of the State or the IO is engaged.

To make the assessment required by this precondition, human agents of the State or the IO need to discern, first, whether or not the conduct forming the employment is attributable to the State or the IO (or to some combination of one or more State(s) or IO(s) or both).[32] If attribution is established, human agents of the State or the IO need to discern whether a breach occurred. This exercise entails assessing the conduct against applicable law. Finally, if the occurrence of a breach is established, human agents of the State or the IO need to assess whether or not the circumstances preclude the wrongfulness of the breach.

If the premise underlying the seventh precondition is valid, the absence of an assessment of whether or not the conduct constitutes a violation — and, if so, the absence of an evaluation of whether the international legal responsibility of the State or the IO is engaged — may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

8 HUMAN AGENTS OF THE STATE OR THE IO FACILITATE INCURRENCE OF RESPONSIBILITY

The eighth precondition concerns situations where a breach — the wrongfulness of which is not precluded by the circumstances — is established. The precondition is that where such a breach is found, one or more human

32 For an analysis of certain legal aspects concerning attribution of autonomous cyber capabilities to a State, see Haataja (n 27)

agents of the State or the IO arguably need to facilitate incurrence of responsibility of the State or the IO with respect to the breach.

As part of the process to facilitate such incurrence of responsibility, human agents of the State or the IO may arguably need to impose relevant consequences on the State or the IO. Those consequences may relate, for example, to cessation or reparation (or both) by the State or the IO.[33]

Suppose the premise underlying the eighth precondition is valid. In that case, an absence of facilitation of incurrence of responsibility — including the imposition of relevant consequences on the State or IO — may be preclusive of an element integral to a satisfactory application of international law by the State or the IO.

IV
PRECONDITIONS CONCERNING THE APPLICATION BY NON-INVOLVED HUMANS OF INTERNATIONAL LAW TO THE CONDUCT OF A STATE OR AN IO

As in the previous section (III), in this section I also focus on employments of autonomous cyber capabilities attributable to one or more States, IOs, or some combination of both. However, in this section, I seek to outline some preconditions underlying elements that are arguably necessary for a satisfactory application of international law to a State or an IO that conducts an employment of autonomous cyber capabilities by humans and entities *not* involved in such conduct. Such non-involved people might include, for example, legal advisers from another State or another IO or judges on an international court seized with proceedings instituted by one State against another State.

33 See DARSIWA (n 11) arts 30–31; DARIO (n 11) arts 30–31.

1 HUMANS ARE LEGAL AGENTS

The first precondition is that arguably humans are the agents for the exercise and implementation of international law applicable to the State or the IO.[34] This precondition is premised on the notion that existing international law presupposes that the functional exercise and implementation of international law to a State or an IO by a human or entity not involved in relevant conduct is reserved solely to humans. In line with the formulated precondition, that exercise and implementation of international law may not be partly or wholly reposed in non-human (artificial) agents.[35]

If the premise underlying the first precondition is valid, with respect to an employment of autonomous cyber capabilities by a State or an IO, the absence of an exercise and implementation of international law applicable to the State or IO by (non-involved) humans or entities may be preclusive of an element integral to a satisfactory application of international law by (non-involved) humans (and by a related entity, if any) to the State or the IO.

2 HUMANS DISCERN THE EXISTENCE OF CONDUCT THAT FORMS AN EMPLOYMENT OF AUTONOMOUS CYBER CAPABILITIES

The second precondition is that one or more humans not involved in the conduct of the State or the IO arguably need to discern the existence of conduct forming an employment of autonomous cyber capabilities attributable to a State or an IO. To instantiate this precondition, the conduct arguably must be susceptible to being discerned by (non-involved) humans.

Suppose the premise underlying the second precondition is valid. In that case, the absence of discernment by (non-involved) humans of the existence of conduct forming an employment of autonomous cyber capabilities attributable to a State or an IO may be preclusive of an element integral to a satisfactory application of international law by (non-involved) humans (and by a related entity, if any) to the State or the IO.

34 See above n 25.
35 For an exploration concerning non-human (artificial) agents, see Haataja (n 27).

3 HUMANS ATTRIBUTE RELEVANT CONDUCT OF A STATE OR AN IO TO THE RELEVANT ENTITY (OR ENTITIES)

The third precondition is that humans not involved in the conduct of the State or the IO arguably need to attribute conduct that forms an employment of autonomous cyber capabilities by or on behalf of a State or an IO to that State or that IO. To instantiate this precondition, the conduct undertaken by or on behalf of a State or an IO arguably must be susceptible to being ascribed by (non-involved) humans to the State or the IO.

If the premise underlying the third precondition is valid, the absence of an attribution by (non-involved) humans of conduct that forms an employment of autonomous cyber capabilities undertaken by or on behalf of the State or the IO may be preclusive of an element integral to a satisfactory application of international law by (non-involved) humans (and by a related entity, if any) to the State or the IO.

4 HUMANS DISCERN THE LAW APPLICABLE TO RELEVANT CONDUCT

The fourth precondition is that one or more humans not involved in the conduct of the State or the IO arguably need to discern the law applicable to conduct that forms an employment of autonomous cyber capabilities attributable to the State or the IO. To instantiate this precondition, the legal provisions applicable to the State or the IO to which the relevant conduct is attributable arguably must be susceptible to being discerned by (non-involved) humans. For example, if an employment of autonomous cyber capabilities by a State occurs in connection with an armed conflict to which the State is a party, humans not involved in that conduct may need to discern whether the State has contracted into Additional Protocol I of 1977 and, if not, whether a possibly relevant rule reflected in that treaty is binding on the State as a matter of customary international law.[36]

Suppose the premise underlying the fourth precondition is valid. In that case, the absence of the discernment by (non-involved) humans of the law applicable to the conduct forming an employment of autonomous cyber

36 On the so-called 'Baxter paradox', see, eg, International Law Commission, 'Third Report on Identification of Customary International Law' (27 March 2015) UN Doc A/CN.4/682, 28–29; International Law Commission, 'Provisional Summary Record of the 3251st meeting of the ILC' (15 May 2015) UN Doc A/CN.4/SR.3251, 7–8.

capabilities attributable to a State or an IO may be preclusive of an element integral to a satisfactory application of international law by (non-involved) humans (and by a related entity, if any) to the State or the IO.

5 HUMANS ASSESS POTENTIAL VIOLATIONS

The fifth precondition is that humans not involved in conduct that forms an employment of autonomous cyber capabilities attributable to the State or the IO arguably need to assess possible violations by the State or the IO concerning that conduct.

To make that assessment, (non-involved) humans arguably need to discern, first, whether or not the relevant conduct is attributable to the State or the IO. To instantiate this aspect of the fifth precondition, the conduct forming the employment of autonomous cyber capabilities arguably must be susceptible to being ascribed by (non-involved) humans to the State or the IO.

If attribution to the State or the IO is established, (non-involved) humans arguably need to discern the existence or not of the occurrence of a breach. To instantiate this aspect of the fifth precondition, the conduct forming the employment of autonomous cyber capabilities by the State or the IO arguably must be susceptible to being evaluated by (non-involved) humans as to whether or not the conduct constitutes a breach.

If the existence of a breach is established, (non-involved) humans arguably need to assess whether or not the circumstances preclude the wrongfulness of the violation. To instantiate this aspect of the fifth precondition, the conduct forming the employment of autonomous cyber capabilities arguably must be susceptible to being evaluated by (non-involved) humans as to whether or not the specific circumstances preclude the wrongfulness of the breach.

If the premise underlying the fifth precondition is valid, the absence of an assessment by (non-involved) humans of possible violations committed by the State or the IO may be preclusive of an element integral to a satisfactory application of international law by (non-involved) humans (and by a related entity, if any) to the State or the IO.

6 HUMANS FACILITATE INCURRENCE OF RESPONSIBILITY

The sixth precondition is that humans not involved in conduct forming an employment of autonomous cyber capabilities attributable to the State or the IO arguably need to facilitate incurrence of responsibility for a breach the wrongfulness of which is not precluded by the circumstances. Responsibility may be incurred through relatively more formal channels (such as through the institution of legal proceedings) or less formal modalities (such as through non-public diplomatic communications between States).

As part of the process to facilitate incurrence of responsibility, (non-involved) humans arguably need to impose relevant consequences on the responsible State or IO. Those humans typically do so by acting through a legal entity to which they are attached, such as another State, another IO, or an international court. The consequences may relate to cessation and reparations (among other forms of consequences).[37]

Regarding cessation, the responsible State or IO is obliged to cease the act, if it is continuing, and to offer appropriate assurances and guarantees of non-repetition, if circumstances so require. To instantiate this aspect of the sixth precondition, the conduct forming the employment of autonomous cyber capabilities arguably must be susceptible to being evaluated by (non-involved) humans as to whether or not the conduct is continuing, and the conduct must also arguably be susceptible to being subject to an offer of appropriate assurances and guarantees of non-repetition, if circumstances so require.

Regarding reparation, the responsible State or IO is obliged to make full reparation for the injury caused by the internationally wrongful act. To instantiate this aspect of the sixth precondition, the conduct forming the employment of autonomous cyber capabilities arguably must be susceptible to a determination by (non-involved) humans of the injury caused and the making of full reparations in respect of the injury.

Suppose the premise underlying the sixth precondition is valid. In that case, the absence of facilitation by (non-involved) humans of incurrence of responsibility of the responsible State or IO for a breach the wrongfulness of which is not precluded by the circumstances may be preclusive of an element integral to a satisfactory application of international law by (non-involved) humans and by a related entity to the State or the IO.

37 See DARSIWA (n 11) arts 30–31; DARIO (n 11) arts 30–31.

V
PRECONDITIONS CONCERNING THE APPLICATION OF INTERNATIONAL LAW TO CONDUCT THAT FORMS AN INTERNATIONAL CRIME UNDER THE ICC STATUTE

In earlier sections, I focused on applying international law to employments of autonomous cyber capabilities by or on behalf of a State or an IO, whether the application is undertaken by those involved in the conduct (section III) or those not involved in it (section IV). In this section, I seek to outline some preconditions underlying elements that are arguably necessary for a satisfactory application of international law to a human who commits an international crime related to an employment of autonomous cyber capabilities. I focus on the imposition of individual criminal responsibility under the ICC Statute.

In this section, I use the phrase 'ICC-related human agents' to mean humans who exercise and implement international law in relation to an application of the ICC Statute. Such agents may include (among others) the court's prosecutors, defense counsel, the registrar, and judges.

1 HUMANS ARE LEGAL AGENTS

The first precondition is that humans are arguably the agents for the exercise and implementation of international law applicable in relation to international crimes.[38] This precondition is premised on the notion that existing international law presupposes that the functional exercise and implementation of international law to the conduct of a natural person is reserved solely to humans. In line with the notion, this exercise and implementation of international law may not be partly or wholly reposed in non-human (artificial) agents.

If the premise underlying the first precondition is valid, the absence of such an exercise and implementation of international law by an ICC-related human agent may be preclusive of an element integral to a satisfactory application of international law to the relevant natural person.

38 See above n 26.

2 HUMANS DISCERN THE EXISTENCE OF POTENTIALLY RELEVANT CONDUCT

The second precondition is that ICC-related humans agents arguably need to discern the existence of conduct that forms an employment of autonomous cyber capabilities ascribable to a natural person. For this precondition to be instantiated, such conduct arguably must be susceptible to being discerned by ICC-related human agents.

Suppose the premise underlying the second precondition is valid. In that case, the absence of such a discernment by an ICC-related human agent may be preclusive of an element integral to a satisfactory application of international law to the relevant natural person.

3 HUMANS DETERMINE WHETHER THE ICC MAY EXERCISE JURISDICTION

The third precondition is that ICC-related human agents arguably need to determine whether or not the court may exercise jurisdiction in relation to an employment of autonomous cyber capabilities ascribable to a natural person. The court may exercise jurisdiction only over natural persons.[39] Furthermore, the ICC may exercise jurisdiction only where the relevant elements of jurisdiction are satisfied. To instantiate the third precondition, conduct that forms an employment of autonomous cyber capabilities ascribable to a natural person arguably must be susceptible to being evaluated by ICC-related human agents as to whether or not the conduct is attributable to a natural person over whom the court may exercise jurisdiction.

If the premise underlying the third precondition is valid, the absence of such a determination by ICC-related human agents may be preclusive of an element integral to a satisfactory application of international law to the relevant natural person.

39 ICC Statute art 25(1).

4 HUMANS ADJUDICATE THE EXISTENCE OR NOT OF AN INTERNATIONAL CRIME

The fourth precondition is that ICC-related human agents arguably need to adjudicate whether or not an employment of autonomous cyber capabilities ascribable to a natural person subject to the jurisdiction of the court constitutes, or otherwise contributes to, an international crime over which the court has jurisdiction. For the fourth precondition to be instantiated, such conduct arguably must be susceptible to being evaluated by ICC-related human agents in pre-trial-, trial-, and appeals-related proceedings as to whether or not (among other things) the conduct satisfies the 'material' and 'mental' elements of one or more crimes and whether the conduct was undertaken through a recognized mode of responsibility.

Suppose the premise underlying the fourth precondition is valid. In that case, the absence of such adjudication by ICC-related human agents may be preclusive of an element integral to a satisfactory application of international law to the relevant natural person.

5 HUMANS FACILITATE INCURRENCE OF INDIVIDUAL CRIMINAL RESPONSIBILITY

The fifth precondition is that ICC-related human agents arguably need to facilitate incurrence of individual criminal responsibility following an adjudication that an employment of autonomous cyber capabilities ascribable to a natural person subject to the jurisdiction of the court constituted, or otherwise contributed to, an international crime over which the court lawfully exercised jurisdiction. As part of the process to facilitate incurrence of such responsibility, ICC-related human humans arguably need to facilitate the imposition of penalties on the responsible natural person.[40]

If the premise underlying the fifth precondition is valid, the absence of such facilitation of incurrence of individual criminal responsibility by ICC-related human agents may be preclusive of an element integral to a satisfactory application of international law to the relevant natural person.

40 ibid art 77.

VI
CONCLUSION

In this chapter, I have sought to span out to frame the application of international law to employments of autonomous cyber capabilities in terms of some arguably necessary preconditions to that application. Among the things at stake include the bringing of a binding rule to bear on a particular employment of autonomous cyber capabilities in a manner that accords with the object and purpose of the relevant provision, that facilitates observance of the provision, and that facilitates incurrence of responsibility in case of breach of the provision.

The arguably necessary preconditions may vary somewhat depending in part on the type of responsibility at issue, be it the responsibility of a State, an IO or a natural person. Yet at least some commonalities may be detected across these types of responsibility, including that humans are arguably the (at least primary) legal agents for the exercise and implementation of international law and that relevant conduct arguably needs to be susceptible to being discerned, attributed, understood and assessed.

The preconditions related to States, IOs and natural persons as formulated in this chapter are by and large generic. Therefore, these preconditions might be useful to consider for the application of international law in relation to employments of other complex socio-technical systems as well. Nevertheless, due to the nature of autonomous cyber capabilities, the preconditions formulated here might be particularly salient for those capabilities.

Whether — and, if so, the extent to which — international actors will commit in practice to instantiating preconditions necessary for satisfactorily applying international law to employments of autonomous cyber capabilities may depend on factors that have not been expressly addressed in this chapter but that warrant consideration.[41]

41 For an argument that algorithmic forms of warfare — which may ostensibly include certain employments of autonomous cyber capabilities — cannot be subject to law, see Gregor Noll, 'War by Algorithm: The End of Law?' in Max Liljefors, Gregor Noll and Daniel Steuer (eds), *War and Algorithm* (Rowman and Littlefield 2019).

International Legal Obligations

Chapter 7

Autonomous Cyber Capabilities and the International Law of Sovereignty and Intervention

Michael N Schmitt

I
INTRODUCTION

The issue of how international law can respond to the advent of autonomous systems and capabilities is fraught and emotive, especially in the context of warfare, with images of 'killer robots' on one side and claims that autonomy will further humanitarian ends on the other. This chapter explores the intersection of autonomous cyber capabilities and two international law primary rules — that requiring respect for the sovereignty of other States and the prohibition on coercive intervention into their internal or external affairs. Of all of the rules of international law, these are the likeliest to be violated through employment of cyber capabilities, whether autonomous or not. The issue at hand in this chapter is whether a cyber operation that involves autonomous capabilities presents unique issues with respect to the application of these two rules. Are these rules up to the task of governing autonomy in cyberspace?

II
INTERNATIONALLY WRONGFUL ACTS

To address this question, it is first necessary to understand the concept of unlawfulness. The legal term for a violation of international law is 'internationally wrongful act'. According to Article 2 of the Articles on State Responsibility, a reliable restatement of the customary law of State responsibility prepared by the International Law Commission, 'There is an internationally wrongful act of a State when conduct consisting of an action or omission: (a) Is attributable to the State under international law; and (b) Constitutes a breach of an international obligation of the State.'[1] Both criteria must be satisfied for any cyber operation to be unlawful.

As to the first, there are a number of bases for attributing a cyber operation to a State. The clearest is that an 'organ' of the State, such as the armed forces, a security service, an intelligence agency, or the State's cyber agency, conducted the autonomous cyber operation in question.[2] A cyber operation is also attributable under law to a State when an individual or non-State group, such as a hacktivist, terrorist group, or private cyber security firm, acts on 'the instructions of, or under the direction or control of, that State in carrying out' the foperation.[3]

In the absence of attribution, a cyber operation will generally not violate international law (although there are limited exceptions, such as violations of international criminal law by individuals). For instance, operations mounted by patriotic hackers or cyber criminals who are not acting at the behest of a State do not qualify as internationally wrongful acts.[4] Even beyond this key limitation, the attribution rules can prove challenging. To take one example, the type of relationship between a State and a non-State group that qualifies as 'instructions or direction or control' is somewhat ambiguous legally, quite aside from the fact that evidence of that nexus may not be iron-clad. In that regard, claims of

1 Draft Articles on Responsibility of States for Internationally Wrongful Acts, UNGA Res 56/83 (28 January 2002), Annex (Articles on State Responsibility) art 2.
2 ibid art 4.
3 ibid art 8. Other attributable cyber operations include those conducted by persons or entities exercising elements of governmental authority (art 5), organs placed at the disposal of a State by another State (art 6), or an insurrectional or other movement that becomes the new government (art 10), and operations carried out in the absence or default of the official authorities (art 9) or that are acknowledged and adopted by a State as its own (art 11).
4 However, the State from or through which the operations are being launched may have an obligation to put an end to them in certain circumstances pursuant to the 'due diligence obligation'. ('*Tallinn Manual 2.0*') and accompanying commentary.

attribution to a State often provoke debates over the requisite standard of evidential sufficiency.

The fact that a cyber operation involves autonomous capabilities can complicate factual attribution, but it does not make attribution more difficult as a matter of law. It is the nature of the relationship between the State and the individual or group conducting the operation that determines whether the attribution criterion for an internationally wrongful act has been satisfied. Taking the most straightforward example, a military cyber unit's cyber operation that employs an autonomous capability is attributable to the unit's State irrespective of the consequences of the operation, including whether the cyber unit anticipated, or even could have reasonably anticipated, those consequences. Those are instead issues that bear on the second criterion of an internationally wrongful act, breach of a legal obligation owed another State.

For the sake of analysis, it will be assumed that the use of the autonomous cyber capabilities under consideration is attributable to a State. Therefore, the remaining analysis will focus on the second criterion of an internationally wrongful act, breach of the primary rules of international law requiring respect for the sovereignty of other States and prohibiting coercive intervention.

III
AUTONOMY

Before proceeding to those issues, it is first necessary to lay the groundwork by considering the concept of autonomy. Unfortunately, discussions of autonomous systems are plagued by a cacophony of definitions. For the purposes of this article, however, the definitional framework provided by Rain Liivoja, Maarja Naagel, and Ann Väljataga works well.

> We consider autonomous operation in its simplest sense to refer to the ability of a system to perform some task without requiring real-time interaction with a human operator. Thus, the way a system performs is not decided, in each instance, by a person, but is the result of the design and programming of the system and the stimuli that it receives from its operational environment.

> [T]his broad definition of autonomy does not mean that an autonomous system is by definition one that is completely beyond human control. Rather, it means that the manner in which a human interacts with the system and exercises control over it differs from a system that is operated manually in real time.
>
> Thus, when we speak in this paper of an autonomous cyber capability, we mean a capability that involves the performance of some significant function with a significant degree of autonomy. What constitutes significant would, however, vary from capability to capability.[5]

By this approach, different capabilities have different degrees of autonomy (ranging from so-called automated to those that are highly autonomous), with the common feature being the lack of real-time human direction.[6] Thus, using common terminology, the autonomous systems referred to in this article include most 'on the loop' and 'out of the loop systems', but not those in which the human is 'in the loop'.

In the context of the law surrounding autonomous cyber capabilities, it also is useful to distinguish cyber operations that are offensive from ones that are defensive. As discussed in this chapter, the former category comprises cyber operations employing autonomous capabilities that are attributable a State, whereas the latter are operations that are a direct response to the ongoing or imminent hostile cyber operations of another State. For instance, an autonomous capability designed to disable cyber infrastructure that is being used to carry out a hostile operation falls into the defensive category, whereas the operation to which it responds is offensive in character. A borderline case in terms of categorization is an autonomous cyber capability employed in response to another State's hostile cyber operation that targets cyber infrastructure other than that

5 Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) 10–11 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>.

6 See discussion at Tim McFarland, 'The Concept of Autonomy', this volume, s IV. See also, in the military context, the definitions found in US Department of Defense, Directive 3000.09: Autonomy in Weapons Systems (21 November 2012, incorporating change 1, 8 May 2017) 13–14:

> autonomous weapon system. A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation.
>
> human-supervised autonomous weapon system. An autonomous weapon system that is designed to provide human operators with the ability to intervene and terminate engagements, including in the event of a weapon system failure, before unacceptable levels of damage occur.
>
> semi-autonomous weapon system. A weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator.

used to conduct the hostile operation. As examined herein, such systems are encompassed in the offensive category, even though their motivation is defensive.

Defensive cyber operations employing autonomy may be further divided into passive and active operations. A passive capability operates within the targeted system. Examples are most firewalls and intrusion detection/prevention systems. Active defensive measures, by contrast, operate beyond the targeted systems, the paradigmatic example being a 'hack back'. As will become apparent, both the offensive–defensive and passive–active distinctions are of relevance in assessing whether the use of an autonomous cyber capability amounts to an internationally wrongful act in violation of the rules governing sovereignty and intervention. It is to those rules, that analysis turns.

IV
SOVEREIGNTY

The existence of a rule of sovereignty in international law was questioned in a 2018 speech at Chatham House by the United Kingdom's then Attorney General, Jeremy Wright:

> Some have sought to argue for the existence of a cyber specific rule of a "violation of territorial sovereignty" in relation to interference in the computer networks of another state without its consent.
>
> Sovereignty is of course fundamental to the international rules-based system. But I am not persuaded that we can currently extrapolate from that general principle a specific rule or additional prohibition for cyber activity beyond that of a prohibited intervention. The UK Government's position is therefore that there is no such rule as a matter of current international law.[7]

By the British approach, cyber operations, whether involving autonomous capabilities or not, never violate the sovereignty of the State into

7 Jeremy Wright, 'Cyber and International Law in the 21st Century' (Address at Chatham House, 23 May 2018) <https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>.

which they are conducted. For the United Kingdom, therefore, analysis typically begins with an assessment of whether a hostile cyber operation constitutes unlawful intervention (see below), or even a use of force in violation of the UN Charter Article 2(4) and its customary analogue.

No other State has publicly taken the same position, although the US Department of Defense's General Counsel expressed a degree of sympathy with elements of the position at an address in 2020.[8] A number of States, including France,[9] the Netherlands,[10] Czech Republic,[11] Austria,[12] and Switzerland,[13] have taken the opposite position. In its 2020 Allied Joint Publication 3.20, *Allied Joint Doctrine for Cyberspace Operations*, NATO States did so as well, although the United Kingdom issued a reservation on that particular element of the doctrine.[14]

That sovereignty is a rule of international law applicable in the cyber context is the more defensible position, one well-founded in treaty law, State practice and *opinio juris*, as well as the subsidiary sources of international law, decisions of tribunals and the work of scholars.[15] Indeed, sovereignty is the rule of international law most likely to be violated by hostile cyber operations attributable to States. The aspect of autonomy changes nothing in this regard.

Sovereignty can be violated based on either territoriality or based on interference or usurpation of inherently governmental functions. For there to be a territorial violation, a cyber operation attributable to a State must cause some effect on another State's territory; it makes no difference whether that effect manifests on government or private cyber-infrastructure. More to the point, it makes no legal difference whether

8 Paul C Ney Jr 'DOD General Counsel Remarks at US Cyber Command Legal Conference' (2 March 2020) <https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/>. For a fuller discussion of the remarks, see Michael N Schmitt, 'The Defense Department's Measured Take on International Law in Cyberspace' (*Just Security*, 11 March 2020) <https://www.justsecurity.org/69119/the-defense-departments-measured-take-on-international-law-in-cyberspace/>.

9 France, Ministry of the Armies, 'International Law Applied to Cyberspace' (2019) 6–7 <https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf>.

10 Netherlands, Ministry of Foreign Affairs, 'Letter to the Parliament on the International Legal Order in Cyberspace: Annex' (5 July 2019) 1–2 <https://www.government.nl/ministries/ministry-of-foreign-affairs/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace>.

11 'Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security — Second Substantive Session' (10–14 February 2020) <http://webtv.un.org/search/3rd-meeting-open-ended-working-group-on-developments-in-the-field-of-information-and-telecommunications-in-the-context-of-international-security-second-substantive-session-10%E2%80%9314-february-2020/6131646836001/?term=%22Open%20Ended%20Working%20Group%22&lan=English&cat=Meetings%2FEvents&sort=date>.

12 ibid.

13 ibid.

14 NATO, *Allied Joint Publication 3.20: Allied Joint Doctrine for Cyberspace Operations* (January 2020) v, 20.

15 Michael N Schmitt and Liis Vihul, 'Respect for Sovereignty in Cyberspace' (2017) 95 Texas Law Review 1638.

the requisite effect is caused by a system with autonomous capabilities. It is the nature of the effect that matters.[16]

The unresolved issue is the type of effects that qualify an operation as a sovereignty violation. It would seem clear that non-*de minimis* physical damage or injury caused by the use of an autonomous cyber capability on another State's territory would do so. Below this threshold, consensus is elusive. The prevailing view appears to be that at least a cyber operation resulting in a permanent loss of functionality of the targeted cyber infrastructure, or systems that rely upon it, qualifies.[17] Similarly, an operation necessitating either replacement or physical repair of that system, as in the case of replacing components, violates sovereignty.[18]

Unfortunately, States have been reticent to set forth their legal positions as to where the threshold for violation of sovereignty lies. To date, only France has done so with any degree of granularity. In a document issued by its Ministry of the Armies, that nation took the position that 'Any cyberattack against French digital systems or any effects produced on French territory by digital means by a State organ, a person or an entity exercising elements of governmental authority or by a person or persons acting on the instructions of or under the direction or control of a State constitutes a breach of sovereignty.'[19] Although the precise parameters of France's approach remain to be determined, it is an extremely broad approach to qualifying cyber operations as violations of sovereignty, one that other States may feel uncomfortable adopting, lest it bar their own cyber operations.

Returning to the operational typology, a passive cyber defensive measure employing autonomous capability will not violate the sovereignty of other States since it takes place on the territory of the State conducting it. However, both active defensive measures and offensive cyber operations involving autonomy raise the prospect of a sovereignty violation. Whether sovereignty is violated is a question of law (the threshold for violation) and one of fact (the scale and nature of the effects caused). Autonomy does not alter the application of either of these determinations.

Sovereignty can also be violated when a cyber operation by one State interferes with, or usurps, an inherently governmental function of another State. Whether this violation can take place outside the territory

16 See discussion in *Tallinn Manual 2.0* (n 4) rule 4 and accompanying commentary.
17 ibid 20–1; France, Ministry of the Armies (n 9) 7.
18 For instance, a 2012 hostile cyber operation targeting Saudi Aramco affected 35,000 computers, necessitating replacement of hard drives: Jose Pagliery, 'The Inside Story of the Biggest Hack in History' (*CNN Business*, 5 August 2015) <https://money.cnn.com/2015/08/05/technology/aramco-hack/>.
19 France, Ministry of the Armies (n 9) 7.

of the State against which the hostile cyber operation is directed remains unsettled in international law.[20] For instance, it is unclear whether a cyber operation that leverages autonomous capabilities to target the Estonian government data stored at a data centre in Luxembourg, thereby impeding Estonian's ability to carry out its inherently governmental functions, violates Estonian sovereignty on this basis.

In most cases, hostile operations are directed against cyberinfrastructure located on the State's territory. There is a key distinction between violations based on interference with an inherently governmental act and territorial effects. In the former, the determinative factor is whether the operation interfered with or usurped an inherently governmental function. There is no requirement that a particular type of harm occur beyond that interference or usurpation. This opens the door to non-destructive and non-injurious cyber operations employing autonomous capabilities, or those that otherwise do not reach the threshold of territorial violation, amounting to a sovereignty violation.

An inherently governmental function may best be understood as a function that States alone have the authority to engage in (or authorize other entities to perform on their behalf). Classic examples include collecting taxes, conducting elections and law enforcement. For instance, take the case of an autonomous capability that searches for systems being used by a particular candidate's campaign and disrupts their use. Irrespective of whether the effects on those systems qualify the operation as a breach on the basis of territoriality, the fact that the candidate's campaign has been disrupted would amount to interference in the conduct of the election by the State concerned.

Or consider an autonomous law enforcement cyber capability that activates when it senses criminal activity. It is programmed to attempt to penetrate the cyber infrastructure being used for the criminal operation in order to disable it or to gather evidence as to the perpetrator. Reliance on autonomous capabilities has no bearing on the lawfulness of the law enforcement activity. Rather, it is the fact that the State using it is engaged in 'enforcement jurisdiction' on another State's territory without that State's consent that constitutes a violation of the latter's sovereignty. It has usurped an inherently governmental function because only the State from which the purported criminal activity emanated enjoys the competency under international law to exercise, or consent to another State's exercise of, law enforcement authority on its territory.

20 *Tallinn Manual 2.0* (n 4) 23.

As with territoriality, the use of an autonomous passive defense capability is unlikely to trigger a violation of another State's sovereignty on the basis of interference with or usurpation of another State's inherently governmental functions because States seldom have a right under international law to engage in those functions abroad (except in the commons). And as with violation of sovereignty on the basis of territoriality, both active cyber defense capabilities and autonomous offensive operations employing autonomous capabilities risk violation should they interfere with or usurp and another State's exclusive right to engage such functions on its own territory.

V
INTERVENTION

Unlike sovereignty, the existence of a rule of non-intervention in the cyber context is uncontroversial, as illustrated by the UN Group of Governmental Experts' confirmation in its 2015 report,[21] a position subsequently endorsed by the General Assembly.[22] Intervention into the internal or external affairs of another State is an internationally wrongful act in both customary international law and certain treaties, such as the Charter of the Organization of American States.[23] The parameters of a treaty violation of the rule are to be found in the text of the instruments themselves, as well as through interpretation consistent with the principles and rules set forth in the Vienna Convention on the Law of Treaties,[24] while the following analysis of intervention by autonomous cyber means is limited to the customary international law rule of non-intervention.[25]

In its *Nicaragua* judgment, the International Court of Justice observed that intervention consists of two elements, both of which must be satisfied for a violation to occur. First, the object of the cyber operation must be another State's internal or external affairs, known as the *domaine*

21 Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (22 July 2015) UN Doc A/70/174, [26], [27(b)].
22 UNGA Res 70/237 (30 December 2015).
23 Charter of the Organization of American States (signed 30 April 1948, entered into force 13 December 1951) 119 UNTS 3, arts 3(e), 19–20.
24 Vienna Convention on the Law of Treaties (adopted 23 May 1969, entered into force on 27 January 1980) 1155 UNTS 331, arts 31–33.
25 See discussion in *Tallinn Manual 2.0* (n 4) rule 66 and accompanying commentary.

réservé in international law. The Court explained, '[T]he principle forbids all States or groups of States to intervene directly or indirectly in internal or external affairs of other States. A prohibited intervention must accordingly be one bearing on matters in which each State is permitted, by the principle of State sovereignty, to decide freely. One of these is the choice of a political, economic, social and cultural system, and the formulation of foreign policy.'[26]

In other words, *domaine réservé* is an area of activity that international law leaves to States to regulate, thereby affording them the discretion to make their own choices about such activities. Although the precise contours of the *domaine réservé* are indistinct, certain activities unambiguously fall within its ambit. For example, language policy, elections, crisis management, the structure of government, and diplomatic activities clearly qualify, thereby opening the door to the possibility that a cyber operation using autonomous capabilities to affect them, as in the case of disrupting the functioning of a nation's response to a pandemic,[27] will run afoul of the non-intervention rule. By contrast, matters that are committed to international law, such as the international human rights to expression and privacy online, do not qualify. Thus, for instance, using autonomous cyber capabilities to disrupt another State's efforts to block lawful on-line expression would not qualify as a violation of the non-intervention rule; it might, however, violate the sovereignty of the State concerned.

Although there is significant overlap with the concept of inherently governmental functions in the law of sovereignty, *domaine réservé* is a broader notion.[28] Most inherently governmental functions qualify as a *domaine réservé*, but certain *domaine réservés* are not inherently governmental functions. An example is the provision of tertiary education, which in many countries is provided by the private sector and thus not inherently governmental. However, it is a *domaine réservé* in the sense that international law generally leaves States free to regulate such education. Accordingly, an offensive cyber operation involving autonomous capabilities that disrupts the functioning of tertiary education would likely not violate sovereignty unless it caused the requisite territorial

26 *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v US)* (Merits) [1986] ICJ Rep 14 ('*Nicaragua*') [205].

27 See discussion of sovereignty and intervention in the context of a pandemic in Marko Milanovic and Michael N Schmitt, 'Cyber Attacks and Cyber (Mis)information Operations during a Pandemic' (2020) 11 Journal of National Security Law & Policy 247.

28 On the relationship between sovereignty and intervention, see Harriet Moynihan, 'The Application of International Law to State Cyberattacks: Sovereignty and Non-intervention' (Chatham House Research Paper, December 2019) ch 5 <https://www.chathamhouse.org/2019/12/application-international-law-state-cyberattacks>.

effects but could constitute prohibited intervention so long as the second element of intervention, coercion, is satisfied.[29]

According to the International Court of Justice in the *Nicaragua* judgment, 'Intervention is wrongful when it uses methods of coercion in regard to such choices, which must remain free ones. The element of coercion, which defines, and indeed forms the very essence of, prohibited intervention, is particularly obvious in the case of an intervention which uses force, either in the direct form of military action, or in the indirect form of support for subversive or terrorist armed activities within another State.'[30] Applying this standard by analogy, using an autonomous offensive cyber capability to support insurgents fighting their government would amount to a clear case of intervention. The question, though, is in what other circumstances is use of an autonomous cyber capability against a *domaine réservé* prohibited by the rule?

As noted by the Netherlands Ministry of Foreign Affairs in 2019, '[t]he precise definition of coercion, and thus of unauthorised intervention, has not yet fully crystallised in international law. In essence it means compelling a State to take a course of action (whether an act or an omission) that it would not otherwise voluntarily pursue. The goal of the intervention must be to effect change in the behaviour of the target state.'[31] Restated, an act of coercion is one that deprives another State of choice by either causing that State to behave in a way it otherwise would not or to refrain from acting in a manner in which it otherwise would act. Merely influencing the other State's choice does not suffice; the choice to act or not has to effectively be taken off the table in the sense that a reasonable State in same or similar circumstances would no longer consider it to be a viable option.

To illustrate, using autonomous cyber capabilities to spread disinformation during an election is a noxious form of influence, but not necessarily coercive, for voters (the State) retain their ability to decide for whom to vote. But using autonomous cyber means to disrupt the operation of voting machinery or alter vote counts would certainly be coercive because the very ability of members of the electorate to exercise political choice has been denied.[32]

29 However, the analysis must be precise. If universities are engaged in developing responses to a pandemic at the behest of or in cooperation with the government, use of an autonomous cyber capability could be a violation of sovereignty on the basis that dealing with a pandemic is an inherently governmental function.

30 *Nicaragua* (n 26) [205].

31 Netherlands, Ministry of Foreign Affairs (n 10) 3.

32 See generally Michael N Schmitt, '"Virtual" Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law' (2018) 19 Chicago Journal of International Law 30.

An often-misunderstood dynamic of the prohibition involves the relationship between the coercion and the *domaine réservé*. The *domaine réservé* is not the physical target of the operation. Rather, it is that area of activity that the cyber operation is meant to coerce. Consider a State's covert cyber operation that employs autonomous capabilities in a ransomware attack against the sole international port facility of another State. To assess whether the operation constitutes unlawful intervention, it is necessary to determine why the former is conducting that hostile activity. If it is merely a criminal attempt to acquire funds, it is not coercive vis-a-vis any *domaine réservé*. However, if designed to force the State to, for instance, alter its trade policy by creating a situation in which there is no choice but to transship through the attacker's logistics network, the relationship between the coercive operation and a *domaine réservé*, here trade policy, exists.

As to the typology of operations, passive defensive cyber operations enabled by autonomy will not violate this rule because there is no *domaine réservé* to coerce; States do not enjoy control over a *domaine réservé* on the territory of other States. In most cases, the same is true with regard to active defensive cyber operations that employ autonomous capabilities. This is because there must be an attempt to deprive the State concerned of its exercise of choice over an area of activity that is not committed to international law. Since the State conducting the initial hostile cyber operation to which the defensive action responds is operating extraterritorially, that operation is committed to international law rules ranging from the requirement to respect the sovereignty of other States to the prohibition on the use of force. It may be that the specific operation does not violate any particular rule, but that extraterritorial cyber operations into another State's territory as such are governed by the rules of international law has long been accepted by the international community.[33] Of course, offensive cyber operations are subject to the rule of non-intervention, whether conducted using autonomous capabilities or not. Beyond attribution, the only question is whether the elements necessary for breach of that primary rule have been satisfied.

33 See, eg, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (24 June 2013) UN Doc A/68/98, [19].

VI
INTENT AND MISTAKE OF FACT

The fact that autonomous cyber capabilities operate without human involvement, and sometimes without immediate human oversight, raises issues of intent and mistake of fact. In this regard, it is necessary to dispense with one red herring at the outset. Just because a cyber capability operates autonomously does not mean that the State that employs it lacks the intent to cause the requisite consequences. Autonomous systems are not independent actors in the legal system. Rather, autonomous capabilities are programmed by humans and, more importantly, humans decide to use them. So long as that decision is attributable to a State as described above, the use of an autonomous cyber capability in no way takes the operation beyond the reach of the rules regarding sovereignty and intervention.

However, that fact the human may not entirely understand how a system with autonomous capabilities might operate, or at least be able to predict the consequences of its use, raises an interesting issue. If the individual or entity deciding to use the capability did not intend an effect that eventuated, but that effect would otherwise qualify the operation as a violation of either the sovereignty or intervention rules, have those rules nevertheless been violated?

Consider a cyber operation that uses autonomous capabilities to map a targeted system in another country. The State conducting the operation harbors no intention of causing any physical effects that would violate sovereignty, and mere cyber espionage is generally not considered to be an internationally wrongful act.[34] However, some damage unexpectedly results to the targeted system. Has the State conducting the operation breached its obligation to respect the target State's sovereignty?

Or consider a State's covert cyber operation employing autonomous means to engage in the theft of intellectual property related to the development of a COVID-19 vaccine. It does not seek to impede the process, but the breach is discovered, and affected laboratories have to shut down temporarily to assess the integrity of the affected research data. As a result, development is slowed. Did the operation violate the rule of nonintervention because 1) a nation's pandemic response falls within its *domaine réservé* and 2) the laboratories were forced to temporarily

34 *Tallinn Manual 2.0* (n 4) rule 32.

interrupt vaccine development? Of course, such situations could arise in the case of a cyber operation not employing autonomous capabilities, but they would seem more likely to surface should autonomy be relied upon.

The International Law Commission addressed the issues of intent and knowledge in its commentary to the Articles on State Responsibility.

> Whether there has been a breach of a rule may depend on the intention or knowledge of relevant State organs or agents and in that sense may be 'subjective'. For example, article II of the Convention on the Prevention and Punishment of the Crime of Genocide states that: 'In the present Convention, genocide means any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such …' In other cases, the standard for breach of an obligation may be 'objective', in the sense that the advertence or otherwise of relevant State organs or agents may be irrelevant. Whether responsibility is 'objective' or 'subjective' in this sense depends on the circumstances, including the content of the primary obligation in question. The articles lay down no general rule in that regard. The same is true of other standards, whether they involve some degree of fault, culpability, negligence or want of due diligence. Such standards vary from one context to another for reasons which essentially relate to the object and purpose of the treaty provision or other rule giving rise to the primary obligation. Nor do the articles lay down any presumption in this regard as between the different possible standards. Establishing these is a matter for the interpretation and application of the primary rules engaged in the given case.[35]

In other words, the role of intent turns on whether it is an element of the breach in question. On the one hand, if it is, as is textually the case with genocide and other rules of international criminal law, the absence of intent will preclude a cyber operation that involves autonomous cyber capabilities from amounting to either an internationally wrongful act by the State concerned or an act generating individual criminal responsibility. Importantly, though, the commentary acknowledges that intent can be a condition precedent to breach of a primary rule in which the requirement is not clear on its face. Thus, in cases of an implicit intent requirement, no breach will lie absent intent.

On the other hand, the absence of an express or implied intent requirement generally opens the door to the possibility of breach even if

35 International Law Commission, Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries (UN 2008) ('*Commentary to Articles on State Responsibility*') 34–5.

the consequences that manifested were unforeseen and unforeseeable. Accordingly, the role of intent in assessing whether a cyber operation employing autonomous capabilities violates international law depends on the presence or absence of a *mens rea* element in the individual primary rules.

However, a degree of caution is merited. As Marko Milanovic has pointed out, certain rules and regimes of international law have developed bespoke standards with respect to mistakes of fact. For instance, he notes that in international human rights law and international humanitarian law an 'honest and reasonable' mistake as to the facts can exonerate the State concerned.[36] This begs the question of whether a similar mistake of fact standard should apply in the case of other rules of international law like sovereignty and intervention.

To illustrate, consider a State A cyber countermeasure (see below on countermeasures) involving autonomous capabilities mounted against State B that unexpectedly bleeds over into State C. The result is a permanent loss of functionality of affected cyber infrastructure in State C, a violation of that State's sovereignty. If State A should have known (constructive knowledge) that bleed over would occur, it has violated State C's sovereignty even though the operation's qualification as a countermeasure precluded its wrongfulness as to State B. The belief that there would be no bleed over was not reasonable. But if the belief was reasonable, should that fact excuse the violation of State C's sovereignty?

The experts who drafted *Tallinn Manual 2.0* concluded that a mistaken but reasonable mistake of fact as to the need to use force in self-defense against another State would excuse that use of force.[37] As Milanovich notes, there is a degree of State practice supporting this position.[38] Yet the International Court of Justice seemed to come to a contrary conclusion in its *Oil Platforms* judgment.[39] And in the context of countermeasures, the International Law Commission, in its commentary to the Articles on State Responsibility, opined that,

> A State taking countermeasures acts at its peril, if its view of the question of wrongfulness turns out not to be well founded.

36 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/>.

37 *Tallinn Manual 2.0* (n 4) 347.

38 Marko Milanovic, 'Mistakes of Fact When Using Lethal Force in International Law: Part II' (*EJIL: Talk!*, 15 January 2020) <https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-ii/>.

39 *Oil Platforms (Iran v US)* [2003] ICJ Rep 161 ('*Oil Platforms*') [73].

> A State which resorts to countermeasures based on its unilateral assessment of the situation does so at its own risk and may incur responsibility for its own wrongful conduct in the event of an incorrect assessment. In this respect, there is no difference between countermeasures and other circumstances precluding wrongfulness.[40]

A majority of the experts who authored *Tallinn Manual 2.0* took the same position. In doing so, they 'emphasised the desirability of preventing a proliferation of countermeasures and the fact that countermeasures, despite being designed to resume lawful relations between the states concerned, nevertheless present a risk of escalation.'[41] The experts distinguished this position from their view with respect to a mistake of fact in the context of self-defense on the basis that States should be afforded a wide degree of discretion to act when the consequences of a failure to do so can be extremely serious, as is the case with respect to a failure to respond to an armed attack.

But that conclusion was not unanimous. Some experts contended that an honest and reasonable mistake of fact should operate to leave the countermeasure's preclusion of wrongfulness intact.[42] In their view, States must be empowered to defend themselves against hostile cyber operations, whether those operations are at the level of an armed attack entitling the victim State to act in self-defense or an internationally wrongful act below that level that opens the door to countermeasures.

As is apparent, the law surrounding the mistake of fact doctrine, beyond discreet bodies of law in which such a doctrine clearly applies, remains unsettled. This is certainly the case with respect to both sovereignty and intervention. The sounder legal position is that it does not excuse a violation of international law unless it negates intent with regard to a primary rule of international law requiring intent as a condition of violation. Otherwise, the State that was the victim of the mistake of fact would have to suffer the consequences of that mistake without the possibility of securing reparations, which are only due in the face of an internationally wrongful act.[43] By rejecting the applicability of a mistake of fact doctrine, the costs of a mistake of fact are appropriately shouldered by a State making it, not the victim of that mistake.

40 Commentary to Articles on State Responsibility (n 35) 130.
41 *Tallinn Manual 2.0* (n 4) at 116.
42 ibid.
43 Articles on State Responsibility (n 1) art 31.

Since intent is not a required element of the breach of the obligation to respect the sovereignty of another State, a cyber operation using autonomous capability that causes unintended qualifying effects would violate international law. As to the unsettled question of whether a mistake of fact doctrine might excuse a sovereignty violation, States are likely to reject its applicability for the aforementioned reason, especially as autonomous, and especially artificial intelligence, cyber capabilities become common. After all, the less control a State exercises over the conduct of an operation, the more logical it is that the State bear the risk of its mistake and the less appropriate it is that victim States should be left less than whole.

By contrast, intent is an implied requirement for the internationally wrongful act of intervention into the internal or external affairs of another State. Recall that there must be a relationship between coercion and the *domaine réservé*; the State conducting the operation has to seek to deprive the target State of choice with respect to its behaviour or policies involving a *domaine réservé*. Therefore, absent intent to do so, there would be no violation of this prohibition if an autonomous cyber capability caused unexpected harm that in fact deprived the affected State of choice.

To take a simple example, consider a case in which a State uses autonomous passive cyber defences to enhance the security of cyber systems on its territory. An insurgent group in another State has been using a social media platform operated from the former for command, control and communications ('C3') in hostilities with the government. The autonomous passive defensive measures significantly improve the security of social media, thereby contributing to the security of the insurgent group's C3. In that there was no intent to enhance the insurgent group's operational capabilities, there is no intervention.

VII
CIRCUMSTANCES PRECLUDING WRONGFULNESS

Even though certain cyber operations employing autonomous capabilities might breach either the obligation to respect the sovereignty of other States or the prohibition on intervention into the internal or external affairs of those States, international law sets forth a number of

circumstances in which international law nevertheless would not be violated. These so-called 'circumstances precluding wrongfulness' include consent, self-defense, qualification of the action as a countermeasure, force majeure, distress, and necessity.[44] The most significant in the context of autonomy are countermeasures, necessity, and self-defense.

A COUNTERMEASURES

A countermeasure is an 'act' (either an action or omission) that would be unlawful but for the fact that it is designed to put an end to another State's (the 'responsible State' in international law terms) operation that is breaching an obligation owed the former (the 'injured State' in international law terms).[45] Nothing bars application of this circumstance precluding wrongfulness to cyber operations that involve autonomous capabilities.

As an example, this basis for precluding the wrongfulness of an internationally wrongful act could allow for active defense, such as an autonomously conducted hack-back or a human launched hack-back involving autonomous capabilities. It could also take the form of an offensive operation employing autonomous capabilities against systems other than those used to conduct the unlawful cyber operation if the objective is to compel the responsible State to desist. This is because a countermeasure need not be directed at the entity conducting the unlawful cyber operation or the cyberinfrastructure from which it originated. For instance, a cyber countermeasure might leverage autonomous capabilities to target vulnerable government or private cyberinfrastructure having nothing to do with the cyber operation to which the injured State is responding. A countermeasure need not even be in-kind; a cyber operation involving autonomous capability may be used in response to a non-cyber internationally wrongful act, as in the case of providing funding or arms to an insurgent group fighting the government.[46] The key limitation on countermeasures is instead that they may only be intended to either put an end to an ongoing unlawful action or to secure reparations for one that has been completed, or both; countermeasures may not be motivated by a desire to punish or retaliate.[47]

44 Articles on State Responsibility (n 1) arts 20–25.
45 See generally, *Tallinn Manual 2.0* (n 4) rules 20–25.
46 Such actions qualify as intervention: *Nicaragua* (n 26) [242].
47 Articles on State Responsibility (n 1) art 49(1).

The prospect of employing an autonomous capability as a countermeasure raises three issues. First, countermeasures must be proportionate. Proportionality is understood in the countermeasures context as meaning 'commensurate with the injury suffered, taking into account the gravity of the internationally wrongful act and the rights in question.'[48] In practical terms, the negative effects of the countermeasure for the responsible State may not be excessive relative to the harm the injured State is suffering. If the autonomous capability causes excessive harm, the State taking the purported countermeasure will have itself violated international law. In this regard, recall that the absence of intent or a mistake of fact often will not excuse the injured State's violation even if the nature and extent of harm caused were unforeseen and unforeseeable. In most cases, a disproportionate countermeasure will violate the responsible State's sovereignty, but other violations might also lie.

Second, the Articles on State Responsibility provide that '[b]efore taking countermeasures, an injured State shall call upon the responsible State…to fulfil its obligations [to cease the operation and offer any appropriate assurances, guarantees and reparations][49] [and] notify the responsible State of any decision to take countermeasures and offer to negotiate with that State.'[50] An absolute notification requirement would not necessarily preclude the post-notice launch of a cyber countermeasure involving autonomous capabilities, but it would bar using autonomous capabilities to launch an automatic response to an incoming hostile cyber operation.

The commentary to the Articles on State Responsibility acknowledges that there may be certain situations requiring 'urgent countermeasures' to preserve an injured State's rights.[51] States that have spoken to the issue have taken a strong stance against a notice requirement in situations in which notice might diminish the countermeasure's likelihood of success, for instance by allowing the responsible State to take measures in anticipation of the action,[52] or because providing notice could reveal sensitive capabilities.[53] This does not necessarily mean that an automatic hack back relying upon autonomous capabilities or a no-notice countermeasure involving autonomy would never run afoul of the purported notice requirement. But it does open the door to no-notice countermeasures so

48 ibid art 51.
49 ibid arts 30, 31.
50 ibid art 52.
51 Commentary to Articles on State Responsibility (n 35) 135–6.
52 France, Ministry of the Armies (n 9) 8; Netherlands, Ministry of Foreign Affairs (n 10) 7; Ney (n 8).
53 Wright (n 7).

long as the State employing the autonomous capability can make a cogent argument that it was necessary to act without notice, as might be the case, for instance, with hostile operations against critical infrastructure that can only be defeated by exploiting a zero day vulnerability in the responsible State's systems.

Third, countermeasures are only available in response to internationally wrongful acts that are attributable to States.[54] Therefore, to be lawful there would have to be a relatively high degree of certainty that a particular State was behind the hostile cyber operation if autonomous means were used to determine whether to launch the countermeasure response or the countermeasure response itself involved autonomous capabilities. This is an important limitation in light of the view expressed above that a mistake of fact does not excuse an internationally wrongful act unless provided for in the body of law or primary rule in question, which is not the case with sovereignty or intervention. Indeed, recall that both the International Law Commission and a majority of the *Tallinn Manual 2.0* experts were of the view that countermeasures are taken at the injured State's risk.

B NECESSITY

A second basis upon which the wrongfulness of a cyber operation utilizing autonomous capability is precluded is in a circumstance of necessity. A cyber operation is 'necessary' when it is 'the only way for the State to safeguard an essential interest against a grave and imminent peril and (d)oes not seriously impair an essential interest of the State or States towards which the obligation exists, or of the international community as a whole.' [55]

This circumstance precluding wrongfulness is especially important, for there is no requirement that the hostile cyber operation to which the cyber operation responds be attributable to a State, or even that the initiator of the operation be known. Moreover, the hostile cyber operation to which the State responds in necessity need not be an internationally wrongful act. Most importantly, a State's cyber operation conducted on

54 Articles on State Responsibility (n 1) art 22. Note that a countermeasure directed at a non-State actor conducting hostile cyber operations might be appropriate if the State from which the operation being mounted is in breach of its due diligence obligation. See Michael N Schmitt, 'In Defense of Due Diligence' [2015] Yale Law Journal Forum 68, 79–80.

55 Articles on State Responsibility (n 1) art 25(1). See generally *Tallinn Manual 2.0* (n 4) rule 26 and accompanying commentary.

the basis of necessity is lawful even though it may breach an obligation such as sovereignty that is owed another State that bears no responsibility whatsoever for the situation, as long as doing so does not seriously affect the latter's essential interests. This makes the possibility of bleed over caused by an autonomous capability less likely to result in a violation of international law. Thus, necessity fills key gaps left by these requirements in the context of countermeasures.

As with countermeasures, there may be practical issues with respect to using autonomous capabilities in situations of necessity, both when they contribute to determining whether to launch a response (perhaps without human involvement), and as to those that form part of the cyber response. With respect to the former, the autonomous capability would have to discern if an essential interest of the State is at stake and determine whether the negative impact on that interest is grave. Part of the challenge is that neither 'essential interest' nor 'grave and imminent peril' are well-defined in international law.

In this regard, policymakers and scholars often speak in terms of hostile cyber operations against critical infrastructure as triggering necessity. However, it is not the infrastructure that must be essential, but rather the interest that an operation against the infrastructure will affect. Moreover, the notion of critical infrastructure is relative; one State's critical infrastructure may not be another's because States have differing needs. And even if it can be agreed that certain cyberinfrastructure is of a nature that an operation conducted against it will always affect an essential interest, as in the case of nuclear facilities, a cyber operation targeting that infrastructure might not gravely affect the interest. Thus, while there could be circumstances in which the employment of autonomous capabilities on the basis of necessity is lawful, the capability would have to be programmed very carefully to ensure it comports with necessity's demanding criteria.

Finally, the requirement that a cyber operation mounted on the basis of necessity not place the essential interests of other States in grave and imminent peril presents a significant obstacle if autonomous capabilities are used. Should the response cause an effect at that level, the fact that the State did not anticipate those consequences, a possibility that is likely exacerbated by autonomous capabilities, would not shield it from responsibility for violations of international law, in particular sovereignty, involving those effects.

C SELF-DEFENCE

A third circumstance precluding wrongfulness is self-defense pursuant to Article 51 of the UN Charter and customary international law.[56] That article provides, in relevant part, 'Nothing in the present Charter shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member of the United Nations, until the Security Council has taken measures necessary to maintain international peace and security.'[57] Although self-defense as a circumstance precluding wrongfulness is usually discussed in the context of the prohibition on the use of force found in UN Charter Article 2(4) and customary international law, most uses of force also violate the sovereignty of the State into which they are conducted and, as noted by the International Court of Justice in it *Nicaragua* judgment, the rule of non-intervention.[58] Thus, if a cyber operation involving autonomous capability qualifies as an act of self-defense, neither of those rules is violated.

In that this circumstance precluding wrongfulness envisions a use of force, it places very strict criteria on its applicability. Most important, self-defense is only available when the operation to which it responds is at the 'armed attack' level. That threshold is somewhat ambiguous in the non-cyber context but very much more so with respect to hostile cyber operations.[59] Cyber operations involving autonomous capabilities that result in significant injury or physical damage clearly qualify, but below that threshold there is a lack of consensus in the international community.[60]

The most robust position taken to date is that of the French Ministry of the Armies, which announced in 2019 that a 'cyberattack could be categorised as an armed attack if it caused substantial loss of life or considerable physical or economic damage. That would be the case of an operation in cyberspace that caused a failure of critical infrastructure with significant consequences or consequences liable to paralyse whole swathes of the country's activity, trigger technological or ecological disasters and claim numerous victims.'[61] Since French position has not yet

56 Articles on State Responsibility (n 1) art 21. See generally *Tallinn Manual 2.0* (n 4) rules 71–5 and accompanying commentary.
57 Charter of the United Nations, art 51.
58 *Nicaragua* (n 26) [205].
59 Michael N Schmitt, 'The Use of Cyber Force and International Law' in Marc Weller (ed), *The Oxford Handbook of the Use of Force in International Law* (Oxford University Press 2015) 1110, 1119–29.
60 Netherlands, Ministry of Foreign Affairs (n 10) 9: 'At present there is no international consensus on qualifying a cyberattack as an armed attack if it does not cause fatalities, physical damage or destruction yet nevertheless has very serious non-material consequences.'
61 France, Ministry of the Armies (n 9) at 8.

been publicly embraced by other States, most of whom have remained silent on the matter, the threshold at which self-defense will preclude the wrongfulness of a cyber operation involving autonomous capabilities remains highly uncertain.

This being so, States resorting to autonomous capabilities must be alert lest they inadvertently respond in self-defense to a cyber operation not reaching the armed attack threshold, wherever it might lie. This prospect is particularly problematic because while, as discussed, it is uncertain whether a mistake of fact excuses a mistaken use of cyber force in self-defense, there is no question that it does not excuse a mistake of the law, such as an error regarding the threshold for breach. And even though the threshold of harm necessary to trigger the right of self-defense is ambiguous, a State operating in the grey zone of normative uncertainty always risks the condemnation of other States. That autonomous capabilities might generate results that are somewhat less predictable than cyber operations not employing such capabilities only complicates matters.

Two further uncertainties in the law of self-defense further complicate cyber operations involving autonomous capabilities. First, there is a longstanding debate as to whether States are entitled to resort to self-defense in the face of hostile operations at the armed attack level that were neither mounted by another State nor, in the words of the International Court of Justice in the *Nicaragua* judgment, conducted 'by or on behalf', or with the 'substantial involvement' of, another State.[62] Although the better view is that the right of self-defense applies to armed attacks by non-State actors,[63] the International Court of Justice has on two occasions confirmed the restrictive position it took in *Nicaragua*.[64] Should that approach prevail as a matter of law, those employing an autonomous capability, or the anonymous capability itself, would have to have the capacity to distinguish operations satisfying the conditions set forth by the Court from those that do not.

Second, this uncertainty relates directly to the so-called 'unwilling-unable' debate.[65] Assuming for the sake of analysis that self-defense is available against non-State actors, consider a case in which non-State actors are acting from the territory of another State without the

62 *Nicaragua* (n 26) [195].
63 Compare, eg, Netherlands, Ministry of Foreign Affairs (n 10) 9 (applies to non-State actors) to France, Ministry of the Armies (n 9) 8 (must be conducted 'directly or indirectly' by a State).
64 *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory* (Advisory Opinion) [2004] ICJ Rep 136, [139]; *Armed Activities in the Congo (Democratic Republic of the Congo v Uganda)* [2005] ICJ Rep 168, [146]–[147].
65 *Tallinn Manual 2.0* (n 4) 347–8.

involvement of that State. May the victim State conduct cyber operations involving autonomous capabilities into the territorial State against the non-State actor without violating the territorial State's sovereignty (or violating the rule of non-intervention)?

It may not do so on the basis of countermeasures because they are unavailable in response to the operations of non-State actors, cyber or otherwise, that are not attributable to a State. Should the non-State actor's operations not affect an essential interest of the victim State in a grave and imminent manner, neither would there be any basis to conduct the operation pursuant to necessity. And if cyber operations involving autonomous capability at the use of force level are needed to address the situation, neither countermeasures nor necessity allow for the use of force.[66] This leaves only self-defense as a possible circumstance precluding the wrongfulness of the cyber response to the non-State actor attacks.

There is substantial disagreement over whether self-defense may preclude the wrongfulness of the violation of sovereignty that would occur should the operation involving autonomy be launched on that basis into a State to which the operation cannot be attributed. Some are of the view that it cannot — that sovereignty is a veil that only may be pierced when the State concerned is considered under international law to have itself directly or indirectly launched the armed attack. However, numerous States have taken the position that the right of self-defense against the actions of a non-State actor located in the territory of another State attaches when the territorial State is either 'unable or unwilling' to put an end to the hostile operations from its territory.[67] In light of this debate, States employing autonomous cyber capabilities into other States against non-State actors pursuant to the right of self-defense run the risk of having their operations characterized as breaches of sovereignty, intervention and perhaps even unlawful uses of force by some States and international law pundits.

Finally, any use of an autonomous cyber capability on the basis of self-defense must comply with the requirements of necessity and proportionality that have been recognized by the International Court of Justice and are uniformly accepted as conditions on the right of self-defense.[68]

66 The possibility is expressly ruled out in Articles on State Responsibility (n 1) art 50(1)(a).

67 The United States, for instance, has long held this position in the non-cyber context. See, eg, Barack Obama, Remarks by the President at the National Defense University (23 May 2013) <https://obamawhitehouse.archives.gov/the-press-office/2013/05/23/remarks-president-national-defense-university>.

68 *Nicaragua* (n 26) [176], [194]; *Legality of the Threat or Use of Nuclear Weapons* (Advisory Opinion) [1996] ICJ Rep 226, [41]; *Oil Platforms* (n 39) [43], [73]–[74], [76]. See also *Tallinn Manual 2.0* (n 4) rule 72 and accompanying commentary.

In the context of self-defense, necessity denotes the requirement that there be no non-forcible means of dealing with the situation effectively, while proportionality refers to the requirement that no more cyber or non-cyber force be used than that which is required to put an end to the armed attack. Defensive responses at the use of force level that employ autonomous capabilities, and the autonomous capabilities themselves, will have to be capable of making such calculations if self-defense is to operate as a circumstance precluding wrongfulness with respect to the sovereignty of the State into which it is conducted, the principle of non-intervention into the internal or external affairs of that State, and the prohibition on the use of force.

VIII
CONCLUDING THOUGHTS

It is *de rigeur* in international law circles to approach new technologies with grave concern. The rebuttable presumption seems to be that international law will fall short in adequately governing them. That was certainly the case with cyber operations. At the time the *Tallinn Manual* project was launched in 2009, claims that cyberspace was a normative Wild West were frequent, and very much in vogue. Yet, by the time of its publication in early 2017, *Tallinn Manual 2.0*'s experts, hailing from around the world, had identified 154 consensus rules and agreed upon nearly 600 pages of commentary.

This does not mean that there are no remaining challenges in the interpretation and application of the extant international law in the cyber context. Nevertheless, States are making significant progress in assessing how international law governs cyberspace, as illustrated by the work of the multiple UN Groups of Governmental Experts, the proceedings of the UN Open-Ended Working Group, and the number of statements on the subject that have been issued in the last two years.[69]

To some extent, the same dynamic is underway with respect to autonomy and international law. Initially, attention centered on lethal autonomous weapons systems, with battle lines drawn between those

69 See fuller discussion in Michael N Schmitt, 'Taming the Lawless Void: Tracking the Evolution of International Law Rules for Cyberspace' (2020) 3(3) Texas National Security Review 32.

who would outlaw the systems and those who argued that international humanitarian law suffices to govern them, primarily through the interpretive process that occurs with all new technologies of war.[70]

This book takes an important step by looking at autonomy in the context of cyber operations, and the organizers are to be congratulated for focusing critical thinking on the subject. As with many other nascent technologies, however, and at least with respect to the international law rules requiring respect for the sovereignty of other States and prohibiting intervention into their internal or external affairs, it would appear that autonomy presents few challenges; the normative architecture appears sound. While there are numerous unsettled issues surrounding application of these two primary rules to cyber operations, the fact that a cyber operation employs autonomous capability has little legal bearing on their resolution. Rather, autonomy simply sometimes make it more difficult to confidently apply the rules because it contributes uncertainty as to consequences. Yet, these are dilemmas of fact, not law, and must be understood and acknowledged as such.

70 Compare Human Rights Watch, 'Losing Humanity: The Case Against Killer Robots' (19 November 2012) <https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots> with Michael N Schmitt and Jeffrey C Thurner, '"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict' (2013) 4 Harvard National Security Journal 231.

Chapter 8

A Moment in Time: Autonomous Cyber Capabilities, Proportionality, and Precautions

Peter Margulies[1]

I
INTRODUCTION

In the fragile domain of computer network security, seconds can mean the difference between responding effectively to an incursion and sustaining devastating damage. Using those precious seconds is a job for machines, not humans. An autonomous computer system — defined as software that chooses particular actions, without specific human pre-approval — can respond quickly.[2] However, reliance on machines has its perils, including

1 I thank Gary Brown, Tim McFarland, Kurt Sanger, and participants in the NATO Cooperative Cyber Defense Centre of Excellence Workshop on Autonomous Cyber Capabilities for comments on previous drafts. This chapter is based on a longer piece. See *Peter Margulies,* 'Autonomous Weapons in the Cyber Domain: Balancing Proportionality and the Need for Speed' (2020) 96 International Law Studies 394.

2 See Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017) ('*Tallinn Manual 2.0*') 128 (noting speed of cyber exchanges); United Nations Institute for Disarmament Research, 'The Weaponization of Increasingly Autonomous Technologies:

ensuring that machines that go beyond mere defense do so in compliance with applicable international law principles such as proportionality.[3]

Proportionality is central in several contexts. First, in the *jus ad bellum*, self-defense must be tailored to the goal of stopping an adversary's attacks.[4] Second, in the *jus in bello*, the rule of proportionality means that the harm to civilians expected cannot be excessive in light of the military advantage that the planner anticipates.[5] Third, a State's countermeasure in response to a violation of sovereignty or a breach of the principle of nonintervention should center on persuading the responsible State to comply with its obligations.[6] This chapter also argues that the duty to

Autonomous Weapon Systems and Cyber Operations' (16 November 2017) <https://unidir.org/publication/weaponization-increasingly-autonomous-technologies-autonomous-weapon-systems-and-cyber> 4 (noting that as part of 2015 'Grand Cyber Challenge' competition, US Department of Defense ('DoD') Defense Advanced Research Projects Agency ('DARPA') noted that it sought '[m]achines… to find and patch [software flaws] within seconds… and find their opponents' weaknesses'). See also Robin Geiss, 'The International-Law Dimension of Autonomous Weapons Systems' (Friedrich Ebert Siftung. October 2015) <http://library.fes.de/pdf-files/id/ipa/11673.pdf> 9 (noting that US. National Security Agency (NSA) is allegedly working on software that will autonomously analyze data inputs and when necessary respond to cyber attacks from abroad); cf Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (Norton 2018) 214–16 (describing autonomous features of Stuxnet, a means allegedly designed and deployed by the United States and Israel to insert a software flaw into the industrial control systems running centrifuges that were part of the Iranian nuclear program).

3 Masahiro Kurosaki, 'Toward the Special Computer Law of Targeting: "Fully Autonomous" Weapons Systems and the Proportionality Test' in Claus Kreß and Robert Lawless (eds), *Necessity and Proportionality in International Peace and Security* (Oxford University Press 2020) 409; Ashley Deeks, Noam Lubell and Daragh Murray, 'Machine Learning, Artificial Intelligence, and the Use of Force by States' (2019) 10 Journal of National Security Law and Policy 1. See also Alan Schuller, 'At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapons Systems with International Humanitarian Law' (2017) 8 Harvard National Security Journal 379 (discussing autonomous systems and law of armed conflict); Charles P Trumbull IV, 'Autonomous Weapons: How Existing Law Can Regulate Future Weapons' (2020) 34 Emory International Law Review 533, 580–8 (discussing application to autonomous agents of law of armed conflict, including rules of proportionality and precautions in attack); John Yoo, 'Embracing the Machines: Rationalist War and New Weapons Technologies' (2017) 105 California Law Review 443, 481–8 (arguing that accuracy of autonomous systems can reduce harm to civilians in armed conflicts).

4 See *Tallinn Manual 2.0* (n 2) 349.

5 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, 1125 UNTS 3 ('AP I') art 51(5)(b).

6 See *Tallinn Manual 2.0* (n 2) 128; *Air Services Agreement of 27 March 1946 (US v France)* (1978) 18 RIAA 417, [83] ('*Air Services Agreement*'); Michael N Schmitt, '"Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law' (2014) 54 Virginia Journal of International Law 697, 715. This chapter takes no position on whether respect for sovereignty *per se* is part of the backdrop of international law or instead constitutes a primary rule. Cf Michael N Schmitt and Liis Vihul, 'Respect for Sovereignty in Cyberspace' (2017) 95 Texas Law Review 1639, 1644–9 (suggesting that prohibition on violations of sovereignty, particularly through incursions on territory of another State short of the actual use of force, constitutes primary rule of international law) with Gary P Corn and Robert Taylor, 'Sovereignty in the Age of Cyberspace' (2017) 111 American Journal of International Law 207, 209–10 (arguing that respect for sovereignty is overarching principle, rather than basis for a separate rule barring incursions on sovereign territory when those incursions are too fleeting or marginal to constitute a use of force). See also Jeremy Wright, 'Cyber and International Law in the 21st Century' (23 May 2018) <https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century> (agreeing that respect for sovereignty *per se* is not a 'specific rule' that exceeds the scope of the principle of non-intervention); cf Eric Talbot Jensen, 'The Tallinn Manual 2.0: Highlights and Insights' (2017) 48 Georgetown Journal of International Law 735, 741–2 (taking middle view in role of sovereignty *per se*, especially content of rules on territorial incursion below the use of force threshold, turns on the 'domain and practical imperatives of States and is subject to adjustment in interstate application').

take feasible precautions — express in the *jus in bello*[7] — is inherent in all proportionality requirements, including those governing the *jus ad bellum* and countermeasures.

Proportionality serves vital purposes. In the *jus in bello*, it limits harm to key interests, including the liberty, safety, and welfare of civilians and the integrity of civilian infrastructure. Proportionality in countermeasures also safeguards State interests, curbing a victim State's impingements on a responsible State's sovereignty after an incursion that may have involved limited impact. In requiring some fit between a response and an initial incursion, proportionality in the *jus ad bellum* and in countermeasures limits escalation that can needlessly expand disputes.

In the *ad bellum* and countermeasures contexts, this chapter argues that a victim State should receive a 'margin of appreciation' — a measure of deference — in crafting an answer to incursions by a responsible State.[8] An unduly strict reading of proportionality can stifle victim States' responses, creating a 'first-mover' advantage when a State uses force unlawfully[9] or breaches the principle of non-intervention.[10] Regaining the initiative for victim States is particularly pressing in the cyber realm, where an initial attack can occur with great speed, while engendering broad effects. Autonomous cyber agents can provide that necessary response capability.[11]

To promote compliance with the proportionality principle in international law, this chapter suggests that States must also take all feasible precautions to reduce harm to civilians, civilian objects, and sovereign interests. While international law makes this duty express in the *jus in bello*, this chapter argues that the duty to take due care through feasible precautions also inherently applies to both the *jus ad bellum* and

7 AP I, art 57(2)(a)(ii); Geoffrey S Corn, 'War, Law, and the Oft Overlooked Value of Process as a Precautionary Measure' (2015) 42 Pepperdine Law Review 419, 459.

8 The European Court of Human Rights has granted States a margin of appreciation in tailoring individual rights such as the right of free expression to each State's society and culture. See *Zana v Turkey* [1997] ECHR 94, [51(ii)]. I suggest in this chapter that a similar concept has informed development of the law of countermeasures, providing a victim State with a measure of flexibility — albeit flexibility within reasonable bounds — in crafting proportional countermeasures. See *Air Services Agreement* (n 6) [83] (suggesting that assessing proportionality in countermeasures necessarily involved an "approximation" of the scale of action by a victim State needed to induce a responsible State to comply with its obligations).

9 See *Tallinn Manual 2.0* (n 2) 331–5 (explaining that use of force in cyber realm entails effects that are akin to kinetic actions in severity, immediacy, directness, and invasiveness).

10 Some commentators have argued that the law of countermeasures is unduly restrictive, hamstringing victim States' responses. Gary Corn and Eric Jensen, 'The Use of Force and Cyber Countermeasures' (2018) 32 Temple International and Comparative Law Journal 127. Ensuring that a countermeasure is an effective remedy — especially in the cyber domain — may require some streamlining and modest revision of legal requirements. For example, because time may be of the essence, States need flexibility in determining whether to provide a responsible State with notice of a pending countermeasure. See *Tallinn Manual 2.0* (n 2) 120.

11 See Deeks, Lubell and Murray, (n 3) 7–8.

countermeasures.[12] That duty is both substantive and evidentiary, with the substantive duty representing *lex ferenda*, and the evidentiary component constituting *lex lata*.

As a substantive matter, both the *jus ad bellum* and the law of countermeasures are gradually moving toward an acknowledgment that due care is a component of proportionality, at least in the interdependent cyber domain. Reflecting this emerging duty, when a State engages in lawful self-defense against another State, unintended spillover effects on a third State's networks arising from that self-defense would constitute an unlawful use of force.[13] Feasible precautions that would reduce that spillover are a logical implication of the *jus ad bellum*'s prohibition on the use of force. Similarly, the 'interdependent nature' of networks makes due care a component of proportionality in countermeasures.[14]

In addition, feasible precautions are important from an evidentiary perspective, as *proof* that a State has exercised the due care that proportionality requires. If a State has made feasible efforts to reduce the consequences of its actions in cyberspace, external audiences — be they other States, tribunals, or scholars and nongovernmental organizations — will be more likely to find that any effects beyond strict proportionality are *de minimis*. To codify this natural tendency, feasible cyber precautions in the *jus ad bellum* and countermeasures should be regarded as a prerequisite for the margin of appreciation that a target State enjoys in these arenas.

Because of the need for speed, a State deploying autonomous cyber capabilities may need to plan feasible precautions *before* a specific incursion from another State. Indeed, that plan is a crucial element in the training of an autonomous agent. For example, gathering intelligence about an adversary may be a necessary component of such training. In addition, to sharpen training, a State should have reviewed any prior actions by the agent and included results of that review in subsequent inputs to the model. A State that engages in countermeasures against a responsible State may also be able to reduce needless damage *after* its countermeasure by patching damaged networks, such as those in a third State. A State should have a plan in place *before* an attack that will enable it to make such subsequent repairs.

This chapter groups feasible precautions into four categories: reconnaissance, coordination, repairs, and review. Reconnaissance entails

12 Consistent with this implication, the *Tallinn Manual* states that States considering countermeasures must 'exercise considerable care' in ensuring compliance with proportionality. See *Tallinn Manual 2.0* (n 2) 128.
13 ibid 333–4.
14 ibid 128.

efforts to map an adversary's network in advance of any incursion by that adversary, since after an incursion time may be elusive.[15] On this view, acts of cyber espionage such as the use of honey pots are not merely permitted, but *required*, at least if they are feasible. Coordination requires that a cyber agent rely on more than one algorithm, machine, or sensor; instead, often it will entail the interaction of multiple systems, including one or more that will keep watch on the primary agent.

In addition, a State must where feasible assist in repair of damage it has caused through a countermeasure, including secondary effects felt by third-party States. Where a responding State can provide a patch to address secondary effects, that patching should in the *jus in bello* reduce the net quantum of harm to civilian persons or objects ascribed to the attack in the proportionality calculus, and play a similar role in the *jus ad bellum* and countermeasures. Precautions required under this approach would include formulating a plan for such repairs and integrating that plan into the current and future performance of autonomous models. Finally, planners must regularly review the performance in the field of autonomous cyber agents. Having a process in place to learn from *past* uses of autonomous cyber capabilities is an essential precaution for *future* uses of this rapidly evolving technology.

These precautions will not ensure compliance with the principle of proportionality in all cases involving autonomous cyber agents. But they will both promote compliance and provide States that take these precautions with a limited safe harbor: a margin of appreciation for effects that would otherwise violate the duty of proportionality in the *jus ad bellum* and countermeasures. In the *jus in bello*, taking the measures described above would comply with the rule of precautions in attack.

This chapter is in three parts. Part II discusses cybersecurity and then segues into an in-depth account of autonomy, including both virtues, such as speedy response and analysis of multiple variables, and flaws, including unintelligibility, brittleness, and bias. Part III notes the principle of distinction and the rule of proportionality in the *jus in bello*, and then analyzes in greater detail the role of proportionality in the *jus ad bellum* and countermeasures. It also discusses the rule of precautions, in its express status under the *jus in bello* and its implied function as a component of proportionality in the other two bodies of law discussed here. Part IV specifically discusses the categories of precautions outlined

15 See *Tallinn Manual 2.0* (n 2) 128 (discussing mapping as prelude to countermeasures, while also suggesting that mapping will typically occur after responsible State's breach of duty that occasioned possible countermeasure).

here: reconnaissance, coordination, repair, and review. This approach will maximize autonomy's tactical strengths in the cyber arena will curbing the effects of autonomy's flaws.

II
TWO CHALLENGING TECHNOLOGICAL ARENAS: CYBER AND AUTONOMY

Both the cyber domain and government resort to autonomous systems feature new technological challenges and capabilities.[16] This section briefly outlines these issues. It stresses challenges facing autonomous systems, to highlight the importance of legal rules to govern autonomous agents.

A CYBER INCURSIONS: THE TURN TOWARD A MORE PROACTIVE RESPONSE

The world increasingly relies on computer networks and the Internet for information, communication, and even acquiring essential goods and services. Without the Internet, both daily life and everyday governance would be far more difficult. As a result of this dependence, incursions on the Internet have taken center-stage.[17]

These incursions have taken a variety of forms. Distributed denial of service ('DDoS') attacks are among the most common, using masses of computers (botnets) to deluge web sites with emails or other communications, effectively rendering those sites dysfunctional for some period of time.[18] In addition, States and non-State actors can launch malicious software (malware) that can exfiltrate data for purposes of identity theft, pilfering of intellectual property, or espionage.[19] In another type of incursion, States and others can use malware to manipulate software or destroy

16 On the challenges posed by new technologies, see Eric Talbot Jensen, 'The Future of the Law of Armed Conflict: Ostriches, Butterflies, and Nanobots' (2014) 35 Michigan Journal of International Law 253, 257–8.
17 See US Cyberspace Solarium Commission, 'Final Report' (March 2020) <https://www.solarium.gov/>.
18 See David A Wallace and Christopher W Jacobs, 'Conflict Classification and Cyber Operations: Gaps, Ambiguities, and Fault Lines' (2019) 40 University of Pennsylvania Journal of International Law 643, 652.
19 See US Cyberspace Solarium Commission (n 17) 8–9.

data stored on other networks.[20] In incursions such as Stuxnet (sometimes called Olympic Games), States or others can manipulate software to compromise industrial control systems ('ICS'), causing kinetic damage.[21] Recently, Russia has launched coordinated information operations that used thousands of computers to impersonate persons and groups on social media, spread inaccurate data, and influence democratic elections, such as the 2016 US presidential campaign.[22]

States have sought to develop timely and effective responses to these incursions. For example, the United States has recently outlined a 'defend forward' component of its 'persistent engagement' strategy.[23] That strategy heralds a more proactive approach to parrying other cyber incursions. As part of that strategy, US cyber forces temporarily deprived a Russian government unit, the Internet Research Agency, of access to the Internet during the 2018 US election.[24] That visible US response is a powerful signal that victim States cannot remain passive in the face of cyber incursions.

B AUTONOMY

This chapter defines autonomy as artificial intelligence that chooses and executives particular actions without specific human pre-approval. Those actions can be substantive decisions that affect commerce, industry, domestic governance, and both conflict and competition between States — in other words, a broad swath of 'human' endeavor.[25] Operating autonomously, models of artificial intelligence draw inferences, discern patterns, and initiate actions based on machine learning.[26] Autonomy

20 See Dan Effrony and Yuval Shany, 'A Rule Book on the Shelf? Tallinn Manual 2.0 on Cyberoperations and Subsequent State Practice' (2018) 112 American Journal of International Law 583, 620–3.

21 In the Stuxnet episode, two States — reported the United States and Israel — introduced malware into ICS that ran Iranian centrifuges used to process uranium for Iran's nuclear program. As a result, the centrifuges overheated and had to be replaced, requiring much time, effort, and expense that set back the Iranian nuclear program. See Wallace and Jacobs (n 18) 655–6.

22 See US Cyberspace Solarium Commission (n 17) 68; Michael N Schmitt, '"Virtual" Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law' (2018) 19 Chicago Journal of International Law 30; Sean Watts and Theodore Richard, 'Baseline Territorial Sovereignty and Cyberspace' (2018) 22 Lewis and Clark Law Review 771, 790 (discussing Russian efforts to use human trolls to influence Ukrainian elections); cf Effrony and Shany (n 20) 609–11 (discussing Russian hacking of U.S. political party as part of election influence operation).

23 See US Department of Defense, 'Cyber Strategy Summary' (September 2018) <https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF>; US Cyberspace Solarium Commission (n 17) 33–4.

24 See Erica Borghard, 'Operationalizing Defend Forward: How the Concept Works to Change Adversary Behavior' (*Lawfare*, 12 March 2020) <https://www.lawfareblog.com/operationalizing-defend-forward-how-concept-works-change-adversary-behavior>.

25 For the working definition of autonomy used in this edited collection, see Tim McFarland, 'The Concept of Autonomy', this volume, ch 2.

26 Pedro Domingos, *The Master Algorithm* (Basic Books 2015); Stuart J Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd edn, Prentice Hall 2010); Peter Margulies, 'Making

offers extraordinary potential benefits, but also carries substantial risks.[27]

Scholars and practitioners of computer science had identified notable flaws in autonomy's current execution. For example, autonomous agents lack contextual judgment. Their reasoning can be 'brittle': changing minor details in their inputs can spur marked changes in outputs.[28] Outputs can also be biased.[29] In addition, because of the vast number of variables that many types of autonomous agent analyze, outputs can be difficult to explain verbally.[30] Moreover, 'automation bias' prompts human beings to overestimate technology's accuracy.[31]

III
AUTONOMY, CYBER, AND PRINCIPLES OF INTERNATIONAL LAW

In the realms of cyber and autonomy, international law applies.[32] International law includes proportionality in the *jus ad bellum*, *jus in bello*, countermeasures, and human rights.[33] Cyber and autonomy may require

Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts' in Jens David Ohlin (ed), *Handbook on Remote Warfare* (Edward Elgar 2017) 415–31; Emily Berman, 'A Government of Laws and Not of Machines' (2018) 98 Boston University Law Review 1277, 1286–90; David Lehr and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 UC Davis Law Review 653.

27 Peter Margulies, 'Surveillance by Algorithm: The NSA, Computerized Intelligence Collection, and Human Rights' (2016) 68 Florida Law Review 1045, 1063–71. See also Shin-Shin Hua, 'Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control' (2019) 51 Georgetown Journal of International Law 117, 124–6 (discussing models of machine learning); see generally Russell and Norvig (n 26).

28 Katherine J Strandburg, 'Rulemaking and Inscrutable Automated Decision Tools' (2019) 119 Columbia Law Review 1851, 1877–8.

29 Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 Proceedings of Machine Learning Research 1.

30 Zachary C Lipton, 'The Mythos of Model Interpretability' (Cornell University, 6 March 2017) <https://arxiv.org/abs/1606.03490> 4.

31 See Claudia E Haupt, 'Artificial Professional Advice' (2019) 21 Yale Journal of Law and Technology 55, 71 (noting that humans reviewing an agent's work — such as a medical diagnosis based on radiological imaging — may do only a cursory job because they believe the agent is virtually always correct).

32 See *Tallinn Manual 2.0* (n 2) 127 (discussing application to cyber of international law regarding countermeasures); Harold Hongju Koh, 'International Law in Cyberspace' (2012) 54 Harvard International Law Journal 1, 8; Brian Egan, 'International Law and Stability in Cyberspace' (2017) 35 Berkeley Journal of International Law 169, 177; Paul C Ney Jr, 'DOD General Counsel Remarks at US Cyber Command Legal Conference' (2 March 2020), <https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/>; Kristen E Eichensehr, 'The Cyber-Law of Nations' (2015) 103 Georgetown Law Journal 317; Michael N Schmitt, 'Wired Warfare 3.0: Protecting the Civilian Population During Cyber Operations' (2019) 101 International Review of the Red Cross 333, 334 (noting 'broad consensus that IHL [international humanitarian law] applies to cyber operations during an armed conflict').

33 This paper leaves the important issue of proportionality and human rights for another day. See Margulies (n 27).

modest revisions in international law rules relevant to traditional kinetic or other means of action and response. Before we address in detail the need for further elaboration or revision, we should outline the relevant international law rules. This section reviews the relevant rules on proportionality in the areas covered by this chapter: the *jus ad bellum*, countermeasures, and the *jus in bello*. I also address the *jus in bello* rule of precautions in attack, and suggest that some version of that rule applies in the other two contexts — the *jus ad bellum* and countermeasures — covered in this chapter.

A DISTINCTION, LETHAL WEAPONS, AND THE CYBER DOMAIN

While the analysis of proportionality here does not directly address the *jus in bello*'s core principle of distinction, clarification of that core principle is a useful first step. The principle of distinction bars the targeting of civilians in an armed conflict.[34] Much of the controversy about autonomy in armed conflict has stemmed from concern that autonomy poses tensions with this principle.[35] Using computers to make targeting decisions with little or no real-time human ability to veto those decisions could result in substantial noncompliance.[36]

34 AP I, arts 48, 51(2).

35 Kenneth Anderson, Daniel Reisner and Matthew Waxman, 'Adapting the Law of Armed Conflict to Autonomous Weapons Systems' (2014) 90 International Law Studies 386, 401–5; Marco Sassòli, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified' (2014) 90 International Law Studies 308; Michael N Schmitt and Jeffrey S Thurnher, '"Out of the Loop": Autonomous Weapons Systems and the Law of Armed Conflict' (2013) 4 Harvard National Security Journal 231. See also *Jeremy Rabkin and John Yoo, Striking Power: How Cyber, Robots, and Space Weapons Change the Rules for War* (Encounter 2017) (praising potential of new technology for making warfighting more precise). Critics of the use of autonomous weapons in armed conflict have outlined comprehensive concerns about compliance with the *jus in bello* and have urged a ban on development of such weapons. See Christof Heyns, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions' (United Nations 2013, A/HRC/23/47) [55] (warning that autonomous agents do not exhibit 'compassion'); Peter Asaro, 'On Banning Autonomous Weapons Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making' (2012) 94(886) International Review of the Red Cross 687 (asserting that use of autonomous agents in targeting during armed conflict may diminish regard for human life). Other scholars have argued that the critics' concerns are misplaced or exaggerated. See Chris Jenks, 'False Rubicons, Moral Panic, and Conceptual Cul-De-Sacs: Critiquing and Reframing the Call to Ban Lethal Autonomous Weapons' (2016) 44 Pepperdine Law Review 1.

36 Compliance with the principle of distinction is a more or less pressing issue depending on the precise nature and purpose of the particular system at issue. See US Department of Defense, 'DoD Directive 3000.09: Autonomy in Weapons Systems' (21 November 2012) 13–14, <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf> (defining autonomous system as one that 'once activated, can select and engage targets without further intervention by a human operator'; noting that some systems 'allow human operators to override [autonomous] operation'); see also McFarland (n 25) 13–17 (discussing conceptions of autonomy). Autonomous weapons do not necessarily target humans and may be stationary and purely defensive in character. For example, the US Navy has long used fixed autonomous weapons to

For example, suppose that an autonomous agent mistakenly 'learned' through inputted data that it was permissible to attack civilians, or drew unreasonable inferences in identifying a civilian as a direct participant in hostilities subject to targeting.[37] Acting on these mistakes would pave the way for violations of international humanitarian law ('IHL'). An autonomous agent's unreasonable decision to use lethal force in an armed conflict would constitute a major challenge to IHL's traditional balance of humanity and military necessity.[38]

Analyzing autonomous agents' compliance with IHL in the cyber realm mutes but does not eliminate the concerns raised by the prospect of agents' violation of the principle of distinction in traditional kinetic conflicts.[39] Operations in the cyber domain do not entail direct targeting of persons. So concerns about the mistaken or unreasonable use of lethal force are less compelling. However, such concerns are still relevant. Cyber attacks on civilian sites, such as hospital, schools, or traffic systems, could still cause bodily harm to civilians, as well as harm to civilian objects.[40] Any comprehensive legal regime for autonomous cyber agents should address those issues.

B CYBER AND PROPORTIONALITY

Violations of the rule of proportionality by autonomous cyber agents in the *jus ad bellum*, *jus in bello*, and countermeasures contexts can cause harm to civilians or civilian objects that is excessive or simply needless in light of those cyber agents' legitimate purposes. For example, as

identify and repel enemy missiles approaching naval vessels. See International Committee of the Red Cross, *Autonomous Weapons Systems: Technical, Military, Legal and Humanitarian Aspects* (Background Paper for Meeting of Experts, 1 November 2014) 65–6, <https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014>. These fixed, defensive applications do not raise the same concerns as mobile, offensive systems about compliance with IHL. The US Navy is also developing autonomous swarming technology for offensive naval operations that the Navy could at some point use for targeting, although the rules for these systems currently require human supervision. See Sasha Radin and Jason Coats, 'Autonomous Weapons Systems and the Threshold of Non-International Armed Conflict' (2016) 30 Temple International and Comparative Law Journal 133, 135.

37 Michael N Schmitt, 'Deconstructing Direct Participation in Hostilities: The Constitutive Elements' (2010) 42 New York University Journal of International Law and Politics 697, 699; Kenneth Watkin, 'Opportunity Lost: Organized Armed Groups and the ICRC "Direct Participation in Hostilities" Interpretive Guidance' (2010) 42 New York University Journal of International Law and Politics 641, 643–4.

38 See Michael N Schmitt, 'Military Necessity and Humanity in International Humanitarian Law: Preserving the Delicate Balance' (2010) 50 Virginia Journal of International Law 795, 796.

39 Duncan B Hollis, 'Autonomous Legal Reasoning in International Humanitarian Law' (2016) 30 Temple International and Comparative Law Journal 1, 10–11.

40 See Oona A Hathaway and others, 'The Law of Cyber-Attack' (2012) 100 California Law Review 817, 848.

suggested above, in an armed conflict an autonomous cyber agent may engage in lawful targeting of a software operating system developed for use by an adversary's military, but in the process may also cause damage to different civilian systems that is foreseeable and excessive in light of the military advantage expected from the underlying attack.[41] Similarly, outside armed conflict, an autonomous cyber agent might take a countermeasure in response to another State's interference. However, the countermeasure might entail effects on the adversary State's sovereign rights that were too far-reaching to comply with proportionality. For these reasons, proportionality's impact on the use of autonomous cyber agents matters, even if the core *jus in bello* principle of distinction is not directly in play.

1 *Jus Ad Bellum Proportionality*

Proportionality in the *jus ad bellum* governs a State's use of force in self-defense against an armed attack.[42] In the cyber realm, the State that has suffered an armed attack (the 'victim State') must first apply the threshold criterion of necessity, asking whether force — as opposed to use of passive means such as firewalls or active measures such as DDoS incursions that do not rise to the level of force — is reasonably *required* to defeat the attack.[43]

Once a victim State has found that the use of force in self-defense is necessary, it assesses proportionality. Proportionality under the *jus ad bellum* has both functional and quantitative aspects. On a functional level, proportionality asks whether a reasonable person would view the 'scale, scope, duration, and intensity' of the force used in self-defense as tailored to the prevention of further attacks.[44] Quantitatively, there will often be some relation in scale, duration, and intensity between an armed attack and force used in self-defense.[45] Those planning the use of force should consider both effects on the initial attacker and collateral impacts on other States, entities, and interests.[46]

41 See *Tallinn Manual 2.0* (n 2) 128 (noting potential for disproportionate harm in countermeasures caused by 'interconnected and interdependent nature of cyber systems').
42 See *Tallinn Manual 2.0* (n 2) 340–4.
43 ibid 348–9.
44 See *Tallinn Manual 2.0* (n 2) 349 (asserting that proportionality limits responses to those 'required to end the situation that has given rise to the right to act in self-defense').
45 See Enzo Cannizaro, 'Contextualizing Proportionality: *jus ad bellum* and *jus in bello* in the Lebanese War' (2006) 88(864) International Review of the Red Cross 779, 784 (noting that a 'State acting in self-defence… [should] maintain a certain level of correspondence between the defensive conduct and the attack which prompted it').
46 *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v USA)* (Merits) [1986] ICJ Rep 14, [194] ('*Nicaragua*'); see also ibid [7], [9], [201]–[214] (Justice Schwebel dissenting).

Assessing collateral impact is crucial for autonomous cyber agents, given the interconnectedness of the internet.[47] Responses that are necessary and proportionate for a country that has engaged in an armed attack may well be unnecessary and disproportionate if those responses spill over into other countries not responsible for the attack. In the kinetic domain, it may often be relatively straightforward to restrict a kinetic response to a particular country. For example, a missile strike in self-defense by Arcadia could target military objectives in Ruritania, if the latter country had engaged in an armed attack on Arcadia. Arcadia's strike on Ruritania would generally not affect the third country of Pacifica. However, in the interconnected world of the internet, in which Pacifica individuals and entities may use servers located in Ruritania, such precision can be more difficult to achieve. The result may be serious impacts on the sovereign interests of Pacifica and other third-party States.

On the other hand, the need for speed in the cyber domain may require greater flexibility in defining both necessity and proportionality. In particular, those concepts should not rigidly require the passage of time between an initial attack by the responsible State and the victim State's response. In the cyber realm, a waiting period of hours or even minutes could mean the difference between preserving the victim State's critical infrastructure and leaving the victim State helpless. Suppose that Ruritania launches an all-out cyber attack on Arcadia's power grid. In this situation, Arcadia may lack the time for a digital forensic investigation to determine whether its passive measures, such as firewalls, have thoroughly blocked Ruritania's attack. Similarly, Arcadia may not have time to ponder whether measures below the use of force threshold will persuade Ruritania to cease its attacks.

Nevertheless, in other situations, Arcadia may have the time to assess whether passive measures or countermeasures will adequately address the threat. For example, suppose that Arcadian officials detect phishing emails sent by Ruritanian agents to employees who inspect and maintain ICS at an Arcadian power plant. Those phishing emails contain malware that Arcadia believes could disable the plant's ICS and therefore do serious physical damage to the plant's machinery. In this situation, Arcadia will have time to require the power company to conduct a sweep of its network and send out an emergency notice to its employees to apprise them of the threat and refrain from opening messages that seem suspicious.

47 See Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) 29 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>.

In this situation, therefore, an immediate use of force by Arcadia against Ruritania would be neither necessary nor proportionate.

Read against this shifting factual backdrop, the *Tallinn Manual*'s discussion of necessity and proportionality provides victim States with the flexibility they need, without giving them unbounded license. While it is true that the Manual's *ad bellum* discussion of the need to assess the efficacy of passive defenses and countermeasures[48] may imply a specific time sequence in which assessment follows attack, that is not the only possible reading of this passage. As we will see in the next section, in some situations a State should be able to calibrate its autonomous cyber agents to detect an all-out attack and respond accordingly. In these situations, the victim State should be able to flip the conventional time sequence of attack followed by a necessity and proportionality assessment, and instead rely on a prior 'beta test' of its passive defenses and active below-the-force-threshold options. In this situation, requiring that a victim State comport with the conventional time sequence might mean that the victim State would lose the ability to respond *at all* — a result that no State would agree to and that international law does not require. The *Tallinn Manual*'s discussion should not be read to mandate this anomalous outcome.[49]

In addition, as this chapter also discusses later in this Section regarding countermeasures, a victim State is entitled to a measure of deference or 'margin of appreciation' in responding to a series of 'pin-prick' attacks in the cyber realm.[50] Consider, for example, a series of phishing attacks by Ruritania on various sensitive government agencies in Arcadia. Assume that those attacks could have kinetic consequences, if the malware that Ruritania had implanted in its phishing emails had invaded Arcadian government networks. In response, Arcadia would not be limited to individual attacks that mimicked the Ruritanian incursions. Instead, Arcadia would be allowed to use force equal to a discrete increment *beyond* aggregation of Ruritania's attacks, as long as that additional increment was reasonable. Assuming digital data and/or intelligence showed that the Ruritanian attacks were related, permitting Arcadia to aggregate the impacts of Ruritania's attacks would be consistent with proportionality.[51]

48 *Tallinn Manual 2.0* (n 2) 349.
49 ibid. If the *Tallinn Manual* were to be read in this narrow way, its guidance would unduly restrict the options available to victim States under the *jus ad bellum*.
50 ibid 342, 823.
51 ibid 342; but see Yoram Dinstein, *War, Aggression, and Self-Defence* (4th edn, Cambridge University Press 2005) 230–1 (noting that at least one State has taken this aggregate approach, while other authorities believe that pin-prick attacks must be escalating in scale to allow a State to go beyond the force necessary to repel any particular attack).

Permitting an additional increment beyond the cumulative impact of Ruritanian attacks would allow Arcadia to mount a robust response and ensure that Ruritania accrued no lasting tactical or strategic advantage. In contrast, given the interdependent nature of cyber networks, confining Arcadia to a response to individual pin-prick attacks or even to a rigidly demarcated aggregate would in practice force Arcadia to stay *below* the level of aggregate impacts. Restricting Arcadia to an aggregate would have that practical effect because an attempt to achieve a precise aggregate in an interconnected online world could well overshoot the mark. Allowing a victim State a margin of appreciation beyond the attacking State's aggregate impacts would ensure that the responsible State's violations of international law did not place the victim State at a permanent disadvantage.

But even with ability to aggregate impacts and a margin of appreciation in that calculation, proportionality would still impose some limits on the victim State's cyber response. For example, suppose Ruritania has attacked an ICS in an Arcadian defense plant and actually damaged plant machinery. Further suppose that this attack appears to be a one-off, with no other attacks in progress. Since Ruritania's attack had kinetic consequences, Arcadia could respond in self-defense.

Arcadia's response could include attacks on the ICS of a Ruritania defense plant. To the extent that a quantitative test for *jus ad bellum* proportionality applies, this response would match the Ruritanian incursion. Indeed, an attack on the ICS of *multiple* Ruritanian defense plants would be within Arcadia's margin of appreciation. So would a targeted and temporary power outage or cyber takedown limited to the Ruritanian military.

However, an Arcadian response that aimed to destroy the Ruritania power grid as a whole would be disproportionate. Without a broader Ruritanian attack, an Arcadian response taking down the entire Ruritanian power grid would exceed any quantitative test for *jus ad bellum* proportionality, and also go beyond what was reasonably necessary to deter further attacks. This reading of the *jus ad bellum* proportionality principle would limit escalation and keep disputes within the cyber domain to the extent possible, curbing spillover into the kinetic realm.

2 *Proportionality and Countermeasures*

This brings us to proportionality in countermeasures. Countermeasures are responses by a victim State to another State's violations of international law.[52] Typically, countermeasures are temporary[53] — a factor that this chapter views as related to proportionality. In addition, countermeasures have often entailed notice to the responsible State, although the notice requirement is flexible enough to respond to the dictates of practicality.[54] Under current understandings of international law, countermeasures are not available against a non-State actor, although a State can target civilian networks — subject to proportionality — in the interest of persuading the responsible State to cease and desist.[55] Countermeasures are not available in collective self-defense, and must be below the level of an armed attack.[56]

In international law regarding countermeasures, proportionality takes into account both a functional aspect — the role of the countermeasure in inducing the responsible State to 'comply with its obligations', and a quantitative aspect — matching the countermeasure with the importance, scale, and duration of the initial action that prompted the countermeasure.[57] More than in the *jus ad bellum*, function and fit are independent criteria. That is, a given countermeasure may be unlawful because it exceeds the importance of the initial action — including its impact on sovereignty — as well as the initial action's scale and duration, *even though* the countermeasure was necessary to induce the responsible State to fulfill its duties.[58]

At the same time, a key arbitral decision on countermeasures recognizes that the fit of a countermeasure need not be precise down to

52 *Tallinn Manual 2.0* (n 2) 116–17; *Gabčíkovo-Nagymaros Project (Hungary v Slovakia) (Judgment)* [1997] ICJ Rep 7, [85].
53 *Tallinn Manual 2.0* (n 2) 119.
54 ibid 120.
55 ibid 112–13.
56 ibid 125–26. Many experts believe that a State cannot employ countermeasures above the threshold for the use of force. Most States place the use of force at a lower threshold than armed attack, although the United States believes the two are identical. Ibid 126. Countermeasures also may not violate fundamental human rights or *jus cogens*. Ibid 123; Rebecca Crootof, 'International Cybertorts: Expanding State Accountability in Cyberspace' (2018) 103 Cornell Law Review 565, 577–8.
57 See '*Tallinn Manual 2.0*' (n 2) 128; International Law Commission, *Report of the International Law Commission on the Work of Its Fifty-Third Session, Draft Articles on Responsibility of States for Internationally Wrongful Acts* (UN GAOR, 56th sess, Supp No 10, UN Doc A/56/10, 2001) art 51, cmt 6 ('*Draft Articles*'); *Air Services Agreement* (n 6) [83]; Schmitt, '"Below the Threshold" Cyber Operations' (n 5) 715.
58 See *Draft Articles*, art 51, cmt 6 (noting that "in every case a countermeasure must be commensurate with the injury suffered, including the importance of the issue of principle involved … partly independent of the question whether the countermeasure was necessary to achieve the result of ensuring compliance").

the last decimal point.[59] As the arbitral tribunal noted in the Air Services case, 'judging the "proportionality" of counter-measures is not an easy task and can at best be accomplished by approximation.'[60] In practice, the willingness to engage in approximation means that the victim State receives a measure of deference — in international law, what is often called a 'margin of appreciation'[61] — in crafting a countermeasure.[62] Even with an appropriate margin of appreciation, a suitably 'commensurate' countermeasure should not interfere with an interest that is markedly more important than the interest that the initial action of the responsible State impaired.[63]

C THE DUTY TO TAKE FEASIBLE PRECAUTIONS: AN EXPRESS OR IMPLICIT DUTY, DEPENDING ON THE RELEVANT LEGAL FRAMEWORK

This chapter argues that whenever proportionality, in each of its guises, is applicable, the duty to take feasible precautions also applies, either expressly or implicitly. The rule of precautions is express in the *jus in bello*,[64] but also holds for the *jus ad bellum* and the law of countermea-

59 *Air Services Agreement* (n 6) [83]; *Draft Articles*, art 51, cmt 3.

60 ibid.

61 *Zana v Turkey* [1997] ECHR 94, [51(ii)] (despite protection for free speech, upholding criminal conviction of an official who used the phrase "national liberation movement" to describe a Kurdish group that Turkey and other States had designated as a terrorist organization); Robert D. Sloane, 'Human Rights for Hedgehogs?: Global Value Pluralism, International Law, and Some Reservations of the Fox' (2010) 90 Boston University Law Review 975, 983.

62 Michael A Newton and Larry May, *Proportionality in International Law* (Oxford University Press 2014) 183 (asserting that proportionality in countermeasures involves a 'rough contextual approximation' of the judgment of 'policymakers acting in light of the information and assessments reasonably available to them to inform good-faith decision-making'); see also ibid 186 (describing proportionality in countermeasures as 'prohibition against excesses rather than a requirement for equivalence'). That space between precise equivalence and prohibited excess is the margin of deference under proportionality. The Rome Statute does something comparable, in classifying as a war crime an attack executed with the 'knowledge' that harm to civilians will be 'clearly excessive' compared with the military advantage anticipated. See Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 2, art 8(2)(b)(iv). Additional Protocol I appears to impose a more rigorous standard, since it does not expressly include the adverb 'clearly' as a modification of 'excessive' harm to civilians. See AP I, art 51(5)(b). Deference inheres in the space between AP I's standard of 'excessive' harm, which could prompt undue second-guessing of commanders operating in the fog of war, and the Rome Statute's 'clearly excessive' standard, which requires the tribunal to find unequivocally that the harm to civilians is excessive before imposing liability. Critics of the Rome Statute's war-crime definition have argued that it undermines the rule of proportionality. See Adil Ahmed Haque, 'Protecting and Respecting Civilians: Correcting the Substantive and Structural Defects of the Rome Statute' (2011) 14 New Criminal Law Review 519, 525. Addressing the appropriateness of the Rome Statute's definition is beyond the scope of this chapter. However, the Rome Statute's drafting speaks to the perceived need for a space in which commanders and planners can operate without fear of second-guessing. This chapter suggests that a similar space is needed for countermeasures.

63 See Draft Articles, art 51, cmt 6.

64 AP I, art 57(2)(a)(ii); Corn (n 7) 459; Geoffrey Corn and James A Schoettler Jr, 'Targeting and Civilian Risk Mitigation: The Essential Role of Precautionary Measures' (2015) 223 Military Law

sures. As we shall see, the duty to take feasible precautions may either stand on its own, as an independent substantive duty layered on top of proportionality, or may be evidentiary in nature, demonstrating a State's *compliance* with the rule of proportionality. Both the substantive and evidentiary conceptions are important in the cyber domain, because the need for speed often makes prompt action necessary but also requires feasible measures to mitigate the harm that speed could cause.

Under the rule of precautions in IHL, a State must take all 'feasible' steps to reduce civilian harm.[65] A feasible step is one that is practicable, given resource constraints, technological limits, and tactical concerns such as the importance of preserving certain means or instrumentalities of warfare (including weapons) for future engagements and the disadvantage of disclosing certain advancements to adversaries or the world at large.[66] A feasible step is not one that is merely *possible*; requiring a State to implement all possible steps would hamstring commanders, undermining the crucial value of military necessity.[67] However, a definition of feasibility that imposed no duties on States would drain all meaning from the rule of precautions.

At the intersection of technology and the rule of precautions in attack, over time resource constraints and tactical concerns recede. As time progresses, mass production of any technology becomes more widespread and hence less expensive. Moreover, over time, knowledge of a once-rare or closely held technology proliferates, as the capacity to construct and deploy nuclear weapons increased from the time that the United States used nuclear weapons at the close of World War II. Decreased expense and increased proliferation ease resource constraints and tactical concerns, making it more feasible to deploy a formerly new technology.

The rise in technology that is evident in cyber and autonomy also makes precautions relevant in areas where they have not traditionally been salient, including countermeasures and the *jus ad bellum*. Technology highlights the need for speed — the importance of responding quickly, to avoid greater damage or disadvantage and increase the probability that a given measure by a victim State will effectively repel an incursion and persuade a responsible State to cease its offending conduct.[68]

Review 785, 837; Jean-Francois Queguiner, 'Precautions under the Law Governing the Conduct of Hostilities' (2006) 88 International Review of the Red Cross, 793, 797.

65 AP I, art 57(2)(a)(ii).

66 See David A Wallace and Shane R Reeves, 'Protecting Critical Infrastructure in Cyber Warfare: Is It Time for States to Reassert Themselves?' (2020) 53 UC Davis Law Review 1607, 1635.

67 cf Schmitt, 'Military Necessity and Humanity in International Humanitarian Law' (n 38).

68 See Dan Saxon, 'A Human Touch: Autonomous Weapons, Directive 3000.09, and the "Appropriate Levels of Human Judgment Over the Use of Force"' (2014) 15(2) Georgetown Journal of

A victim State that is slow to respond encourages other States to violate international law, either with an armed attack that violates the *jus ad bellum* or with an action that violates the principle of sovereignty or constitutes an unlawful interference under the use of force threshold. However, the importance of speed in a response may also produce greater adverse impacts for the responsible State and for third party States.

Here, the rule of precautions has a substantive role to play, not only in the *jus in bello*, but also in the *jus ad bellum* and countermeasures. Suppose a State can feasibly deploy technology to craft a timely, effective countermeasure that is also more precise than the response otherwise available. Given this assumption, this chapter argues that the rule of precautions applies. Therefore, States have a duty to deploy that more tailored technology.

Both the US Government and a spectrum of international law scholars have indicated support for a rule of precautions that would apply regarding the use of force, the conduct of hostilities, and countermeasures below the use of force threshold. For example, the US Department of Defense has indicated that even below the use of force threshold, a cyber operation 'should not be conducted in a way that unnecessarily causes inconvenience to civilians or neutral persons'.[69] One distinguished commentator has criticized this statement by the US Department of Defense as lacking adequate support or as merely stating a US policy preference rather than articulating a binding legal requirement.[70] However, the US Department of Defense's unqualified statement of a duty to avoid needless inconvenience to civilians through cyber operations is consistent with both the substantive conception of precautions outlined here.[71]

The *Tallinn Manual*'s International Group of Experts ('IGE') seems to endorse such a role for precautions in countermeasures, by citing the need to employ 'considerable care' in crafting a proportionate countermeasure.[72] Indeed, the IGE suggests that prior to initiating countermeasures,

International Affairs 100, 103–4.

69 US Department of Defense, *Law of War Manual* (31 May 2016) [16.5.2] <https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf?ver=2016-12-13-172036-190> ('US Law of War Manual').

70 See Gary D Brown, 'Commentary on the Law of Cyber Operations and the DoD Law of War Manual' in Michael A Newton (ed), *The United States Department of Defense Law of War Manual: Commentary and Critique* (Cambridge University Press 2018) 337, 346 (stating that military lawyers who rely on the DoD Law of War Manual 'would be better served if the Manual made clear that.. [avoiding unnecessary inconvenience to civilians or neutrals] is a US policy rather than the law,' at most warranting placement in US rules of engagement); Schmitt, 'Wired Warfare 3.0' (n 32) 82 and 349 (describing US Department of Defense Law of War Manual Statement as addressing policy).

71 The analysis in this chapter supplies additional analytical support for the US Department of Defense position.

72 *Tallinn Manual 2.0* (n 2) 128.

a victim State must conduct a 'full assessment' that includes 'mapping the targeted system' and 'reviewing relevant intelligence'.[73] As this chapter argued earlier in this Part and discusses further in the next Part, taking such precautions does not necessarily lock a victim State into a rigid time sequence. A State can engage in such precautions *before* an attack or other action by a responsible State. Indeed, a prudent State would continually acquire cyber, signals, and human intelligence about its adversaries. The key point is that States have a duty to take such measures where feasible to temper the State's response or that response's effects. The emphasis on such steps suggests that for international law, countermeasures include a precautionary element.[74]

The Tallinn Manual's editor, Professor Michael Schmitt, has also recently outlined a comparable view of the importance of precautions. Discussing contexts at or below the use of force threshold, Schmitt has urged that as a matter of policy, States not engage in cyber incursions in which the 'expected concrete negative effects on… the civilian population are excessive relative to the concrete benefit… anticipated' by the incursion.[75] Although this advice adopts the language of proportionality, it also suggests a role for precautions.

To discern the role of precautions, suppose a State can reap a particular benefit with a cyber countermeasure, at the cost of inconvenience to civilians at level X. Now supposed that the State can feasibly achieve the same benefit with a technological precaution that would reduce negative effects on civilians to 1/2 X. If the State decides to proceed *without* employing the feasible technological precaution — even though using the precaution would reap the same benefit — it is reasonable to view the difference between X and 1/2 X as 'excessive'. Professor Schmitt's description of the results of his balancing test supports this reading. For example, Professor Schmitt has noted that as a matter of policy a State should reject a cyber action that would yield significant civilian inconvenience when the expected benefit was 'trifling'.[76] Such a State decision, according to Professor Schmitt, would seem petty and mean-spirited.[77] Indeed, such a decision would not serve the criterion of military necessity that interacts with the principle of humanity to

73 ibid.
74 Perhaps the IGE suggestion here largely pertains to the evidentiary conception — proving that the State engaging in countermeasures relies on best practices to facilitate compliance with the rule of proportionality. However, one can also read the *Tallinn Manual*'s analysis as recognizing that a substantive view of precautions is inherent in the requirement that countermeasures be proportionate.
75 Schmitt, 'Wired Warfare 3.0' (n 32) 347.
76 ibid 349.
77 ibid (noting that such an incursion would 'smack of mere maliciousness').

form IHL's crucial balance.[78] At least when the cost of employing a feasible precaution is de minimis because of economies of scale, a failure to employ that precaution would similarly fail Professor Schmitt's test.

Moreover, even if precautions do not have the freestanding substantive significance in countermeasures that they possess in IHL — imposing duties *beyond* proportionality when added safety steps are 'feasible' — precautions also have an *evidentiary* significance. That is, a State that takes precautions prior to engaging in countermeasures can cite those precautions as evidence that its response is proportionate. For example, suppose a victim State's countermeasure includes collateral damage to the responsible State's systems or to neutral States that is more substantial in scale than the impact of the responsible State's initial action. As noted earlier, a State should receive a margin of appreciation that would cover modest increments beyond the initial action's effects. Suppose, however, that a margin of appreciation is only available when the victim State has demonstrated good faith or even reasonable care. In that event, the victim State's care in mapping the responsible State's system can constitute evidence that the victim State's actions causing the additional damage were not intentional, knowing, or even negligent. In this sense, the good faith and due care that a victim State shows through the taking of precautions is probative evidence of compliance with international law.

D SUMMARY

This discussion has analyzed proportionality in the cyber domain in the contexts of the *jus ad bellum* and countermeasures. As noted in this section, proportionality is both functional and substantive. Analyzing proportionality includes assessing whether the proposed countermeasure will, 1) elicit compliance with international law by the responsible State, and, 2) correspond with the importance, scale, and duration of the initial incursion. Under the view expressed here and supported by the *Air Services Agreement* arbitral award, a victim State has a margin of appreciation on criterion #2. Without that flexibility, a victim State facing rigid legal norms in an uncertain operational environment may choose a *more modest response* than the initial incursion. That restricted response would not effectively signal to the responsible State that the latter should cease its

78 ibid.

violation of international law. In addition, this Part has argued that in addition to being an independent requirement under IHL, the need to take feasible precautions is inherent in both the *jus ad bellum* and countermeasures. In each case, precautions have a substantive dimension, imposing additional duties even when a State has satisfied proportionality, and evidentiary significance, demonstrating that the State has procedures in place that will promote its compliance with the proportionality rule.

IV
PRECAUTIONS IN THE USE OF AUTONOMOUS CYBER AGENTS

Now that we have discussed the test for proportionality and made the case for an inherent rule of feasible precautions in the *jus ad bellum* and countermeasures joining the express rule in the *jus in bello*, it is time to focus more specifically on the criteria guiding the rule of precautions. The use of autonomous agents in the cyber domain poses special challenges, because of autonomous agents' brittleness, bias, and unintelligibility, as well as humans' tendency toward automation bias.[79] *Lex lata* has not yet caught up with the demands in this emerging arena. So the following discussion ventures into the venue of *lex ferenda*, although the discussion proceeds on the assumption that norms will move in this direction. By way of criteria for precautions, the chapter suggests four pillars: reconnaissance, coordination, repairs, and review. I address each in turn.

A RECONNAISSANCE AND THE IMPERATIVE OF ESPIONAGE

While espionage and reconnaissance are mainstays of State behavior in peace and war, this chapter goes further: intelligence collection, including espionage, is not merely permitted but *required* in the use of autonomous

79 US Department of Defense has recognized the importance of these issues. See US Defense Innovation Board, 'AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense' (31 October 2019) 31–3, <https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF> (discussing need to combat bias); ibid at 33–8 (discussing need to understand data inputs and explain and review outputs for autonomous agents).

cyber agents. That requirement extends not merely to the *jus in bello*, which mandates consideration of 'reasonably available' information in targeting decisions,[80] but also to the *jus ad bellum* and countermeasures. In the cyber realm, intelligence collection — which I refer to as reconnaissance here — will often be virtual.[81] But on occasion human aid to such efforts is necessary for their success — as in the case of human insertion of a thumb drive to introduce a worm for exfiltration of data[82] — and is feasible. In such situations, the approach taken here would require such human aid.

Virtual reconnaissance, supplemented as needed by human and signals methods of collection, is necessary to ensure that autonomous cyber agents comply with international law. Without the capacity to map an adversary's network and associated systems,[83] an autonomous cyber agent will be 'flying blind', without the ability to either accurately target adversaries or avoid excessive collateral damage. As noted in the previous section, waiting until after an attack or other action has occurred will often hinder an effective response in each of the contexts examined here, including the *jus ad bellum* and countermeasures as well as the *jus in bello*. Because of the need for speed, collecting cyber intelligence on potential adversaries before an attack will often be the only way to ensure that a response is both effective and tailored to avoid needless harm. Without that precaution, the brittleness and bias of autonomous agents will produce errors that both reduce the reasonably anticipated benefits of the attack or countermeasure and increase the harms that a reasonable decisionmaker would expect.

Under the substantive conception of precautions outlined in the previous section, these concerns about foreseeably reduced benefits and increased harms dictate that when prior collection through reconnaissance — including espionage — is feasible, it is required. When a victim State used reconnaissance as well as the other steps suggested below — such as coordination, review, and repair — under an evidentiary view of precautions such measures would be presumptive evidence of proportionality, entitling a victim State to a margin of appreciation.

80 See *Tallinn Manual 2.0* (n 2) 424, citing UK Ministry of Defence, *The Joint Service Manual of the Law of Armed Conflict* (2004) [5.3.4].

81 ibid 168. Militaries often refer to the gathering of information as encompassing intelligence, surveillance, and reconnaissance (ISR). See *Michael N Schmitt and Sean Watts,* 'The Decline of International Humanitarian Law *Opinio Juris* and the Law of Cyber Warfare' (2015) 50 Texas International Law Journal 180, 210–11. Purely for ease of reference, this chapter uses the term 'reconnaissance' to connote the full range of intelligence collection, including espionage.

82 *Tallinn Manual 2.0* (n 2) 171.

83 ibid 128.

On the other hand, suppose that reconnaissance is *not* feasible, and that in its absence a victim State cannot be reasonably certain that an autonomous cyber agent will be sufficiently precise to avoid excessive harm. Under the evidentiary view of precautions, use of the agent despite this concern would provide a basis to infer that the victim State had violated the rule of proportionality.

An example will be helpful. Suppose that Arcadia implants malware in various government networks of Pacifica. The malware has autonomous capabilities: it has been trained both to observe Pacifica's networks and react to particular inputs from those networks. Suppose further that Arcadia's autonomous malware receives inputs indicating that Pacifica has just commenced an attack on Arcadia's networks. Based on inputs about the operation of Pacifica's networks that Arcadia's malware has already gathered, Arcadia's autonomous cyber agent will be able to launch corresponding attacks on Pacifica's networks, echoing the scale, scope, duration, and importance of the attacks on Arcadia. Without the autonomous malware already in place, Arcadia would have had to 'start from scratch' in both attributing and responding to the attack. Absent the autonomous cyber agent that Arcadia had already implanted in Pacifica's networks, Arcadia might have mistakenly attributed the attacks to another rival, Ruritania. By virtue of the malware it had previously implanted, Arcadia has the capacity to both correctly attribute the attacks to Pacifica and respond appropriately. If placement of the malware in Pacifica's networks is feasible, this would be a necessary precaution on Arcadia's part.

B COORDINATION OF AUTONOMOUS METHODS

The brittleness and bias of autonomous agents, in addition to requiring increased reconnaissance, also mandates expanded coordination. By coordination, this chapter refers to the use of different autonomous methods simultaneously or in close succession to refine outputs.[84] The interaction of different modes acts as a check on agents' errors. A model with a particular strength or training in specific data can blunt the impact of arbitrary or biased outputs from other models with different strengths

84 See Scharre (n 2) 20–1; Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking 2019) (discussing importance of checks on outputs of any single autonomous learner).

and training. Working together, coordinated models can perform more functions without gaps or mistakes.[85]

In cyber, coordination amounts to an autonomous 'red team'. Just as a 'red team' makes human decisions better by posing objections and presenting alternatives, the autonomous equivalent lowers the risk of false positives while ensuring that attribution is precise.[86] The autonomous use of coordination aids greatly in both detecting anomalies and misuse in networks that may signal a cyber intrusion.

The coordination factor described here is not prescriptive regarding a *particular* autonomous methodology — the ensemble learning approach described above is merely an illustration. Coordination's core is an autonomous capability to conduct different inquiries of inputs simultaneously or in tight succession, to test preliminary hypotheses rapidly and weed out the effects of brittleness and bias. A State that has deployed an autonomous cyber agent with this coordination capability has checked another box in the precautionary matrix.

Coordination like this is nothing new in a State or commander's lexicon. Commanders regularly use a range of inputs from intelligence, surveillance, and reconnaissance ('ISR') and strive to weigh disparate inputs in a balanced fashion, to reduce the chance of reliance on a single flawed source. Redundancy is also a common feature of automotive and aircraft software, weapons systems, and other advanced technology.[87] The coordination criterion merely builds on this foundation.

In the Boeing Max 737 episode, a single sensor operated to exaggerate the risk of a stall and thus trigger a downward plunge in the aircraft's nose that resulted in two catastrophic crashes.[88] Additional sensors would have more readily detected ambiguous information, suggesting that a stall was not imminent and that automatic depression of the aircraft's nose

85 See Russell and Norvig (n 26) 1005; Domingos (n 26) 238.

86 Anna L Buczak and Ethan Guven, 'A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection' (2016) 18(2) IEEE Communications Surveys and Tutorials 1153, 1162–70; see generally Mark Raymond, 'Engaging Security and Intelligence Practitioners in the Emerging Cyber Regime Complex' (2016) 1(2) Cyber Defense Review 81, 92 (discussing human red-teaming in cyber arena); US Cyberspace Solarium Commission (n 17) 22 (discussing red-teaming of preliminary policy proposals to identify their weaknesses). On issues of attribution in cyber incursions, see Dennis Broeders, Els De Busser and Patryk Pawlak, 'Three Tales of Attribution in Cyberspace: Criminal Law, International Law, and Policy Debates' (Hague Program for Cyber Norms Policy Brief 2020) 7–8, <https://www.universiteitleiden.nl/en/research/research-output/governance-and-global-affairs/three-tales-of-attribution-in-cyberspace.-criminal-law-international-law-and-policy-debates>; Nicholas Tsagourias, 'Cyber Attacks, Self-Defence and the Problem of Attribution' (2012) 17 Journal of Conflict and Security Law 229.

87 See Andre Kohn and others, 'Markov Chain-based Reliability Analysis for Automotive Fail-Operational Systems' (2017) 5(1) SAE International Journal of Transportation Safety 30, 32. Fail-safe capabilities are a common feature of advanced systems. These features minimize risk in the event of malfunction.

88 See Chris Hamby, 'How Boeing's Responsibility in a Deadly Crash "Got Buried"' (*New York Times*, 20 January 2020) <https://www.nytimes.com/2020/01/20/business/boeing-737-accidents.html>.

was not necessary. That redundancy in autonomous systems is one way that victim States can properly gauge connections between systems in States responsible for initial actions, thus ensuring that countermeasures minimize harm to unrelated systems in the responsible State.

C REPAIRS: A PATCH IN TIME

As another modification that is appropriate for the cyber domain, the approach taken in this chapter requires — where feasible — that a State assist in repairs of collateral damage caused by autonomous cyber agents. In cyber's virtual arena, a patch may often remedy damage quickly, in contrast with the time-consuming physical repairs that may be required for the effects of kinetic attacks. Suppose that a victim State's response has caused collateral harm to third-party States or unrelated or civilian networks in a responsible State. If the victim State can feasibly provide a timely and effective patch, the approach to precautions taken here would require that action. In addition, if the collateral harms fell below the use of force threshold, the law of countermeasures would require that — where feasible — the victim State provide a patch as part of its duty to ensure that the effects of countermeasures are both temporary and reversible.

In an armed conflict or even peacetime cyber exchanges with adversary States, sharing patches may be more difficult. For example, the US may be wary of sharing software patches with adversaries such as Russia and China. In these situations, an international organization might be needed as an intermediary. Of course, IHL already uses trusted intermediaries for matters such as providing aid to civilians in war zones. Organizations analogous to the International Committee of the Red Cross or Doctors Without Borders could be established to act as clearinghouses for patching information. While sharing information may still not be practicable in such situations, the evanescence of vulnerabilities once used should encourage more information-sharing, much as States share information with international health and humanitarian groups.[89]

As an example, suppose that Arcadia has used malware as part of a cyber countermeasure responding to an intrusion by Pacifica, but in the process has impaired the functionality of software in Ruritania. Arcadia

89 Of the steps suggested here, the provision of repairs is both the least practicable and the greatest departure from the *lex lata*. It may be useful to consider State commitment to the three other steps outlined here — reconaissance, coordination, and review — as an alternative approach that would still yield a margin of appreciation.

promptly acknowledged responsibility for the harm to Ruritania, and provided a patch that restored the functionality of adversely affected operating systems. Assuming that Arcadia incurred no substantial costs through this action that might have reduced its feasibility, the approach to precautions taken in this chapter would require that Arcadia provide the patch to Ruritania. Furthermore, while prompt provision of an effective patch would not completely remove the harm to Ruritania from the proportionality calculus applicable to countermeasures, it would reduce the quantum of harm subject to this calculus.

D REVIEW: UNPACKING THE UNINTELLIGIBLE

In assessing how reconnaissance, coordination, and repairs have performed, review is essential. In IHL, review is part of a State's duty to exhibit 'constant care' in reducing needless harm to civilians.[90] Proportionality in the *jus ad bellum* and countermeasures requires review, as well. A State that has engaged in methodical review of past operations inspires trust that it will learn the right lessons from previous missteps. That review should be independent, to avoid the groupthink that can undermine neutral evaluation. Moreover, review in the autonomous cyber context depends on a State ensuring that its agents' outputs are sufficiently explainable to facilitate review.

In the *jus in bello*, review of a weapon starts prior to deployment with Article 36 of Additional Protocol I, which requires a finding that a weapon is not inherently indiscriminate.[91] Article 36 reviews have a low threshold: a State need only find that some use of a weapon is consistent with IHL. For example, if a State can show that in a particular context, it can use a weapon to target an adversary's force, that weapon has met the requirements of Article 36. Review under Article 36 is vital where this duty applies, but it is more limited than the concept of review advanced here. First, a State's use of cyber may not be a weapon in the Article 36

90 See AP I, art 57(1).

91 AP I, art 36; see also *William H Boothby, Weapons and the Law of Armed Conflict* (2nd edn, Oxford University Press 2016) 347–8; Michael W Meier, 'Lethal Autonomous Weapons Systems (LAWS): Conducting a Comprehensive Weapons Review' (2016) 30 Temple International and Comparative Law Journal 119, 124–6. Even if an autonomous cyber agent passes a weapons review, designers will need to validate its use for particular purposes. Cf Margaret Hu, 'Small Data Surveillance v. Big Data Cybersurveillance' (2015) 42 Pepperdine Law Review 773, 812–16 (urging use of a rigorous test to validate machine learning models).

sense of the term.[92] Second, Article 36 does not apply to countermeasures or other actions taken outside armed conflicts. Third, review here stems from the concept of after-action review in IHL.

After-action review in IHL and provisions for review under international human rights law ('IHRL')[93] are more expansive in scope than pre-deployment Article 36 review. Because the combination of cyber and autonomy is so new, review should be systemic, not merely focused on a specific incident. A State investigating an alleged war crime by one of its service members has no duty to consider whether it should forego the use of humans in future military engagements. The use of humans is sufficiently well-established to render any such inquiry unnecessary. In contrast, depending on the seriousness of the outcomes a review would assess, the novel technology of autonomous cyber agents may require a more searching review of the appropriateness of their deployment.

Such reviews entail a more robust form of independence than the fact-specific detachment required under customary IHL.[94] Under the *lex lata*, an investigation of alleged war crimes is sufficiently independent if it does not suffer from command influence that skews the investigation's analysis and conclusions. However, IHRL has been moving toward a more robust conception that requires greater structural independence from the chain of command.[95] Under the approach taken in this chapter, a more robust structural approach would be required in IHL — recognizing the move in that direction in State practice — and in the *jus ad bellum* and countermeasures.

Review must include efforts to explain the outputs of autonomous agents. As noted earlier, explainability is a challenge for certain forms of artificial intelligence. In particular, neural networks generate outputs that are difficult to explain through conventional verbal means, since the layers that contribute to neural networks' accuracy sift through so many variables. To accommodate this concern, review will need to define explainability more broadly, while ensuring that review is rigorous. A workable conception of review should recognize that there are many ways of enhancing explainability and addressing errors. Moreover, a State should be able to show that it is continually working on more effective means for addressing this concern. Commitment to a reasonable

92 See Jeffrey T Biller and Michael N Schmitt, 'Classification of Cyber Capabilities and Operations as Weapons, Means, or Methods of Warfare (2019) 95 International Law Studies 179.
93 Michael N Schmitt, 'Investigating Violations of International Law in Armed Conflict' (2011) 2 Harvard National Security Journal 31, 80.
94 ibid 50–1.
95 ibid 49–51.

framework of review is more productive than prescribing or prohibiting a particular technology.

In our malware hypothetical, review might be required to determine the cause of mistakes and seek to correct tactics, techniques, and procedures ('TTP') in the future. For example, suppose that Arcadia had used malware embedded in Pacifica's networks to respond to a Pacifica incursion, but that malware had targeted civilian networks in a fashion that manifestly violated the *jus in bello*, *jus ad bellum*, or the rule of proportionality in countermeasures. Arcadia would be required to conduct a review to determine the cause of its mistake and discern means to avoid comparable mistakes in the future. Conducting that review would entail the capacity to discern *why* the agent made the mistake. For example, designers reviewing the agent's performance could seek to reverse-engineer that performance with counter-factuals, to determine what inputs or architecture would have had to change to lead to a different result.

Upon review, designers could determine that they needed to use more elaborate coordination between autonomous learners to detect potential errors and modify the agent's outputs before they created harm. Under the approach taken here, designers would then have to implement the findings of their review. That dedication to review would facilitate continual improvement in compliance with international law.

V
CONCLUSION

In a cyber world where the need for speed is paramount, observing proportionality is crucial. Human designers and operators lack the agility to respond to every-mounting cyber incursions. Autonomous cyber agents can pick up the slack.

In the *jus ad bellum*, *jus in bello*, and the law of countermeasures, proportionality plays an important role in reducing harm and the risk of escalation. However, the amorphous character of proportionality makes it difficult to implement this value across each of the legal arenas described above.

Attributes of autonomy also hinder that mission. Along with their extraordinary speed and analytical prowess, autonomous agents have

notable flaws, including brittleness, bias, and unintelligibility. Beset by automation bias, human designers and operators struggle to acknowledge and address these flaws.

Hamstringing victim States is no answer to autonomy's deficits. In decisions about the use of force, the conduct of armed conflict, and the launching of countermeasures, undue restrictions will force victim States to cede the initiative to first movers who violate international law in search of an edge. States will reject any legal duty that yields this perverse result. A balance is necessary that encompasses the need for speed in victim State responses while ensuring that those responses remain within reasonable bounds.

The approach taken in this chapter seeks to accomplish that goal. It confers a margin of appreciation on victim States' responses. However, that margin of appreciation requires victim States to observe feasible precautions. Those precautions have both independent substantive significance as a component of proportionality and evidentiary value as proof of a victim State's compliance with international law. Necessary precautions are reconnaissance, coordination, repair, and review. Fulfilling those conditions will allow victim States to wrest the initiative from responsible States, while keeping their own responses in check. That balance will preserve stability in the cyber domain and international law.

Chapter 9

Autonomy and Precautions in the Law of Armed Conflict

Eric Talbot Jensen[1]

I
INTRODUCTION

Fixating on what amount of human control is required in the employment of autonomous weapons, including autonomous cyber capabilities, erroneously disregards the most important question with respect to autonomy in armed conflict. The question is whether autonomous weapons can 'select' and 'attack' targets in a manner that complies with the law of armed conflict ('LOAC').[2] Some argue that to comply with the LOAC, selecting and targeting *requires* human judgment. There is no consensus on that assertion. Indeed, States that are Parties to the Certain Conventional Weapons Convention ('CCW') have not acknowledged

1 The author would like to thank Summer Crockett and Carolyn Sharp for excellent research and review assistance.

2 For purposes of this chapter, I will use LOAC and IHL interchangeably, though I recognize that some may argue that they are different in both content and approach to regulation during armed conflict.

that human involvement in selecting and engaging targets is required under the LOAC.[3] Rather, the views of States vary widely on this issue, precluding the assertion that there is a current prohibition.

This chapter analyzes the specific LOAC rules on precautions in the attack, as codified in Article 57 of Additional Protocol I ('AP I'), and asserts that these rules do not require human judgment in targeting decisions. Rather, these rules prescribe a particular analysis that must be completed. That analysis is one, which, in the future, may be done just as effectively (if not more effectively) by weapons systems using autonomous functions.

Part II of this chapter briefly discusses what 'autonomy' means and highlights that there is no single agreed upon definition. For the purposes of this article, the key aspect of autonomy is that a weapon system can select and attack targets without human intervention. Part III analyzes the argument that human judgment is required for selecting and attacking targets and contrasts that position against the current practice of States and their statements on the issue. Further, this Part looks specifically at the requirements of precautions, as codified in Article 57 of AP I. Part IV concludes finding that no requirement for human judgment in selecting and attacking targets currently exists.

II
AUTONOMY

Autonomy, in particular the use of autonomy in weapon systems, is a major point of discussion between States. As Masahiro Kurosaki writes, '[A]utonomy in unmanned systems will be critical to future conflicts

3 One example is the Israeli Harpy NG. According to Shelby Smith, 'Technology Explainer: Automated Defense Technology' (2019) 3 Georgetown Law Technology Review 492, 499:

> One example of an autonomous weapon system is the loitering munition. Loitering munitions, which hover over a human-designated area and strike at targets that match specific parameters, are currently only employed in Israel. The Harpy NG, the most commonly used and advanced model manufactured by Israel Aerospace Industries, is designed to attack enemy radar systems. These loitering munitions resemble drones, or UAVs, and can stay in the air for up to nine hours. Because loitering munitions are set up with specific limits to their range, they may offer a model for future development of autonomous weapons that afford an element of control without the need for human monitoring.

Other systems include the Phalanx CIWS and the C-RAM. See US Army, 'Counter-Rocket Artillery Mortar (C-RAM) Intercept Land-Based Phalanx Weapon System (LPWS)' <https://asc.army.mil/web/portfolio-item/ms-c-ram_lpws/> accessed 9 October 2020; Raytheon Missiles and Defense, 'Phalanx Weapon System' <https://www.raytheonmissilesanddefense.com/capabilities/products/phalanx-close-in-weapon-system> accessed 9 October 2020.

that will be fought and won with technology.'[4] Many of these unmanned systems will be either assisted by or based almost completely on cyber capabilities.

Within the last ten years, formal discussions on autonomous weapons, or weapons that rely on autonomous functions such as machine learning or artificial intelligence, have failed to produce a common understanding of what 'autonomy' even means.[5] As Chris Jenks notes, 'the international community cannot even agree about what they disagree about.'[6]

To some degree, the position individuals or States take on autonomous weapons may be influenced by the definitional decision on autonomy. For example, the ICRC defines autonomous weapon systems as 'weapon systems with autonomy in their "critical functions" of selecting and attacking targets.' Further, the organization takes the approach that such systems would are 'an immediate concern from a humanitarian, legal and ethical perspective, given the risk of loss of human control over weapons and the use of force.'[7]

By contrast, the United Kingdom approaches autonomy more broadly, stating, '[f]ocusing solely on specific — or "critical" — functions or activity in the lifecycle of a weapon is unlikely to be sufficient to ensure there is human control.'[8] The United Kingdom argues that basing regulation on the characterization of a system's function is unhelpful. Instead, it asserts that 'it is the cumulative effect of multiple safeguards across the development and operational lifecycle that establish human control of weapon systems. Therefore, human control should be considered and exercised throughout this lifecycle and in a way that is appropriate to the operational context.'[9]

Such disparate views cause legal experts like Chris Jenks and Rain Liivoja to conclude that 'autonomy is better thought of across several

4 Masahiro Kurosaki, 'Toward the Special Computer Law of Targeting: "Fully Autonomous" Weapons Systems and the Proportionality Test' in Claus Kreß and Robert Lawless (eds), *Necessity and Proportionality in International Peace and Security Law* (Oxford University Press 2020).

5 Chris Jenks, 'False Rubicons, Moral Panic, and Conceptual Cul-De-Sacs: Critiquing & Re-framing the Call to Ban Lethal Autonomous Weapons' (2016) 44 Pepperdine Law Review 1, 12.

6 ibid. See also Heather M Roff and Richard Moyes, 'Meaningful Human Control, Artificial Intelligence and Autonomous Weapons' (Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, 11–15 April 2016) <http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf>.

7 ICRC, 'Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centered Approach' (6 June 2019) <https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach>.

8 ibid.

9 United Kingdom, Statement regarding Agenda Item 5(d) (Meeting of the Group of Governmental Experts on Lethal Autonomous Weapons Systems, 25–29 March 2019) <https://www.unog.ch/80256EDD006B8954/(httpAssets)/1ED3972D40AE53B5C12583D3003F8E5E/$file/20190318-5(a)_IHL_Statement.pdf> ('UK Statement').

different spectrums.' They further add that 'attempts at overall system categorization based on only one of the spectrums — machine complexity — lack practical utility.'[10]

For the purposes of this chapter, a weapon system is autonomous 'when it possesses both an intent (an encoded representation of a goal, a purpose, or a task to be completed) and the ability to act within its environment in furtherance of that goal.'[11] Under this definition, autonomous weapons systems, including autonomous cyber capabilities, could be subject to human control, but also may function without constant or even decisive human control, including during the processes of selecting and engaging targets.

Weapons and weapons systems that are autonomous in this sense have raised the ire of many nongovernmental organizations ('NGOs') and the International Committee of the Red Cross ('ICRC'), and have become the basis for much of the debate among States — particularly in the meetings of States Party to the CCW. Some scholars have argued that the CCW is the perfect forum to hear these debates and regulate autonomous weapons.[12] Part III will analyze these discussions.

One additional definitional caveat is important. The consideration here of autonomous weapon systems is distinct from the question of weapons that may in the future utilize artificial intelligence. Artificial intelligence includes cognition.[13] While an autonomous weapon system

10 Chris Jenks and Rain Liivoja, 'Machine Autonomy and the Constant Care Obligation' (*Humanitarian Law & Policy*, 11 December 2018) <https://blogs.icrc.org/law-and-policy/2018/12/11/machine-autonomy-constant-care-obligation/>.

11 Tim McFarland, 'The Concept of Autonomy', this volume, ch 2, 21.

12 Qiang Li and Dan Xie, 'Legal Regulation of AI Weapons Under International Humanitarian Law: A Chinese Perspective' (*Humanitarian Law & Policy*, 2 May 2019) <https://blogs.icrc.org/law-and-policy/2019/05/02/ai-weapon-ihl-legal-regulation-chinese-perspective/>. Where the authors argue:

> Moreover, the targeting of AI weapon systems is closely tied to their design and programming. The more autonomy they have, the higher the design and programming standards must be in order to meet the IHL requirements. For this purpose, the international community is encouraged to adopt a new convention specific to AI weapons, such as the Convention on Conventional Weapons and its Protocols, or the Convention against Anti-personnel Mines and Convention on Cluster Munitions.

13 As Dustin Lewis has written:

> AI science pertains in part to the development of computationally based understandings of intelligent behavior, typically through two interrelated steps. One of those steps concerns the determination of cognitive structures and processes and the corresponding design of ways to represent and reason effectively. The other step relates to the development of theories, models, data, equations, algorithms and/or systems that embody that understanding. …
>
> So defined, AI systems are typically conceived as incorporating techniques — and leading to the development of tools — that enable systems to 'reason' more or less 'intelligently' and to 'act' more or less 'autonomously'. The systems might do so by, for example, interpreting natural languages and visual scenes; 'learning' (or, perhaps more commonly, training); drawing inferences; and making 'decisions' and taking action on those 'decisions'. The techniques and tools might be rooted in one or more of the following methods: those rooted in logical reasoning broadly conceived, which are sometimes also referred to as 'symbolic AI' (as a form of model-based methods); those rooted in probability (also as a form of model-based methods); and/or those

does not necessarily require cognition to 'learn' on the battlefield, this chapter considers weapons systems that use artificial intelligence or machine learning to adjust decision-making processes, but not weapons systems that are cognitive.

III
HUMAN JUDGMENT AND PRECAUTIONS

I argue more in-depth elsewhere that the LOAC does not require weapons that utilize machine learning or artificial intelligence to be limited by some inclusion of human judgement in the processes of selecting and engaging targets.[14] I will briefly restate various views on this question to facilitate a discussion of how weapons that use machine learning and artificial intelligence, including cyber weapons, might be governed by the LOAC, including the rules on precautions.

Initially, it is important to confirm that the LOAC applies to the use of emerging technologies in general and to autonomous weapon systems or weapons that use machine learning and artificial intelligence in particular. This view is shared by both States[15] and by NGOs.[16] However, great debate exists as to how those weapon systems might comply with the LOAC.

rooted in statistical reasoning and data (as a form of data-dependent or data-driven methods). 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/> (footnotes omitted).

14 Eric Talbot Jensen, 'The (Erroneous) Requirement for Human Judgment (and Error) in the Law of Armed Conflict' (2020) 96 International Law Studies 26.

15 See, eg, Brazil, Statement regarding Agenda Item 5(a) (Group of Governmental Experts on Lethal Autonomous Weapons Systems, 25–29 March 2019) 1–2 <https://www.unog.ch/80256EDD006B8954/(httpAssets)/122DF2DAEE334DDBC12583CC003EFD6F/$file/Brazil+GGE+-LAWS+2019+-+Item+5+a+-+IHL.pdf>; The Netherlands, Statement on Agenda Item 5(a) (Group of Governmental Experts on Lethal Autonomous Weapons Systems, 26 April 2019) <https://www.unog.ch/80256EDD006B8954/(httpAssets)/A2E0497EE93C232AC12583CB0037813B/$file/5a+NL+-Statement+Legal+Challenges-final.pdf>; Poland, 'General Comments' (Statement to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, 25 March 2019) 1 <https://www.unog.ch/80256EDD006B8954/(httpAssets)/5CAD5A1367E305A5C12583CC004CA205/$file/1.+GGE_LAWS_March+2019_PL+Statement_General+comments_25.03.2019.pdf >; European Union, 'An Exploration of the Potential Challenges Posed by Emerging Technologies in the Area of Lethal Autonomous Weapons Systems to International Humanitarian Law' (Statement to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, 25–29 March 2019) <https://www.unog.ch/80256EDD006B8954/(httpAssets)/EA84B3C2340F877DC12583CB003727F3/$file/ALIGNED+-+LAWS+GGE+EU+statement+IHL.pdf> ('EU Statement'); ICRC (n 7) 2.

16 Netta Goussac, 'Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting' (*Humanitarian Law & Policy*, 18 April 2019) <https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/>. See 'Campaign to Stop Killer

In the past decade, various organizations have argued that any use of autonomous weapons would be unlawful because of the non-human element and have called for a ban on research and development of these weapons.[17] The ICRC, while acknowledging the key role of States in this discussion,[18] takes the following view:

> These rules require context-specific judgements to be taken by those who plan, decide upon and carry out attacks to ensure: distinction — between military objectives, which may lawfully be attacked, and civilians or civilian objects, which must not be attacked; proportionality — in terms of ensuring that the incidental civilian harm expected from an attack will not be excessive in relation to the concrete and direct military advantage anticipated; and to enable precautions in attack — so that risks to civilians can be further minimized.
>
> Where AI systems are used in attacks — whether as part of physical or cyber-weapon systems, or in decision-support systems — their design and use must enable combatants to make these judgements.[19]

In response to this argument, Masahiro Kurosaki counters that '[T]he existing human-centered paradigm is merely a product of the history of LOAC and does not exist a priori, an alternative approach to adjust to changing times, should be explored.'[20]

States have taken widely disparate views on these questions. For example, in response to the call for a ban on autonomous weapon systems, the United Kingdom argues:

> in the absence of any clearly articulated empirical evidence as to why existing regulation — including IHL — is inadequate to control developments in emerging technologies, the issue may well lie not with the processes themselves, but with the perceived

Robots' <https://www.stopkillerrobots.org/> accessed 2 March 2020, where the author states: What remains beyond question is that all weapons used in war must be used, and be capable of being used, in compliance with IHL. This means that each State that develops or acquires weapons that utilize AI must be satisfied that these weapons can be used in compliance with existing rules of warfare.

17 See ibid.
18 ICRC (n 7) 2.
19 ibid.
20 Kurosaki (n 4) 15.

> ability of machines to assimilate, understand and meet the relevant extant legal and ethical standards.[21]

Greece[22] and Germany[23] have supported the view that the LOAC requires a degree of human control in selecting and engaging targets, but as Rebecca Crootof notes there is little clarity on the specifics of that control.[24] As an example of the differing views on how human control might manifest in an autonomous weapon system, the United Kingdom states:

> [D]irect human involvement in every detailed action of a system or platform may not be practical or desirable under all circumstances. Instead a human-centred approach to autonomous technologies must take into account the operational context as well as the capabilities and limitations of the personnel deploying the weapon system.[25]

This operational context might include considerations such as whether the system is a land, air or sea-based system and the specific circumstances of both the development and the deployment of the system.[26]

In 2019, the US Department of Defense General Counsel Paul Ney argued that autonomy makes and will continue to make weapons systems more accurate, more precise, and able to perform much more quickly.[27] In perhaps the strongest statement against the fixation on human control, Ney stated:

21 UK Statement (n 9). The UK goes on to argue:
We argue that weapons systems that cannot meet these standards will remain incapable of legal use as set out in existing national and international normative frameworks and will not be developed, fielded and used. All states should look to ensure they meet the basic obligations already set out in the relevant articles of IHL before pressing for bespoke legislation for as-yet undefined capabilities.

22 Greece, 'Potential Challenges Posed by Emerging Technologies in the Area of Lethal Autonomous Weapons Systems to International Humanitarian Law' (Statement to the Group of Governmental Experts on Lethal Autonomous Weapon Systems, 25–29 March 2019) <https://www.unog.ch/80256EDD006B8954/(httpAssets)/D1B935 800DF5F04DC12583CC002F3DD1/$file/GGE+LAWS+-STATEMENT+by+ GREECE-+Challenges+to+IHL.pdf>.

23 Germany, Statement on Agenda Item 5(b) (Group of Governmental Experts on Lethal Autonomous Weapons Systems, 26 March 2019) <https://www.unog.ch/80256EDD006B8954/(httpAssets)/2B8E772610C0F552C12583CB003A4192/$file/20190326+Statement3+Germany+GGE+LAWS.pdf>.

24 Rebecca Crootof, 'A Meaningful Floor for "Meaningful Human Control"' (2016) 30 Temple International and Comparative Law Journal 53, 54.

25 UK Statement (n 9).

26 ibid.

27 Paul C Ney Jr, 'Keynote Address at the Israel Def. Forces 3rd International Conference on the Law of Armed Conflict' (*Lawfare*, 28 May 2019) <https://www.lawfareblog.com/defense-department-gen-eral-counsel-remarks-idf-conference>.

> In the US. perspective, there is nothing intrinsically valuable about manually operating a weapon system as opposed to operating it with an autonomous function. For example, existing law of war treaties do not seek to enhance "human control" as such. Rather, these treaties seek, among other things, to ensure the use of weapons consistent with the fundamental principles of distinction and proportionality, and with the obligation to take feasible precautions for the protection of the civilian population. Although "human control" can be a useful means in implementing these principles, "human control" as such is not, and should not be, an end in itself. In our view, we should not be developing novel principles that stigmatize the use of emerging technologies, when these technologies could significantly enhance how the existing principles of the law of war are implemented in military operations.[28]

Two points appear clear from this brief review of State perspectives. First, there is a consensus that all autonomous weapons systems developed and employed must comply with the LOAC. Second, there is not a consensus as to the degree of human control necessary to comply with the LOAC.

Echoing Ney's statement above, the focus of international regulation should be on LOAC compliance, and not on who or what is bringing about that compliance. As I conclude elsewhere:

> [T]he legal standard for weapon systems using machine learning and artificial intelligence should be the "best application possible" rather than the "best application humanly possible." International focus on the decisions of warfare, rather than the decision-makers, will benefit all concerned and result in greater protections for the participants in and the victims of armed conflict.[29]

With that foundation, a more specific analysis of precautions in the attack, as codified in Article 57 of AP I, is in order to determine key focus areas in ensuring compliance with the LOAC — particularly by militaries that develop and employ autonomous systems to select and engage targets.

28 ibid. For additional statements by the United States in the context of the CCW discussions, see Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (23 October 2018) UN Doc CCW/GGE.1/2018/3.

29 Jensen (n 14) 57.

A ARTICLE 57(1) — 'CONSTANT CARE'

Although Article 57(1) falls under the heading of 'Precautions in Attack', its broad coverage, to includes the conduct of 'military operations' generally.[30] The obligation on States is one of 'constant care'. Autonomous systems have already been created to take an active role in non-combat military operations (for example logistics).[31] Although this chapter focuses on the use of autonomy in combat situations, non-combat autonomous systems can also cause death or injury and therefore deserve some comment here.

Jenks and Liivoja have addressed the issue of autonomy with non-combat vehicles. They argue:

> Article 57(1) would require that autonomous vehicles be designed and relied upon with the safety of the civilian population in mind. Thus, an autonomous ground vehicle should avoid, for example, injuring civilians or damaging civilian building[s] and infrastructure. Likewise, an autonomous aerial vehicle should be capable of avoiding civilian air traffic and not crash into and damage civilian objects upon a failure of the communication link to its operator.[32]

This quote highlights the fact that non-combat autonomous systems may still lead to death or injury and thus commanders need to employ them with constant care for the civilian population.

The constant care obligation applies equally to autonomous cyber operations. As I write elsewhere, 'commanders and all persons conducting cyber operations must recognize and accept the legal obligation to exercise constant care in all military operations, including cyber operations.'[33]

The *Tallinn Manual 2.0* also takes this position stating, 'During hostilities involving cyber operations, constant care shall be taken to spare the civilian population, individual civilians, and civilian objects.'[34] The

30 The Commentary to AP I states that '[t]he term "military operations" should be understood to mean any movements, manoeuvres and other activities whatsoever carried out by the armed forces with a view to combat.' Claude Pilloud and Jean de Preux, 'Protocol I — Article 57 — Precautions in Attack' in Yves Sandoz, Christophe Swinarski and Bruno Zimmermann (eds), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (ICRC 1987) 680, [2191].

31 Jon Harper, 'Autonomous Helicopters Seen as Wave of the Future' (*National Defense Magazine*, 2 February 2018) <https://www.nationaldefensemagazine.org/articles/2018/2/20/autonomous-helicopters-seen-as-wave-of-the-future>.

32 Jenks and Liivoja (n 10) 5.

33 Eric Talbot Jensen, 'Cyber Attacks: Proportionality and Precautions in Attack' (2013) 89 International Law Studies 198, 204.

34 Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*

Group of Experts was unanimous in the formulation of this rule and also argued that

> Use of the word 'constant' denotes that the duty to take care to protect civilians and civilian objects is of a continuing nature throughout all cyber operations; all those involved in the operation must discharge the duty. The law admits of no situation in which, or time when, individuals involved in the planning and execution process may ignore the effects of their operations on civilians or civilian objects. In the cyber context, this requires situational awareness at all times, not merely during the preparatory stage of an operation.[35]

The 'constant' nature of this requirement applies equally to autonomous cyber systems. In designing and utilizing such systems, even outside the context of an attack, military operators must ensure that the autonomous system can exercise constant care.

B ARTICLE 57(2)

Article 57(2) codifies the current codification of the customary law on applying 'precautions in the attack'. These provisions are among those that are recognized as binding on all States that desire to utilize weapons — whether autonomous or not. The question raised by autonomous weapon systems is whether these systems can comply with the requirements as stated in Article 57. Professor Suresh Venkatasubramanian perhaps best describes this question:

> If we look at the principles of distinction, proportionality and precautions under international humanitarian law as guidance for when an attack is considered permissible, we see a lot of judgement framed in terms that to a computer scientist seem imprecise. One might argue that the vagueness in these terms is by design: it allows for nuance and context as well as human expert judgement to play in a role in a decision, much like how the discretion of a

(Cambridge University Press 2017) 476 ('*Tallinn Manual 2.0*'). Note that the author was a member of the Group of Experts for both the *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Cambridge University Press 2013) and the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017).

35 ibid 477 (citations omitted).

> judge plays a role in judging the severity of a sentence. Another view of this 'vagueness by design' is that it allows for future contestability: if commanders are forced to defend a decision later on, they can do so by appealing to their own experience and judgement in interpreting a situation. But what of algorithm-driven targeting? How is a system supposed to learn what targets satisfy principles of proportionality, distinction and precaution when to do so it must rely on a precise labeling that almost cannot exist by design.[36]

Accordingly, this section will analyze the legal requirements contained in the subsections of Article 57(2) and (3) argue that despite potential technological and conceptual limitations, none of those sections present an insurmountable legal obstacle to the use of autonomous weapon systems.

1 *'Those who plan or decide upon an attack …'*

Beginning with Article 57(2), the first provision bearing on the use of autonomous weapon systems, including weapons with autonomous cyber capability, is subparagraph (a). That subparagraph specifically regulates those who plan or decide upon an attack. The ICRC takes the following view:

> International humanitarian law (IHL) requires that those who plan, decide upon and carry out attacks make certain judgements in applying the norms when launching an attack. Ethical considerations parallel this requirement — demanding that human agency and intention be retained in decisions to use force.[37]

Although it is not clear from the text of this provision that human judgment is required, the ICRC argues that both the legal and ethical considerations require human judgment. Others have taken the same approach, arguing specifically that this provision establishes an accountability mechanism that precludes autonomous systems.

36 Suresh Venkatasubramanian, 'Structural Disconnects Between Algorithmic Decision-Making and the Law' (*Humanitarian Law & Policy*, 25 April 2019) <https://blogs.icrc.org/law-and-policy/2019/04/25/structural-disconnects-algorithmic-decision-making-law/>.

37 ICRC, 'Towards Limits on Autonomy in Weapon Systems' (Statement to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, 9 April 2018) s 7 <https: //www.icrc.org/en/document/towards-limits-autonomous-weapons>. The ICRC continues:
From the ICRC's perspective, ethical considerations parallel the requirement for a minimum level of human control over weapon systems and the use of force to ensure legal compliance. From an ethical viewpoint, "meaningful", "effective" or "appropriate" human control would be the type and degree of control that preserves human agency and upholds moral responsibility in decisions to use force. This requires a sufficiently direct and close connection to be maintained between the human intent of the user and the eventual consequences of the operation of the weapon system in a specific attack.

For example, Roff and Moyes have described accountability as follows:

> [A]n ex post process to locate responsibility or liability with human agents, … [that] also establishes a framework of expectation that can guide human agents to align their behavior with expected and appropriate standards. Standards for accountability, moreover, need to ensure that responsibility and liability will be apportioned equitably, and that sanctions will be applied that are commensurate with the wrongdoing (whether intentional or inadvertent) and with the severity of harm that may have been caused.[38]

As inferred above, some complain that autonomous weapons systems that select and engage targets would leave no method of accountability for decisions that violated the LOAC. Others have countered that accountability is not, and has not been, focused solely on the person pulling the trigger, which, in the case of autonomous weapons systems, would mean the system itself.[39] The language of those who 'plan or decide' is obviously meant to include not just the trigger puller, but also those at all levels of command and decision making. This would include, in particular, those who order autonomous weapons systems into battle. As shown below, the Commentary and the statements of the delegations to the negotiating conference that led to AP I confirm this understanding.

The use of the phrase 'plan or decide' was a topic of discussion at the AP I negotiating conference. As the 1987 ICRC Commentary states:

> The terminology used in this provision led to some criticism and explanatory statements. Some considered that the introductory words ("those who plan or decide upon an attack") could lay a heavy burden of responsibility on subordinate officers who are not always capable of taking such decisions, which should really fall upon higher ranking officers. This view is not without grounds, but it is clear that a very large majority of delegations at the Diplomatic Conference wished to cover all situations with a single provision, including those which may arise during close

38 Roff and Moyes (n 6) 3.

39 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/>.

> combat where commanding officers, even those of subordinate rank, may have to take very serious decisions regarding the fate of the civilian population and civilian objects. It clearly follows that the high command of an army has the duty to instruct personnel adequately so that the latter, even if of low rank, can act correctly in the situations envisaged.[40]

Many statements made by the delegations at the conference support this view. For example, the Swiss delegation stated that it 'was critical of paragraphs 2 and 3 of Article 50 because they lacked clarity; particularly the words 'Those who plan or decide upon an attack …' in paragraph 2 (a).'[41] Others, including Afghanistan,[42] Austria,[43] Netherlands,[44] and Sweden[45] echoed this statement.

Contemporary commentators express the same concerns. For example, Rebecca Crootof, in speaking about the command levels at which human control should be exercised, writes:

> [T]here is still no agreement as to the level of decision-making at which human control must occur. The commander determining the rules of engagement is exercising a certain kind of control, the commander ordering a particular attack is exercising another, and the individual implementing that order might exercise yet another kind of control.
>
> Given the difficulty in pinning down what "meaningful human control" actually requires, "[s]everal states [have] expressed skepticism over the added value of the suggested concept, assessing it as being too vague, subjective and unclear.[46]

Moreover, Roff and Moyes have also argued that

40 Pilloud and de Preux (n 30) 681, [2197].
41 *Official Records of the Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applied in Armed Conflicts (Geneva, 1974–1977)* CDDH/SR.42, 212, [43]. He continued, 'That ambiguous wording might well place a burden or responsibility on junior military personnel which ought normally to be borne by those of higher rank. The obligations set out in Article 50 could concern the high commands only — the higher grades of the military hierarchy, and it was thus that Switzerland would interpret that provision.'
42 ibid 219.
43 ibid 212, [46].
44 ibid 205, [1].
45 ibid 236–7.
46 Crootof (n 24) 58.

> At its most basic level, the requirement for [meaningful human control] develops from two premises: 1. That a machine applying force and operating without any human control whatsoever is broadly considered unacceptable. 2. That a human simply pressing a 'fire' button in response to indications from a computer, without cognitive clarity or awareness, is not sufficient to be considered 'human control' in a substantive sense.[47]

In responding to the second point raised by Roff and Moyes, Merel Ekelhof poses an interesting scenario in which a fighter pilot is sent on an attack mission to deliver ordnance on an enemy position. As is normal in military operations, a targeting cell, which includes a lawyer, reviewed the target prior to its approval. The pilot is then assigned the mission, briefed on the intelligence situation, and given specific details about the target — all of which is also loaded into the aircrafts targeting systems. In this particular example, poor weather prevents the pilot from having good visibility of the target. Ekelhof argues that in such circumstances, the pilot will have to 'rely on the aircraft's systems, the weapons guidance systems, and the validation procedure at the operational level to ensure s/he is striking a legitimate military objective in a lawful manner.'[48] Accordingly, she continues:

> Thus, the information about the lawfulness of the action largely depends on the operator's trust in his or her superiors in the chain of command (to provide proper briefing materials and conduct target validation during the planning phase), the F-16 onboard computer (suggesting the appropriate time for weapon's release) and the weapon's guidance system (navigating the munitions to the target). At no point during our F2T2EA process will the pilot gather intelligence about the target or conduct legal analyses.[49]

Ekelhof's scenario aptly illustrates the point that the deliverer of the ordnance — the individual attacking — is doing so having neither seen the target, nor verified the intelligence. Such attacks take place all the time in modern warfare. Similar scenarios can be described with respect to artillery and most 'beyond the line of sight' weapons.

After analyzing this common scenario, Ekelhof concludes:

47 Roff and Moyes (n 6) 1.
48 Ekelhof (n 39).
49 ibid.

> the concept of meaningful human control is not the only, or perhaps the most fitting, approach to analyzing (the effect of autonomous technologies on) human control over critical targeting decisions. Instead, the more appropriate analytical lens would be one that recognizes the distributed nature of control in military decision-making in order to pay due regard to a practice that has shaped operations over the past decades and continues to be standard in contemporary targeting.[50]

Ekelhof's scenario and her conclusions highlight the importance of the language in Article 57, which places responsibility for ensuring precautions not only with the 'trigger puller', but also with many others in the military decision-making process. This would, of course, also apply to commanders who employ autonomous weapon systems, including cyber systems.

Arguing that autonomous weapon systems cannot be utilized in conformity with the LOAC because they lack an accountability mechanism is an overly narrow reading of the words in Article 57. The responsibility falls not only to those who execute the attacks (including an autonomous weapons system), but also to those in 'higher commands' such as the local, operational, and strategic military commanders who will employ those weapons systems on the battlefield, and those in the research, production, review, and approval processes. A more holistic understanding of "those who plan or decide upon an attack" leaves no accountability gap.

This analysis applies equally to weapons utilizing autonomous cyber capabilities. In the commentary discussing Rule 115,[51] the *Tallinn Manual 2.0* states

> An important feature of Rule 115 is its focus on planners and decision-makers. Those who execute cyber attacks may sometimes also be the ones who approve them. In the case of certain attacks, the individual actually executing the attack has the capability to determine the nature of the target and to cancel the operation. … On other occasions, the person executing the attack may not be privy to information as to its character or even the identity of the target. He or she may simply be carrying out instructions to deliver the cyber weapon against predetermined.

50 ibid.

51 *Tallinn Manual* (n 34) 478, rule 115 states: 'Those who plan or decide upon a cyber attack shall do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection.'

> Under these circumstances, the duty of the individual carrying out the cyber attack to verify would be limited to those measures that are feasible in the circumstances.[52]

Because of the technology required for cyber attacks, a combination of individuals likely designed and built the cyber tool, determined the accessibility of the target, mapped the 'surrounding' cyber network, installed the malware, and executed the payload. Consider also the additional leaders and commanders at the tactical, operational and strategic level who are not cyber experts but will make significant decisions concerning the employment of cyber tools in their area of operations. To the extent that they 'plan or decide upon' the attack, they all have the legal obligation to comply with this precaution. Despite this potentially expanded field of players in a cyber attack, there is nothing inherent in the technology that would prevent a full and thorough analysis under Article 57. As the *Tallinn Manual 2.0* states:

> The limitation of this Rule to those who plan or decide upon cyber attacks should not be interpreted as relieving others of the obligation to take appropriate steps should information come to their attention that suggests an intended target of a cyber attack is a protected person or object, or that the attack would otherwise be prohibited.

One last comment on this point is important before moving on to further provisions of Article 57. In a recent publication, Laura Dickinson argues that administrative accountability can also play a key role in the lawful use of cyber capabilities during military operations.[53] Dickinson contends that discussions about the potential of administrative accountability to regulate and ensure the compliance of cyber operations with the LOAC have been largely absent. She asserts:

> Such accountability includes multiple administrative procedures, inquiries, sanctions, and reforms that can be deployed within the military or the administrative state more broadly to respond to an incident in which a violation of IHL/LOAC may have occurred.

52 ibid.

53 Laura A Dickinson, 'Lethal Autonomous Weapon Systems: The Overlooked Importance of Administrative Accountability' in Eric Talbot Jensen and Ronald Alcala (eds), *The Impact of Emerging Technologies on the Law of Armed Conflict* (Oxford University Press 2019).

> This form of accountability may be particularly useful in the case of LAWS, because the restrictions of criminal law, such as the intent requirement for most crimes, may not apply in many circumstances. Administrative accountability is flexible both in the process by which it unfolds and in the remedies available, offering the prospect of both individual sanctions as well as broader organizational reforms.[54]

Dickinson's argument for including administrative accountability in the review process further supports an expansive view of accountability. Too narrow a view on accountability unnecessarily limits the application of legal norms to autonomy on the battlefield.

2 *'Do everything feasible to verify' — Distinction*

Article 57(2)(a)(i) effectively restates the LOAC principle of distinction and requires those who plan or decide upon attacks to do everything feasible to verify that the targets are appropriate military objectives. The content of this rule needs no explanation here. The important question for this discussion is whether autonomous weapons systems can apply the principle of distinction, and how that might be assured.

Distinction is often believed to be a principle that requires human judgement and discretion because of the complexity of the decisions on the modern battlefield. Rather, meaningful adherence to distinction is both a technological question and a legal question. An analysis on whether autonomous weapons systems and those which utilize autonomous cyber capabilities are able to satisfactorily comply with the rules of distinction must be assessed through this latter framework. Moreover, whether or when technology will be capable of applying human judgement is beyond the scope of this paper, and not vital to the current discussion. Recent technological developments may allow the integration of biologically realistic neural networks with computer hardware in a way that could create an autonomous weapon with thinking and processing elements.[55] Such developments might significantly alter the discussion concerning the application of human judgment by weapon systems.

However, accepting that technology is, at present, incapable of human-like judgment, the question at hand is what legal obligation, if

54 ibid 71.
55 Carolyn Sharp, 'Status of the Operator: Biologically Inspired Computing as Both a Weapon and an Effector of Laws of War Compliance' (on file with author).

any, stipulates that an autonomous weapon system could not comply with distinction? Recall Ney's 2019 remarks which stated:

> In the U.S. perspective, there is nothing intrinsically valuable about manually operating a weapon system as opposed to operating it with an autonomous function. For example, existing law of war treaties do not seek to enhance "human control" as such. Rather, these treaties seek, among other things, to ensure the use of weapons consistent with the fundamental principles of distinction and proportionality, and with the obligation to take feasible precautions for the protection of the civilian population. Although "human control" can be a useful means in implementing these principles, "human control" as such is not, and should not be, an end in itself.[56]

Other than the assertion that humans must be involved in any decision to select or engage targets–an assertion that has not been accepted by the international community as legally binding — there is no legal basis for arguing that autonomous systems cannot achieve compliance with the LOAC, including the principle of distinction.

With respect to cyber operations, cyber actors have used both indiscriminate[57] and very carefully tailored[58] tools in conducting cyber operations. As with all autonomous weapon systems, autonomous cyber tools would have to be able to apply force discriminately.

States that develop autonomous systems do not abrogate their legal duty to ensure that every weapon system employed by its armed forces complies with the LOAC. While not always strictly observed,[59] States comply with this requirement through a weapons review process[60] that has been well documented and discussed. This weapons review process

56 Ney (n 27).

57 'Statement from the Press Secretary' (*White House*, 15 February 2018) <https://www.whitehouse.gov/briefings-statements/statement-press-secretary-25/>.

58 David E Sanger, 'Obama Order Sped Up Wave of Cyberattacks Against Iran' (*NY Times*, 1 June 2012) <https://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html>.

59 Kathleen Lawand, 'A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977' (ICRC 2006) 5.

60 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 ('API') art 36; US Department of Defense, 'Directive 5000.01: The Defense Acquisition System' (12 May 2003, incorporating change 2, 31 August 2018) <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/500001p.pdf> ('DoD Directive 5000.01'); US Department of Defense, *Law of War Manual* (31 May 2016) s 6(2) <https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf?ver=2016-12-13-172036-190> .

includes an initial review as well as any necessary follow-up reviews for weapon systems that might change, adapt or 'learn' on the battlefield.[61] States can neither develop nor employ an autonomous weapon system, whether cyber or otherwise, that cannot apply precautions, including the principle of distinction.

3 *'Take all feasible precautions in the choice of means and methods' — Weaponeering*

Article 57(2)(a)(ii) requires States to 'take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life, injury to civilians and damage to civilian objects.' Accordingly, the weapons and tactics armed forces utilize in armed conflict, including potential autonomous weapons systems and those that utilize artificial intelligence or machine learning, must be capable of complying with this rule.

Though this is a significant and necessarily burdensome requirement that clearly affecting the research, development, and employment of weapons and tactics, it is important to note that these provisions equally apply to autonomous weapon systems, including autonomous cyber weapons. As Rain Liivoja points out, most LOAC rules are 'technology-neutral' or 'technology-indifferent', meaning that they need not change with every new technological development.[62] Echoing Liivoja, Marco Longobardo states:

> [T]he rules on the protection of civilians are the same regardless of whether hostilities are conducted with swords, bows, muskets, bombers, drones, or robots; simply, civilians must not be made the object of attacks, period. In this sense, most international humanitarian law rules are 'technology-indifferent', that is, they govern 'the conduct of hostilities and offer[] protection to persons not taking part in hostilities [] all quite irrespective of the means and methods of warfare the belligerents adopt and other technology that they use.[63]

61 See US Department of Defense, 'Directive 3000.09: Autonomy in Weapon Systems' (21 November 2012, incorporating change 1, 8 May 2017) <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>, for information on US weapons reviews on autonomous weapon systems.

62 Rain Liivoja, 'Technological Change and the Evolution of the Law of War' (2016) 97 International Review of the Red Cross 1157.

63 Marco Longobardo, 'Training and Education of Armed Forces in the Age of High-Tech Hostilities' in Elena Carpanelli and Nicole Lazzerini (eds), *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law* (Springer 2019) 77 .

As Longobardo argues, this requirement is technology-neutral and therefore poses no additional limitation on the use of autonomous weapon systems, whether cyber or non-cyber. Any weapons review process must account for this provision of the law and must ensure that autonomous weapons are capable of applying this rule. No additional legal requirement exists.

4 *'Refrain from deciding to launch an attack' — Proportionality*

Article 57(2)(a)(iii) is commonly known as the 'proportionality rule'. This provision is explicitly clear that applying the principle of proportionality (one of the general protections for civilians)[64] is a legal requirement for all those who plan or decide upon attacks.

Masahiro Kurosaki writes extensively on the application of proportionality to autonomous weapon systems. He argues that the principle of proportionality applies to autonomous weapon systems and would equally apply to 'computer-centered' systems.

> The principle of proportionality in modern LOAC has developed through the "reasonable military commander" standard. However, it is not intrinsically tied to, or at least not being limited to, the judgment of military commanders. It could be subject to adaptation in its application to a given circumstance by way of legal standards reflecting the sophisticated characteristics of fully AWS.[65]

As Kurosaki notes, there is no legal limitation on having an autonomous weapon system apply the principle of proportionality in selecting and engaging targets, assuming it could adequately apply the rule. In an interesting twist of argument, Kurosaki further asserts that the Martens Clause — a principle of law often used by opponents of autonomous weapon systems[66] — actually supports the use of autonomous weapons, particularly if it could limit the impacts on civilians.

> [I]t should be recalled that, as the Martens Clause enunciates, the humanitarian purpose of LOAC consists in protecting "the inhabitants and belligerents," no more and no less. The ICTY similarly opined that "[t]he basic obligation to spare civilians and civilian

64 API art 51(5)(b).
65 Kurosaki (n 4) 17–18.
66 Rob Sparrow, 'Ethics as a Source of Law: The Martens Clause and Autonomous Weapons' (*Humanitarian Law & Policy*, 15 August 2018) <https://blogs.icrc.org/law-and-policy/2017/11/14/ethics-source-law-martens-clause-autonomous-weapons/>.

> objects as much as possible must guide the attacking party when considering the proportionality of an attack."[67]

The content of the proportionality rule is not disputed with respect to autonomous weapons systems. Rather, the question is whether such systems can apply the rule. As mentioned above, it is unclear now what technological advancements might allow. What is clear is that any State intending to field an autonomous weapon system that selects and engages targets must meet the LOAC requirement of applying the rule of proportionality as part of the precautions in the attack.

In the context of the LOAC, cyber tools are rarely been used and there is no public record of fully autonomous cyber tools being used. However, the application of the principle of proportionality applies to both cyber tools utilized under the direct control of humans, as advocated in the *Tallinn Manual 2.0*[68] and by others,[69] and to autonomous cyber capabilities.

5 *'An Attack shall be canceled or suspended'*

Article 57(2)(b) requires the attacker cancel or suspend a planned attack when the proportionality calculus changes such that the attack would violate the LOAC. Certainly, there are some attacks that, once triggered, cannot be canceled or suspended (for example, the launching of a missile or the shooting of a field cannon). Rebecca Crootof writes:

> As CNAS notes, 'humans have been employing weapons where they lack perfect, real-time situational awareness of the target area since at least the invention of the catapult' and 'the essence of a projectile weapon, since the first time a human hurled a rock in anger, is the inability to suspend and abort the attack after launch.'[70]

Until the point that the attack is actually launched, the targeter must continue to apply the LOAC and cancel or suspend any attack that, due to a change in circumstances, becomes unlawful.

All autonomous weapon systems, including autonomous cyber systems, must have the capacity to cancel or suspend an attack based on either evidence provided externally or on evidence developed internally. The *Tallinn Manual 2.0* illustrates this point with the following example:

67 Kurosaki (n 4) 14.
68 *Tallinn Manual* (n 34) 481.
69 Jensen (n 33) 204–9.
70 Crootof (n 24) 61.

> For example, assume that a cyber attack is planned and all preparations are completed, including mapping the network and determining the nature of the target system. The attackers are awaiting authorization by the approving authority. Assume further that an operator is continuously monitoring the network. Any material changes in the cyber environment of the proposed target must be relayed to the commander and other relevant personnel as soon as possible.[71]

This is at least in part a design requirement that would be reviewed and tested as part of the weapons review process. While a legal requirement with which States must comply, there is nothing inherent in the construction of autonomous weapons that would prevent them from complying with this rule.

C ARTICLE 57(3) — 'WHEN A CHOICE IS POSSIBLE BETWEEN SEVERAL MILITARY OBJECTIVES'

The last provision of Article 57 that is likely to impact the deployment of autonomous weapon systems, including autonomous cyber systems, is Article 57(3), which states: 'When a choice is possible between several military objectives for obtaining a similar military advantage, the objective to be selected shall be that the attack on which may be expected to cause the least danger to civilian lives and to civilian objects.'[72]

Two aspects of this provision deserve consideration here. First, as this type of a decision would be one that inherently requires judgement, any autonomous weapon system would have to be capable of correctly making decisions that comply with law and policy. Or autonomous weapons, assessing this capability would likely need to be a part of the weapons review process for autonomous weapons.

Second, the legal requirements of this provision strongly argue for the presence of autonomous weapons systems on the battlefield, and the use of autonomous systems more generally. As Ashley Deeks states:

> One reason for the military's attraction to AI is that it can help manage doubt. Every day, especially on the urban battlefield, militaries confront questions about what they are seeing: is that

71 *Tallinn Manual* (n 34) 479.
72 API art 57(3).

> person holding a video camera or a rocket launcher? Why is there very little pedestrian traffic in the market today? Is the person I just detained likely to endanger our forces if released? Will a strike on that warehouse using a joint direct attack munition produce excessive collateral damage? Each of these questions requires decision-making in the face of uncertainty. AI tools can help categorize objects, identify anomalies, and make predictions up to a particular confidence level. These algorithms will be especially useful if they produce recommendations that are sensitive to the precise questions that LOAC requires militaries to answer.[73]

Autonomous systems are systems that utilize ongoing machine learning and artificial intelligence. Therefore, the ability of such systems to accurately assess data concerning a wide variety of battlefield questions will continually increase. The interconnection of sensors, data processors, and algorithmic assessments will assuredly enhance the battlefield commander's ability to gather, assess, and exploit intelligence.

The same holds true when assisting commanders in surveying which targets are the least dangerous to civilians. Furthermore, the structural survivability of non-cyber autonomous systems[74] increases the ability to loiter and gather intelligence — thereby allowing for more comprehensive and thoughtful determinations about selecting and engaging targets. As Charles Trumbull states:

> Advances in robotics and AI will lead to weapons with far greater endurance than humans. Machines "do not get tired, frightened, bored, or angry." They do not suffer the effects of post-traumatic stress disorder or seek revenge after witnessing their fellow soldiers killed in action. Accordingly, autonomous weapons are not susceptible to the human frailties that often lead to war crimes.[75]

73 Ashley Deeks, 'Coding the Law of Armed Conflict: First Steps' in Matthew C Waxman and Thomas W Oakley (eds), *The Future Law of Armed Conflict* (Oxford University Press forthcoming).

74 See Andrew Feickert, Jennifer K Elsea, Lawrence Kapp, and Laurie A Harris, 'US Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress' (Congressional Research Service Report R45392, 20 November 2018) 34 <https://crsreports.congress.gov/product/pdf/R/R45392/4>, where the authors state: '[P]roponents of such systems argue that human emotions — fear, anger, and the instinct for self-preservation — may lead to adverse consequences on the battlefield. Robots, they posit, may not be subject to human errors or unlawful behavior induced by human emotions.'

75 Charles P Trumbull IV, 'Autonomous Weapons: How Existing Law Can Regulate Future Weapons' (2020) 34 Emory International Law Review 533, 545–6. See also Kurosaki (n 4) 18–19 where the author argues: 'LOAC cannot go so far as to strictly demand human soldiers to protect civilians at the sacrifice of their own lives. Machines, however, may be exposed to the risk of destruction, hereby creating more opportunities for saving innocent civilians.'

To the extent that autonomous weapons live up to these expectations, they may prove to be a significant aid in complying with Article 57(3).

D CONCLUSION

Because there is currently no consensus on the required level of human control in the research, development, and employment of autonomous systems, the general standard of weapons review remains sufficient, as long as there is a methodology for continued data collection and compliance.

IV
CONCLUSION

As the analysis above has indicates, the requirement to take precautions in attack do not present unassailable legal impediments to the research, development, or employment of autonomous weapon systems, including autonomous cyber systems, provided such systems are subject to a rigorous weapons review. Furthermore, because Article 57 of Additional Protocol I applies, without prejudice, to all who plan or decide to attack, autonomous weapons remain within the confines of the LOAC requirements. Therefore, with rigorous weapons review processes in place that continually examines the autonomous system's continued 'learning' and absent any legal preclusion to compliant systems, proposed autonomous weapons bans are unlikely to be successful — especially considering the present success of autonomous weapons already in use.

Chapter 10

Reviewing Autonomous Cyber Capabilities

Alec Tattersall and Damian Copeland[1]

I
INTRODUCTION

The law of armed conflict ('LOAC') rests on a premise that, '[i]n any armed conflict, the right of the Parties to the conflict to choose methods or means of warfare is not unlimited'.[2] A key obligation giving effect to this premise is the duty of States Parties to Additional Protocol I to the Geneva Conventions ('AP I') to determine whether the employment of a new weapon, means or method of warfare would comply with the protocol and

1 This chapter reflects the personal views of the authors. The opinions and conclusions offered do not necessarily reflect official positions or views of the Australian Defence Force or the Government of Australia.

2 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 ('AP I'). See also Jean de Preux, 'Protocol I — Article 35' in Yves Sandoz, Christophe Swinarski and Bruno Zimmerman (eds), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (ICRC 1987) ('*ICRC Commentary*') [1385].

other applicable international law.[3] States not Party to AP I are arguably under a more general obligation to ensure the compliance of new weapons and means of warfare with LOAC and other relevant rules of international law.[4] We refer to the relevant processes as 'weapons reviews', which we prefer over the alternative term 'legal reviews' because they involves matters of both legal obligation and State policy.

Autonomous cyber capabilities ('ACC') encompass software (and, where relevant, the necessary accompanying hardware) that operates on or against computers or computer networks by algorithmically executing actions within pre-determined (albeit potentially broad) parameters, without human intervention. Where the ACC are designed or expected to cause damage or destruction in armed conflict, they create novel capabilities that will challenge traditional weapons review practices. This chapter will examine the constituent elements of a weapons review and address some potential tensions that the unique nature of ACC may cause for conventional weapons review practice.

A WHAT IS THE WEAPONS REVIEW REQUIREMENT?

For the 174 States Parties to AP I, the weapons review obligation is articulated in Article 36 as follows:

> In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

Article 36 outlines a requirement to review the legality of the use of 'a new weapon, means or method of warfare' and the phrasing 'study, development, acquisition or adoption' provides the triggers (temporal signposts) for this obligation. While the express language of Article 36 provides States Parties with insight into the scope of their mandatory review obligation, it does not specify how each of the review elements are to be interpreted, the relevant standards to be applied, or designate any

3 AP I art 36.
4 See, eg, Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017) ('*Tallinn Manual 2.0*') 464–7, rule 110.

specific form of review methodology. An enduring challenge is that neither AP I, nor LOAC more broadly, define the terms 'weapons, means and methods of warfare'.[5] Furthermore, there appears to be no international consensus on defining these terms.[6] Acknowledging this unsettled definitional position, a 'weapon' can generally be described as an instrument designed or used to cause damage or harm;[7] a 'means of warfare'[8] recognised broadly as being the weapons, 'platforms and equipment which make possible an attack';[9] and a 'method of warfare' as referring to the manner a means of warfare (usually a weapon) is used, and is commonly articulated as being the 'tactics, techniques and procedures ('TTPs') for carrying out military operations'.[10]

Consistent with the international law principle of sovereignty,[11] and absent specific provision in international law, reviewing States are permitted to determine if an Article 36 review obligation has arisen,[12] and

5 'Weapon' in *Weapons Law Encyclopedia* (Geneva Academy of International Humanitarian Law and Human Rights, 20 November 2013) <http://www.weaponslaw.org/glossary/weapon>.

6 That being said, an inherent characteristic for most definitions is usually the ability to causally (even indirectly) effect objects or persons. Many States define a weapon and/or means based on it being an instrument designed to cause injury, death or property destruction — no requirement is specified that the means cause the effect directly or indirectly. See, eg, Rain Liivoja and Luke Chircop, 'Are Enhanced Warfighters Weapons, Means, or Methods of Warfare?' (2018) 94 International Law Studies 161, 176, suggesting that that something 'will … constitute a weapon when used as an instrument to cause injury, death, damage, or destruction'.

7 Program on Humanitarian Policy and Conflict Research at Harvard University, *HPCR Manual on International Law Applicable to Air and Missile Warfare* (Cambridge University Press 2013) ('*HPCR Manual*') rule 1(ff) defines a weapon as, 'a means of warfare used in combat operations, including a gun, missile, bomb or other munitions, that is capable of causing either (i) injury to, or death of, persons; or (ii) damage to, or destruction of, objects.' Efforts to define weapons (or means) have yielded approaches ranging from characterising weapons by offensive capability, intentionality/design purpose, nature, or deterministic characteristic. For offensive capability, see Justin McClelland, 'The Legal Review of Weapons under Article 36 of Additional Protocol 1' (2003) 850 International Law Review of the Red Cross 397, 404; William Boothby, 'Conflict Law: The Influence of New Weapons Technology, Human Rights and Emerging Actors' (TMC Asser Press 2014) 169; but see also Liivoja and Chircop (n 6) 175, arguing that it is unnecessary to expand the definition to include reference to offensive or defensive capability. For intentionality, see US Department of the Army, Army Regulation 27–53: Review of Legality of Weapons under International Law (1 January 1979; major revision, 23 September 2019). For nature, see William Boothby, *Weapons and the Law of Armed Conflict* (Oxford University Press 2009) 169. For deterministic characteristic, see Jeffrey T Biller and Michael N Schmitt, 'Classification of Cyber Capabilities and Operations as Weapons, Means or Methods of Warfare'' (2019) 95 International Law Studies 179.

8 *HPCR Manual* (n 7) rule 1(t); see also Claude Pilloud and Jean Pictet, 'Protocol I — Article 51' in *ICRC Commentary* (n 2) [1957]: 'The term "means of combat" or "means of warfare" generally refers to the weapons being used'.

9 *HPCR Manual* rules 1(t) and (ff); see also Pilloud and Pictet (n 8) [1957]: '"methods of combat" generally refers to the way in which … weapons are used'.

10 *HPCR Manual* (n 7) rule 1(v); *Tallinn Manual 2.0* (n 4) rule 103(b). Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law* (Cambridge University Press 2005) vol 2, rules 15–19. While we believe that Biller and Schmitt (n 7) 203 correctly identify that a method of warfare does not require a weapon or means, we respectfully suggest that instances where a means is not involved (including the threat of the use of a means) are more limited than what Biller and Schmitt suggest.

11 Helmut Steinberger, 'Sovereignty' in *Max Planck Institute for Comparative Public Law and International Law* (Oxford University Press 1987) vol 10, 414.

12 W Hays Parks, 'Conventional Weapons and Weapons Reviews' (2005) 8 Yearbook of International Humanitarian Law 55, 113.

how they fulfil it.[13] Consequently, each State is required to develop its own definitions and determine whether a cyber capability or activity falls within their definition of what constitutes a weapon, means or method of warfare. Ultimately, a capability that falls outside a State's definition is not required to be legally reviewed as a matter of law.[14] Currently only a limited number of States that undertake Article 36 reviews have publicly acknowledged that their review obligation encapsulates cyber capabilities.[15]

The review obligation is less clear for States not Party to AP I. While few claim the Article 36 review obligation has been accepted as part of customary international law, there is academic support, and recognition by some international and domestic courts,[16] for the view that a narrow customary international law obligation exists.[17] This obligation has been described as requiring the good faith review of means of warfare before they are fielded in armed conflict,[18] or as the slightly broader requirement to ensure compliance with LOAC for means of warfare that are acquired or used.[19] The extent of this obligation would be limited to reviewing means

13 Vienna Convention on the Law of Treaties (adopted 23 May 1969, entered into force on 27 January 1980) 1155 UNTS 331, art 31(1) provides a general rule of interpretation: 'A treaty shall be interpreted in good faith in accordance with the ordinary meaning to be given to the terms of the treaty in their context and in light of its object and purpose'.

14 Biller and Schmitt (n 7) 195.

15 The exact number of States undertaking art 36 reviews is unclear; the number of that review cyber capabilities more so. The United Kingdom (UK), Netherlands, Norway, Sweden, Switzerland, Belgium, Canada, Germany, New Zealand, Australia, Austria, Denmark, Israel, and France have all publicly acknowledged undertaking art 36 reviews. The following States have claimed in the Group of Governmental Experts on Lethal Autonomous Weapons that they undertake reviews: China, Russia, Japan, South Korea. The UK identified that they review cyber capabilities: UK Ministry of Defence, Development Concepts and Doctrine Centre, 'UK Weapons Reviews' (8 March 2016) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/507319/20160308-UK_weapon_reviews.pdf>. Australia identified that they consider cyber capabilities to be reviewable: Australia, 'The Australian Article 36 Review Process' (Group of Governmental Experts on Lethal Autonomous Weapons Systems, 27–31 August 2018) CCW/GGE.2/2018/WP.6, fn 6.

16 *Legality of the Threat or Use of Nuclear Weapons* (Advisory Opinion) [1996] ICJ Rep 226 ('*Nuclear Weapons Advisory Opinion*'); *Shimoda v State of Japan* (1963) 355 Hanrei Jiho 17, translated in (1964) 8 Japanese Annual of International Law 212, 242.

17 See also Natalia Jevglevskaja, 'Weapons Review Obligation under Customary International Law' (2018) 94 International Law Studies 186, referring to the undertaking to respect and ensure respect for the LOAC, the obligation for legal advisers to advise on the applicability of the LOAC, the duty to instruct armed forces on the compatibility with the LOAC, and the prohibition to employ weapons causing superfluous injury or unnecessary suffering and indiscriminate weapons. See generally Hague Convention (II) with Respect to the Laws and Customs of War on Land (adopted 29 July 1899, entered into force 4 September 1900) 189 CTS 429 (1899 Hague Convention II) art 1 and annexed Regulations art 23(e); Hague Convention (IV) with Respect to the Laws and Customs of War on Land (adopted 18 October 1907, entered into force 26 January 1910) 205 CTS 277 (1907 Hague Convention IV) art 1 and annexed Regulations art 23(e).

18 Biller and Schmitt (n 7) 186; *HPCR Manual* (n 7). See also Parks (n 12) 57: 'Under the international law maxim *pacta sunt servanda* states have a general duty to engage in good faith performance of their treaty obligations.' See further ibid and 106–7.

19 *Tallinn Manual 2.0* (n 4) 464, rule 110, where the authors describe a customary international law requirement to '[e]nsure that the cyber means of warfare that they acquire or use comply with the rules of the law of armed conflict that bind them'. Such obligations are argued as being a consequence of the secondary application of rules of LOAC, in particular 1899 Hague Convention II art 1; 1907 Hague Convention IV art 1; Geneva Convention (I) for the Amelioration of the

of warfare with respect to the customary international law prohibitions on weapons causing superfluous injury or unnecessary suffering, and indiscriminacy.[20] Of the States not party to AP I, only the United States have indicated that they will review cyber capabilities.[21]

B WHAT IS AN ACC?

An ACC is in simple terms a cyber capability that can execute tasks without human interaction.[22] The two key elements of an ACC under this definition — both primarily code based — are cyber capability and autonomous functionality.

A cyber capability is software (potentially combined with hardware and a human operator) that operates by digital communication — either by communicating with another cyber device, or through infiltrating software into another cyber device and then using that software to communicate (internally or externally) from that location.[23]

Cyber capabilities that function by communicating with another cyber device are akin to military capabilities that influence or temporally restrict functionality.[24] These capabilities are usually restricted to affecting 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, art 1; Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 85; Geneva Convention (III) relative to the Treatment of Prisoners of War (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 135, art 1; Convention (IV) relative to the Protection of Civilian Persons in Time of War (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 287, art 1; and, AP I, arts 35(2), 51(4) 80(2) and 82. In accordance with common art 1 of the four Geneva Conventions, State Parties are legally bound to respect and ensure respect for the Geneva Conventions in all circumstances, including a good faith performance of the legal obligations contained with the Geneva Conventions, and their Additional Protocols, and accordingly, the obligation provided in AP I art 36.

20 *Tallinn Manual 2.0* (n 4) 241; Jevglevskaja (n 17) 214.

21 Even though in the US armed forces weapons reviews are a matter for each service branch, the US Department of Defense, *Law of War Manual* (June 2015) 1008 confirms that the US will review cyber capabilities where they are weapons or part of weapons. Israel also undertakes weapons reviews, but it is not known whether they include review of cyber capabilities.

22 Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) 10 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>, suggesting that autonomous operation 'in its simplest sense to refer to the ability of a system to perform some task without requiring real-time interaction with a human operator'. Acknowledging there is an unresolved debate on what autonomy is, we have adopted a broad understanding of autonomy — a system or thing (including software) that provides capability to perform tasks without (real-time) human interaction.

23 These communications are not like human communication. Rather they are the 'transmission of data in coded form'. See Robert Elliott Smith, *Rage Inside the Machine: The Prejudice of Algorithms, and How to Stop the Internet Making Bigots of Us All* (Bloomsbury 2019).

24 Influencing a computer function refers to information operations that provide inaccurate data to the computer akin to traditional information operations. Restricting computer function refers to cyber capabilities that overwhelm a targeted computer with communications so that it cannot communicate with other computers, akin to a radio frequency jammer — a military tool that restricts the functioning of a targeted device by limiting the ability of that device to communicate or receive communication.

functionality of a targeted device for a limited period of time.[25] Cyber capabilities that function by infiltrating another cyber device or network operate as 'control' devices in that they take control of a component of the targeted device or network and use this control to communicate instructions to the target device or connected systems (including physical systems) to achieve an effect.[26] The key aspect of these types of capabilities is that the instructions or communications come from the software operating as an individual entity; it is separate to the computer it was released from.[27]

Often cyber capabilities will involve a combination of both elements — communications to a targeted device, and embedded software communicating from within. The first element is primarily for access and the second is primarily to achieve the intended effect(s).[28] Either way, the consequences of such interactions come from instructions that are issued to the targeted device, which result in the effect(s). This interaction principally relies on the targeted device reacting to the instructions with some form of resultant effect(s). Cyber capabilities therefore potentially possess qualities of means (causing damage or destruction) and methods (manner of using means) of warfare.

Cyber capabilities contain an inherent tendency towards autonomous functionality. This is because autonomous functionality reflects the ability to execute tasks absent 'real-time interaction with a human operator'[29] and it is the nature of most cyber capabilities that they operate without direct human manipulation. The autonomous functionality in an ACC is achieved through algorithms, a set of coded instructions that compute instructions for the ACC to implement. Where the algorithms are used to control cyber tasks undertaking combat functionality in an armed conflict the executed tasks are governed by LOAC.[30] Such autonomous function-

25 Key types of cyber capabilities that utilise communications to (or from) the target include denial of service (works by overwhelming a system's resources so either communications can't get in or out, or the device cannot respond to communications), phishing (sending a communication, which appears to be from a trusted source, with the intent that the targeted device or user of the targeted device will release information or respond in a certain way), man-in-the-middle (inserting operator or cyber capability between communications of a targeted device and a device it communicates with), or influence operations.

26 See *Tallinn Manual 2.0* (n 4) 451–2 regarding control for the purposes of damaging the controlled device versus using the controlled device to damage another military objective.

27 Key types of cyber capabilities that communicate from within a targeted device include viruses, trojans, worms, access blockers, or erasers.

28 BlackEnergy was a botnet capability that worked in the opposite way — it enslaved computers (infecting them with software that takes control of the computer) to use in a distributed denial of service attack (DDoS: the enslaved computers communicated — making false requests for information — to a target computer).

29 Liivoja, Naagel and Väljataga (n 22) 10.

30 Tim McFarland, 'Autonomous Weapons and Human Control' (*Humanitarian Law & Policy*, 18 July 2018) <https://blogs.icrc.org/law-and-policy/2018/07/18/autonomous-weapons-and-human-control/>.

ality is therefore claimed to be a method of warfare, typically reviewed as an integral part of the review of a weapon or means of warfare.[31]

C REVIEWABILITY OF ACC

At the present time, the classification of cyber capabilities (used to undertake attacks in armed conflict) is disputed. Legitimate arguments exist for classifying such capabilities as weapons or means,[32] or only as methods.[33] It is not the intent of the authors to resolve this dispute. As such, for the purposes of this Chapter, ACC designed or expected to cause damage or harm in armed conflict can be a weapon or means of warfare that necessitates a full weapons review.[34]

II
WEAPONS REVIEW COMPONENTS

While States are permitted to determine how they fulfil their weapons review obligation, we would suggest a thorough approach would contain the following Parts:

31 There is a subtle difference between the element constituting the cyber capability and the element that provides autonomy. That is the cyber capability provides instructions to a vector or target device of the target state, whereas the autonomy element provides instructions to a weapon capability of the attacking state.

32 States and academics have defined cyber capabilities as 'cyber weapons'. See, for instance, Thomas Rid and Peter McBurney, 'Cyber Weapons' (2012) 157(1) RUSI Journal 6, 7. Rid later noted concerns with defining cyber weapons in Thomas Rid, *Cyber War Will Not Take Place* (2nd ed, Hurst 2017) foreword; *Tallinn Manual 2.0* (n 4) 452, rule 103; Tom Uren, Bart Hogeveen and Fergus Hanson, 'Defining Offensive Cyber Capabilities' (Australian Strategic Policy Institute, 4 July 2018) <https://www.aspi.org.au/report/defining-offensive-cyber-capabilities>. The common feature in these definitional variations of cyber weapons is that they are designed to cause damage or destruction.

33 Biller and Schmitt (n 7) 211–212, for instance, argue that for a cyber capability to be characterised as a means of warfare it must possess the determinative characteristic common to traditional means of warfare — in this instance direct causation of harm. Accordingly, as cyber capabilities do not have a direct harming mechanism, but rather rely on the targeted cyber device responding to instructions given to it, they are not means of warfare. Interestingly, a number of States that appear to have characterised cyber capabilities as being methods of warfare have also recognised that such capabilities should be subject of a full weapons review: see US Department of Defense (n 21) and UK Ministry of Defence (n 15).

34 While AP I art 36 requires weapons review reviewing methods of warfare, the extent of such reviews (when not conducted in conjunction with the review of a weapon) is usually limited to a generalised TTPs review of categories of cyber operations, and not a review of each individual cyber capability. For non-AP I States it means that there is no legal requirement to undertake a review, as the customary requirement — if it exists — only extends to means of warfare.

Part 1 Analysis of the design, technical and performance
Part 2 Determination of 'normal or expected use'
Part 3 Determination that the capability is a 'new weapon, means or method' that requires review under Article 36
Part 4 Specific law
Part 5 General law
    Unnecessary suffering
    Indiscriminacy
    Proportionality (where relevant)
    Environmental damage
    Other LOAC considerations relevant to the weapon, means or method (as required)
Part 6 Other relevant international law
Part 7 Public interest and social conscience — the Martens Clause
Part 8 Domestic law (if necessary)
Part 9 Conclusion

Below is a brief discussion of each Part, and specific issues for ACC.

PART 1 ANALYSIS OF DESIGN, TECHNICAL AND PERFORMANCE CHARACTERISTICS

Requirement. Consistent with the International Committee of the Red Cross ('ICRC') guidance,[35] Part 1 of the weapons review commences with a detailed description (normally provided by relevant experts) of the new weapon, means or method of warfare. This could include technical guidance (design, manufacturing process, material composition, fusing system, guidance system, integrated safety procedures and safeguards), ballistics information (speed, accuracy, damage mechanism, delivery mechanism, effects etc.), analysis and assessments of weapons effects, and appropriate subject matter expert advice on the design and technical characteristics of the weapon or means.

This is followed by detailing its performance characteristics, including an analysis of how it 'operates',[36] and any relevant health and environmental considerations. The ICRC suggests that 'relevant factors would

35 International Committee of the Red Cross (ICRC), 'A Guide to the Legal Review of New Weapons, Means and Methods of Warfare Measures to Implement Article 36 of Additional Protocol I of 1977' (January 2006) 18.
36 Ibid.

include: the accuracy and reliability of the targeting mechanism (including e.g. failure rates, sensitivity of unexploded ordnance, etc.); the area covered by the weapon; whether the weapons' foreseeable effects are capable of being limited to the target or of being controlled in time or space (including the degree to which a weapon will present a risk to the civilian population after its military purpose is served).'[37]

Analysis. For an ACC, this part would need to describe the relevant components (such as harming mechanism, propagation method, and nature of the exploit), and any autonomous functionality (decision-implementation component or 'autonomy algorithms'). A particular issue in reviewing a cyber capability is going to be separating the capability from any operation it is being used for. Cyber operations generally comprise of the following steps: target identification, target reconnaissance, target engagement (or penetration for access), internal target reconnaissance, target establishment (for permanent presence), target exploitation, target effect (harm in the form of damage or destruction), and target extraction.[38] Some or all of these steps could be undertaken by an ACC, or provided by separate but linked cyber capabilities. For instance, the components of a Stuxnet-style ACC, if considered by Part 1 of a weapons review, would include the design and operation of the infection/ propagation technique (e.g. the worm), the algorithm that triggers the payload when specific configuration requirements are found (e.g. the LNK file) and the payload/ harming implement (e.g. the rootkit). The analysis would also include the technical characteristics and specifications of the intended target (e.g. the supervisory control and data acquisition system) to ensure a complete understanding of how the ACC would access and achieve its affect. This analysis then informs the subsequent parts of the weapons review.

A weapons review, however, will only consider those steps (performed by the ACC) which are either an integral part of the attack, or an integral part of the attacking capability,[39] and then only with respect to the ACC's algorithms that undertake or enable combat functionality. The two elements of combat functionality are 'combat' (the use of violence by

37 Ibid.

38 Gary Brown, 'Spying and Fighting in Cyberspace: What is Which?' (2016) 8 Journal of National Security Law & Policy 621, 631-634. It is widely recognised that many of these steps, bar the target effect step, could be undertaken as part of espionage or information gathering exercises.

39 This is actually no different to what is required for the conduct of traditional kinetic operations where the steps can be undertaken by a combination of devices (where intelligence is gathered using civilian human intelligence and surveillance devices, military components extract target information and develop the target, and the attack is undertaken by a separate means or method), or conversely the entire attack may be undertaken by one device (i.e. a fighter aircraft).

armed forces) and 'functionality' (the purpose/task that the instrument is designed or expected to undertake).[40] The combination of these elements results in combat functionality representing the range of actions or functions that a weapon is capable of undertaking to apply violence. Violence in this sense ranges from actions preparatory to a use of force (such as authorising, searching, detecting, tracking, identifying, selecting, cueing, prioritising, determining fire control) through to those using the harming mechanism during use of force actions (applying kinetic or non-kinetic force where violence is intended or expected to neutralise, damage, destroy, detain, injure, or kill).[41]

Determining combat functionality is not always simple, as the utilised code may be able to create different effects — some damaging, some not — as determined by the nature of the target computer or network.[42] This position can be further complicated (and simplified) where aspects of a cyber operation are automated, but only to the extent that the autonomous element undertakes or enables combat functionality. Like with cyber more generally, the complexity comes with determining which code providing autonomy is also relevant to combat functionality. Identifying those components that are part of the ACC and integral to causing damage or harm is essential to completing Part 1.

This is especially the case for the ACC's 'autonomy algorithms', which is relevant to Part 1 for its unusual but significant:

- design features — including the code or algorithm specifications (such as the programming language, incorporation of performance standards *vice* legal standards), modes or levels of autonomy that can be selected, the human interface, and integration of the algorithms (and human-machine interaction) into the combat functionality of the ACC;

- technical characteristics — such as how the algorithms work (the mathematical logic to permit adaptation or learning with algorithmic parameters), how humans interface with it, inbuilt

40 *Johnson v United States*, 170 F 2d 767 (9th Cir 1948).
41 This is not an exhaustive list, nor indeed are these terms the only way of describing the functions listed.
42 For instance, temporary encryption is used in ransomware with the threat of permanent encryption, the latter being akin to a complete loss/destruction of data (target effect). In reality, there may be no difference between intelligence for exfiltration versus information for creating an effect on the target (target exploitation). Instructions to remove all record of the cyber intrusion by deleting the code using an eraser may with limited re-direction be used to remove data from the host computer or network (target extraction).

control mechanisms, and how it will interact with its environment (such as gathering information); and

- performance characteristics — addressing the capability of the algorithms to implement engagement and other decisions against target devices or networks, level and nature of autonomy, digital and procedural safeguards, human interaction and overrides, and relevant standards (legal, reliability and performance).

PART 2 DETERMINATION OF 'NORMAL OR EXPECTED USE'

Requirement. Article 36 requires weapons reviews to encompass the use of a weapon or means 'in some or all circumstances'. A plain reading of this phrase suggests an onerous responsibility on State Parties to consider 'all circumstances' in which the use may be unlawful.[43] The ICRC, academic discourse and States that undertake weapons reviews, primarily appear to have adopted a 'normal or anticipated' or 'normal or expected use' interpretation.[44] This interpretation is a practical recognition that while there are potentially innumerable uses, there is no requirement to foresee or analyse all possible uses (including unlawful misuses) or effects of a weapon, means or method.[45] Of course, a means or method may have multiple roles that fall within the 'normal or expected use' but that list is finite.[46] Part 2 of a weapons review should therefore contain a succinct description of the 'normal or expected use'[47] of the weapon or

43 Parks (n 12) 119. See also ICRC (n 35) 4, stating that the purpose of art 36 is to 'prevent the use of weapons that would violate international law in all circumstances and to impose restrictions on the use of weapons that would violate international law in some circumstances'.

44 See Jean de Preux, 'Protocol I — Article 36' in *ICRC Commentary* (n 2) [1469]: 'The determination is to be made on the basis of normal use of the weapon as anticipated at the time of evaluation'; Report of Committee III, Second Session, Geneva, 3 February–18 April 1975, CDDH/215/Rev.1, in *Official Records of the Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts* (Federal Political Department of Switzerland 1978) vol xv, 259, [31]: '[T]he article is intended to require States to analyse whether the employment of a weapon for its normal or expected use would be prohibited under some or all circumstances. A State is not required to foresee or analyse all possible misuses of a weapon, for almost any weapon can be misused in ways that would be prohibited.' But see JD Fry 'Contextualised Legal Reviews for Means and Methods of Warfare: Cave Combat and IHL' (2006) 44 *Columbia Journal of Transnational Law* 453, 471, suggesting that the traditional 'normal or expected' use approach is too limited i.e. it requires greater consideration of context.

45 Nor is a review required if there is no intention to use a weapon, means or method in armed conflict. As such, domestic use of an ACC, in law enforcement for example, does not require a review to be conducted.

46 This does not mean that a State cannot adopt a broad analysis of a weapon or munition beyond what a manufacturer describes as the weapon's use.

47 ICRC (n 35) 18. This covers 'the use for which the weapon is designed or intended, including the types of targets (e.g. personnel or materiel; specific target or area; etc.); and its means of destruction, damage or injury.'

means,[48] and — particularly where autonomy is present — the manner in which that use occurs (method).[49]

Analysis. In practice, during the development of a conventional weapon, the reviewing State will describe its 'normal or expected use' with increasing fidelity. This would commence with the identified task, effect or solution for an identified gap in a State's military capability driving the procurement or development process,[50] and refined through the capabilities life cycle (documentation, training, certification, and authorisation).[51] In contrast, the somewhat atypical lifecycle of cyber capabilities when combined with autonomous functionality means that the 'normal or expected use' of an ACC would need to be identified closer to capability finalisation.[52] There are a number of factors requiring careful application when determining the 'normal or expected use' of an ACC.

1 *Inclusion of Autonomy*

The addition of an autonomy element may complicate identifying and defining the 'normal or expected' use in a couple of significant ways. First, a change in context may result in a change in the 'normal or expected' use. For instance, 'normal or expected use' is usually assessed in terms of both the intended effect of the capability and the method by which it achieves that effect. The use of autonomy may not change the

48 ibid 17: 'In assessing the legality of a particular weapon, the reviewing authority must examine not only the weapon's design and characteristics (the 'means' of warfare) but also how it is to be used (the 'method' of warfare), bearing in mind that the weapon's effects will result from a combination of its design and the manner in which it is to be used'. See also Kathleen Lawand, 'Reviewing the Legality of New Weapons, Means and Methods of Warfare' (2006) 88(864) International Review of the Red Cross 925, 927–8 who, commenting on the Guide, notes that 'a new weapon — that is, a proposed means of warfare, cannot be examined in isolation from the way in which it is to be used — that is, without also taking into account the method of warfare associated with it'.

49 ICRC (n 35) 10, noting that '[a] weapon or means of warfare cannot be assessed in isolation from the method of warfare by which it is to be used. It follows that the legality of a weapon does not depend solely on its design or intended purpose, but also on the manner in which it is expected to be used on the battlefield.' See also Jean de Preux, 'Protocol I — Article 35' in *ICRC Commentary* (n 2) [1402], emphasising that 'the words "methods and means" include weapons in the widest sense, as well as the way in which they are used'. For non-AP I States the review obligation is limited to the means of warfare.

50 Generally, a capability requirement for a conventional weapon will be described in weapon design and performance specifications that are either the subject of a commercial tender process for the purchase of existing military capabilities or the basis for the development of a new weapon. The weapon chosen or developed for acquisition will be capable of fulfilling the State's capability gap and the manner in which it does so will be weapon's 'normal or expected use'.

51 Following the procurement of the weapon, the normal or expected use will be described in the weapons use and training manuals. Weapon operators and military commanders are taught the technical characteristics of the weapon, their functional operation and intended use.

52 A weapons review of an ACC designed to destroy computer files in an adversary's IT system — e.g. similar to the 2012 Shamoon attack against Saudi Aramco, an oil company — would require careful analysis of the propagation mechanism to ensure that its effects were capable of being restricted to the intended military objective. The analysis would include consideration of the target's IT system architecture to determine whether its specifications are sufficiently unique to ensure connected, civilian IT systems are not affected by the attack. If the propagation mechanism was unable to analysed the Weapon Review may recommend restrictions on the ACC use to address risks of indiscriminate effects.

intended effect, but it may materially alter the method with which the intended effect is achieved in the context of its use. Second, a change in the 'normal or expected' use may also result from autonomy driven by code that has adaptive capacity (i.e. learning, modification, or optimisation). Consideration would need to be given to variables such as:

- the 'nature' of the algorithms, such as the level and complexity of code, permitted code adaptability, and decision-implementation parameters;
- the level and use of reliability, performance and legal standards applied to a decision-implementation capability; and
- the 'nurture' of the algorithms from data diet, training regimes, through to human interaction and application of internal and external controls.

2 *Design — 'Normally' a Single Use Bespoke Capability*

Unlike most conventional weapons, the digital nature of cyber capabilities and elevated security surrounding their use means that they are generally developed internally by States for a specific task or operation, rather than commercially acquired. This has several important consequences. The bespoke nature of the capability will limit the scope of the weapons review to analysing the legal issues relating to the design purpose or performance of the specific task.[53] It will also narrow the parameters of any included autonomy element in line with the specific tasking. Unfortunately, such a capability is unlikely to return or report home on task completion and thus unable to supply sufficient operational performance data — this will have to come from testing and training.

3 *Identifying the Reviewable 'Harming Mechanism'*

What is 'normal' is generally identified by reference to design purpose or intended effect. The design purpose is recognised as an important element in the identification of a weapon or means as it excludes those objects which are capable of causing injury, death or destruction, such as a truck, but which are not designed for that purpose.[54] The design pur-

53 This will expedite the weapons review process, as the analysis will focus only on the legal consequences of achieving the specific design purpose. If the capability were to be re-rolled for a different operational context again, or reused, a further review would likely be required.

54 Michael Schmitt, 'Cyber Operations and the *Jus in Bello*: Key Issues' (2011) 87 International Law Studies 89; Charles Dunlap, 'Perspective for Cyber Strategists on Law for Cyberwar' (2011) Strategic Studies Quarterly 81, 85.

pose is given effect by the harming mechanism. In conventional means, it is normally 'a relatively straightforward process' to identify a harming mechanism; for example, the blade of the knife is its harming mechanism.[55] This is not always going to be as clear for an ACC, either because the harming mechanism is not obvious, or because it relies on the target device to achieve effects.[56] The opaque nature of ACC results partly from the separate elements of cyber (including propagation, exploit and payload, and their integration) and autonomy, as well as the way in which they combine to achieve their design purpose.

First, some effects, such as data destruction, might not be recognised as harm or be the ultimate design purpose.[57] Cyber capabilities, such as malware, that operate by infiltrating a target device or network and taking control of an aspect of that target, cause direct harm by physically altering the state of elements of a targeted device through the transfer of energy.[58] This type or level of harm — whether caused through the employment of an exploit, or by taking control of an aspect of the target and permitting communications directing the host to harm itself or others — is often not recognised as sufficient harm for consideration, or is not the ultimate design purpose or intended effect. A question arises, then, regarding the exploit and the code that permits taking control: is it a means in its own right or is it integral to the ultimate design purpose?[59]

Second, cyber capabilities could have multiple potential 'harming mechanisms'. There are two common positional variants of harming mechanisms in cyber capabilities — those linked to the targeted device, and those relating to the exploit or the payload in the code of the cyber capability. The harming mechanism for most cyber capabilities comes from controlling an aspect of a targeted device (a computer or a network) and using such control to issue commands to the targeted device that result in 'effects'. The potential harming mechanism options viewed from the perspective of the targeted device will be determined by the nature, location and interconnectedness of the targeted device. Not only may they be many and varied, they may also be unknown at the time of the review.

55 McClelland (n 7) 404.
56 Identifying the harming mechanism on conventional weapons is perhaps more important that identifying the harming mechanism in cyber capabilities. For conventional weapons there is an important body of Weapons Law that regulates weapons and means based on the harming mechanism used.
57 US Department of Defense (n 21) 1004.
58 While such alteration may be on an atomic scale, it does not mean it does not occur.
59 A key issue here is recognising the division between the cyber capability (which exercises an element of control of the target) and the instructions it issues dependent upon that control. The ultimate design purpose or intended effect comes from issuing communications and not infiltrating and taking control of the target device. In this case the issue may be one of expected effects as opposed to the intended effects.

Alternatively, the harming mechanism can be viewed from the perspective of the code and algorithms in the cyber capability that are used to achieve or maintain access through a vulnerability in the targeted device, or by issuing commands from within the targeted device that, by design or expected operation, cause damage or destruction. The positional variants are not mutually exclusive. For instance, the 'harming mechanism' in Stuxnet was code (payload) that issued a rather innocuous command (to the targeted device) varying the speed of specific centrifuges resulting in the damage.[60] Of course, similar code aimed at different aspects of a target device may achieve significantly different effects. Code (such as KillDisk) that erases the operating system of a target may be strikingly similar to code that erases the ACC after it has achieved its designated effect, but only the former would potentially be reviewed.

Third, the potential reviewable aspect of autonomy extends beyond algorithmic control of the harming mechanism, which effects human intent, to cover those elements of autonomy that are integral to combat functionality. Indeed, as the extent or complexity of the autonomy increases, the full contours of the autonomous control that requires review may need to be carefully worked through. In Stuxnet, for instance, the autonomous element included propagation, identifying a specific target (SCADA and PLC systems used in Natanz, Iran), determining authority to engage target device, and issuing commands to the target device.[61]

Fourth, when autonomy and cyber are combined, it will become increasingly difficult to identify when and in what way a harming mechanism will be engaged. In situations where the code utilised to achieve the effect is chosen by a controlling algorithm, the harming mechanism will not always be identifiable. This is especially the case where variability exists in ACC — whether from adaptive capacity in the controlling algorithm, the level of autonomy applied to a task, or in the actual code

60 Stamatis Karnouskos, 'Stuxnet Worm Impact on Industrial Cyber-Physical System Security' (Paper presented at IECON 2011 — 37th Annual Conference of the IEEE Industrial Electronics Society, Melbourne, 7–10 November 2011) <https://ieeexplore.ieee.org/document/6120048>. Interestingly, the design and architecture of the Stuxnet worm is not target specific. That is, it the code used in Stuxnet could be, and may subsequently have been, re-purposed against a large number of other SCADA and PLC systems.

61 ibid; Katharina Ziolkowski, 'Stuxnet — Legal Considerations' (NATO CCDCOE 2012) <https://ccdcoe.org/library/publications/stuxnet-legal-considerations/>. While Stuxnet can be distinguished from other less discriminating malware such as that used in the Wannacry, Notpetya or BadRabbit 'purported ransomware' attacks, this does not mean that the autonomous elements of such malware should not be assessed as part of the means or method. The relevance of autonomous elements — whether access, propagation (such as Eternal Blue and MimiKatz), target identification, or damaging mechanism — of code that are means or methods will be determined by the normal or expected use.

used as the harming mechanism (such as through re-writing).[62] In such instances, the range of decision execution options provided by autonomy paired with variability of damaging mechanism, results in an exponential increase in combinations of achieving the design purpose. This potential complexity raises the question as to whether an advanced ACC could be reviewed.

In many instances, not being able to identify the harming mechanism(s) will likely add complexity to the weapons review, making initial acceptance and any ongoing certification more arduous. It should not, however, prevent an ACC from being reviewed. In these situations, the effect of the ACC — its ultimate goal — would need to be assessed for compliance with LOAC. As novel as this may seem, the reality is that the weapons review must focus more on the predictable effect and less on the nature of the harming mechanism (outside of any specific prohibitions). That is, in reviewing a capability, the key is the predictable effect. Reviewing ACCs will therefore be possible where the effects are known, even if the harming mechanism is not entirely obvious at the time of the review.[63]

4 *LOAC and Weapons Review Obligations as Design Criteria*

An ACC designed to cause damage or harm should be capable of being reviewed. If some or all of an ACC's 'normal or expected use' is unable to be appropriately analysed during a weapons review, its use is likely to be either limited or prohibited as the reviewing State is unable to ensure it can be used consistently with its LOAC obligations. At a minimum, reviewability requires that the effects of the ACC are predictable or capable of being limited. Aspects of the ACC's design, code and function may also need to be identifiable, measurable, and explainable/understandable,[64] to permit the assessment of compliance with LOAC. To this end, the weapons review obligations may need to be considered as design criteria. Indeed, given the fundamental importance of the 'normal or expected use' of a capability complying with LOAC, States may regard their LOAC obligations as essential criteria for the design and training of ACC.[65] For

62 Tim McFarland, 'The Concept of Autonomy', this volume, ch 2, 33: 'autonomous capabilities are likely to exist in specific sub-systems performing specific functions rather than be applied to the system as a whole'.

63 ACC are designed to achieve a particular effect, but the level of autonomy they are given permits the ACC's algorithms to determine how the effect is achieved. It is this end result that must be the focus of the weapons review, albeit that the method (that is, the manner in which the code autonomously selects to complete the task) must not be specifically prohibited by LOAC either.

64 Arthur Holland Michel, 'The Black Box, Unlocked: Predictability and Understandability in Military AI' (UNIDIR 2020) 9 <https://unidir.org/sites/default/files/2020-09/BlackBoxUnlocked.pdf>.

65 Legal advisors must actively participate in the study and development of ACC and advise technical

example, an essential performance specification may include the LOAC rules, principles and concepts described in Part 1 of the ICRC Guide.[66]

PART 3 DETERMINATION THAT THE CAPABILITY IS A 'NEW WEAPON, MEANS OR METHOD' THAT REQUIRES REVIEW[67]

Requirement. Part 3 of a weapons review entails a determination as to whether a capability in its 'normal or expected use' qualifies as a weapon or means of warfare.[68] For AP I States the Article 36 obligation extends to the review of methods of warfare[69] beyond the contextual use of a means that is being reviewed.[70] Methods of warfare are traditionally only reviewed as TTPs (how operations for conducted in armed conflict), rather than being a review of any particular operation.[71] Method reviews are consequently usually undertaken without consideration for the context of any use. Such narrow-focussed reviews would not be appropriate for reviewing the interaction of algorithms required for an ACC's cyber and autonomous functionality.

Analysis. There are two facets to the determination in Part 3: First, is the capability being reviewed a 'weapon, means or method' for the State? Second, if it is, what aspect of it requires reviewing?

1 *Determination that it is a Weapon, Means or Method of Warfare*

Central to determining whether an ACC is a means or method of warfare is a State's interpretation of autonomous functionality and the underlying cyber capability, whether individually or in combination.[72]

personnel to ensure LOAC principles, for example the principles of distinction and discrimination, underpin the design or development of a capability.

66 ICRC (n 35) 9–20.

67 In cyber capabilities the ability to determine that it is a means of warfare will often require analysis of the design and technical features and the 'normal and expected use' or effect. Conventional means are usually a little more obvious, thus making it easier to determine their status earlier in the weapons review.

68 As identified above, AP I States must review weapons, means and methods of warfare that their respective military forces intend to study, develop, acquire or adopt, and non-AP I States must — potentially — review means of warfare prior to them being fielded: *Tallinn Manual 2.0* (n 4) 465, rule 110. This indicates that customary international law requires all States to review all means of warfare including munitions, weapons and weapon systems their respective military forces intend to use). For a more comprehensive analysis of (and contrary view on) the customary international position, see Jevglevskaja (n 17).

69 'Methods of warfare' includes attacks and other activities designed to adversely affect the enemy's military operations or military capacity.

70 That is, new methods of warfare are required to be reviewed where they are not a supplementary consideration in the review of a means of warfare (in the context of its 'normal or expected use').

71 Biller and Schmitt (n 7) 221.

72 In section I it was identified that an ACC with a normal or expected use to cause damage or harm

Typically, identification of the 'cause, damage or harm' aspect would come from the design or specific purpose behind the creation of a capability, and be evidenced by a harming mechanism. Unfortunately, as discussed in the previous section, identifying the harming mechanism(s) may not be as straightforward as with conventional means.

The 'autonomy' element of an ACC with a 'normal or expected use' of controlling a harming mechanism and using such control to enable or undertake combat functionality would be classified as a method requiring review of contextual use. [73] Combat functions, in this sense, would not include autonomy in tasks or functions relating to areas such as weapon system mobility (for example, aircraft navigation), system management, or system interaction, unless and to the extent that they are integral to a combat function.

2 *What Elements of an ACC are Reviewable?*

Having determined that an ACC should be the subject of a weapons review, the next question States need to consider is whether the obligation, as a matter of international law, attaches to some or all of the elements or systems of an ACC. That is, which parts of the ACC require review — is it the entire capability or merely elements thereof? For example, does the method of propagation, the exploit method, or any data gathering capability require consideration in the weapons review? International law provides no specific assistance, but given that the purpose of weapons reviews is to ensure that capabilities used in armed conflict are lawful, it could reasonably be argued that a holistic evaluation of a capability is appropriate.[74] Furthermore, with ACC, automating elements integral to an attack will necessarily create a greater connection between the various steps in a cyber operation because a human is not directly manipulating each step of the operation but rather has built the parameters of the operation (and the attack) into the ACC to execute. In other words, to make a cyber capability autonomous, the enhanced connectivity of the elements will require a conglomerate payload (and exploit) capability that is able

in armed conflict can be a means of warfare. Where found to be a means or method of warfare it must be reviewed by AP I States (and potentially also by non-AP I States who determine it is a means). It would behove non-AP I States who determine it is a means or method to undertake a weapons review.

73 Thompson Chengeta, 'Are Autonomous Weapon Systems the Subject of Article 36 of Additional Protocol 1 to the Geneva Conventions?' (2016) 23(1) UC Davis Journal of International Law & Policy 65, 75.

74 This issue is particularly relevant to cyber technology that employ AI to perform specific aspects of its 'normal or expected use'. This could include AI systems for monitoring movement or propagation, identification of the target, execute action in response to variable situations and, depending the extend of programmed parameters, either engage the target independently or cue for human decision making.

to undertake several steps/phases of a cyber operation independent of direct human manipulation. This increases the likelihood that otherwise distinguishable elements of an ACC require review.

This does not mean that all elements of an ACC would need to be reviewed and, from a practical perspective, it may not be desirable to do so.[75] Aspects of the cyber capability that are not integral to targeting are unlikely to be the subject of a weapons review. For example, those aspects related to emplacing an ACC — such as reconnaissance to gather intelligence and access — would not normally need to be assessed, but internal target reconnaissance and target identification likely would be.

(a) Approaches to Identifying Reviewable Elements of an ACC

Given the silence of international law, the question of whether some or all aspects of a particular ACC should be subjected to review will be answered by reference to State practice, articulated in State policy. There are several potential approaches that a State could consider in identifying reviewable elements: focusing on whether the element is integral to 'normal or expected use', whether it directly enables combat functions, or whether it involves functions that engage legal obligations; or, alternatively, considering the risk of causing significant harm to persons or property.

The first approach a State could apply involves identifying and assessing those elements that are integral to the ACC's 'normal or expected use'. This broad approach would theoretically capture all elements that enable the ACC to undertake or enable combat functions. This would include payload elements that directly permit the engagement of combat functions, such as selecting and attacking a target device, as well as payload elements that indirectly enable the weapon's 'normal or expected use', such as reconnaissance or intelligence gathering.[76]

Alternatively, a State could elect to review only those components of an ACC that directly enable combat functions. This approach would exclude consideration of the weapon platform (propagation and exploit) and other systems (payloads) that are not directly involved in undertaking or enabling combat functionality even if they are integral to overall system

75 For example, where the autonomy element is created or supported by artificial intelligence (AI), it is likely to be constructed as systems of systems. The cyber 'payload' and its means of its delivery may be only one of a number of systems that enable the autonomous operation of the cyber means or method of warfare. Not all of the systems will necessarily require review.

76 Consider an unmanned aerial system that employs an autonomous pilot system, removing the requirement for a remote human pilot. As the aerial delivery of the missile is essential to its normal or expected use, the art 36 review of the autonomous pilot system would be necessary.

functionality.[77] Conversely, the accuracy and reliably of the included autonomous systems (algorithms) and the associated sensors that identify the specific target of an ACC would be covered by the weapons review.

A variation of the first two approaches is for a State to review only those elements of an ACC that engage the State's obligations under LOAC or international law more broadly. Such an approach would ensure that all elements relevant to the traditional weapons review obligation are considered. The review would include elements that directly or indirectly undertake or enable combat functionality (payload), but also elements (exploits and propagation/delivery method) that engage legal obligations not relating to use (such as communications, neutrality, and even domestic legal considerations).

Finally, States may adopt a risk methodology to identify reviewable elements of an ACC. This approach is consistent with the regulation of 'safety-critical' software in the aviation, space and automotive industries. Industry standards distinguish different classes of 'safety integrity levels' to 'reflect the degree of severity of possible consequences of failures'.[78] In this way, those elements of an ACC that can cause significant harm to persons or property would be the subject of the weapons review. This approach is of particular interest for ACC where elements of code that have potential for significant harmful consequences would be reviewed, where they would normally be excluded from a weapons review. This would potentially require a review of zero-day exploits, propagation methodology, residual code (akin to explosive remnants of war), pre-positioning elements, and other similar components of an ACC that are relevant to getting in or out, or that are left behind, but are not directly or even indirectly relevant to combat functionality.

(b) 'New'

Article 36 provides a further consideration for ACC through the inclusion of the adjective 'new'. Historically, this word created a temporal trigger to a State's weapons review obligation. This practical expedient allows States to determine as a matter of policy rather than law whether they would review a weapon already in their possession at the time of ratification of AP I.[79] For conventional weapons without autonomy, the

77 For instance, the elements of an ACC that undertake target reconnaissance and identification to acquire any SCADA system may not be directly relevant to any subsequent use of a specific SCADA system.

78 Martin Hagström, 'Complex Critical Systems: Can LAWS Be Fielded?' in Robin Geiss (ed), *Lethal Autonomous Weapon Systems: Technology, Definition, Ethics, Law & Security* (Federal Foreign Office of Germany 2016) 132.

79 Effectively, it required States to only review weapons procured after their ratification of AP I.

question of when a means or method of warfare is new for the purpose of Article 36 is reasonably well settled amongst States and academics. Thus, a weapon needs to be reviewed:

- when it is first studied, developed, acquired, or adopted;
- when an existing weapon is substantively modified; or
- where a State's treaty obligations alter the legality (or otherwise) of a weapon.[80]

The latter two are also triggers for a re-review obligation, although it is unclear how this applies in practice. Unfortunately, this lack of clarity is directly relevant to ACC, which include cyber and autonomy elements.

As mentioned earlier, most cyber capabilities are designed for a specific purpose. They are unique in that, for obvious reasons, they are unlikely to be re-used in their entirety.[81] While it is reasonable to expect that a weapons review would be the first review of a cyber capability, it is possible that aspects of the code comprising the capability will be re-used, substantively altered through self-correction[82] or re-purposed.[83] If a substantive component of the code were to be re-applied,[84] at the very least the State would be obligated to reconsider the original review to determine whether the modification, or new method of use, renders the original review obsolete. The result of such reconsideration may vary.[85]

The inclusion of autonomy in a cyber capability could further trigger the requirement to review it as a 'new' weapon and means in two ways. The first and primary concern is where the ACC is driven by some form of code — or even artificial intelligence ('AI') — with adaptive capability. Adaptive capability in the autonomy element raises the spectre,

This marker avoided the need for a grace period and the prospect of States being immediately non-compliant with the art 36 weapon review obligation. On this, see Parks (n 12) 114.

80 McClelland (n 7) 404; ICRC (n 35) 8–9.

81 Fahmida Y Rashid, 'Stuxnet Worms May Come from Different Authors' (*Eweek*, 20 October 2011) <https://www.eweek.com/security/duqu-stuxnet-worms-may-come-from-different-authors>: 'The general approach among malware developers is to "hit once, then dispose of the code," [BitDefender researcher Bogdan] Botezatu wrote. Code "reuse" is a bad practice among malware developers because most major antivirus vendors would have already developed heuristics and other detection capabilities for that code sample.'

82 McFarland (n 62) 23. Code that can re-write, correct or repair itself warrants ongoing review.

83 Samuele De Tomas Colatin and Ann Väljataga, 'Data as a Weapon: Refined Cyber Capabilities under Weapon Reviews and International Human Rights Law' (NATO CCDCOE 2020) 9 <https://ccdcoe.org/library/publications/data-as-a-weapon/>.

84 For instance, some of the code used in Stuxnet has been re-used in other cyber capabilities (such as Duqu).

85 Many people incorrectly assume that a modification will result in a deleterious change in the performance of a weapon. Rather, a weapon modification will often enhance the performance of a weapon and thus improve its compliance with a State's international law obligations.

potentially, of a means that could modify or vary its methods of warfare (and therefore its 'normal or expected use') constantly. One way of addressing this potential is provided in Part 5 below ('General Law'). The second, albeit less likely, novel application of the 'new' requirement is where the particular use of the ACC will result in a variation of the effect expected (even where 'normal or expected' use remains the same). This would entail a variation in the target device or domain, so as to sufficiently alter the potential effects from using the ACC, and therefore require further review to determine that the ACC complies with the law.[86]

PART 4 **SPECIFIC LAW**

Requirement. Part 4 is the commencement of the legal analysis of the means or method of warfare, comprising of the first test in the two-fold test recommended by the International Court of Justice ('ICJ') in the *Nuclear Weapons* Advisory Opinion.[87] In this part, a reviewer determines if a means or method is specifically prohibited or restricted through obligations assumed by the relevant State under applicable treaties[88] or customary international law as it applies to that State.[89] Specific prohibitions or limitations normally attach to a particular instrument or type of weapon.

At the time of writing, there are no specific prohibitions[90] on ACC as a means or method of warfare, although there are separate ongoing discussions at the United Nations regarding 'lethal autonomous weapon systems' and cyber.[91] As noted above, a reviewer would need to consider

86 A change in target device or target domain would usually require a new or adapted ACC to be developed. This would trigger an original review of the 'new' ACC rather than a re-review of a modified ACC. The autonomy element of an ACC may permit variation in the target device or domain without substantive variation in the ACC. An operational legal review (OLR) — a review of the legal aspects of an operation or activity — may not be sufficient to identify and address all of the legal issues, necessitating a weapons review.

87 *Nuclear Weapons Advisory Opinion* (n 16). The Court adopted a two-fold test: (1) Is there any customary or treaty law that contains a *specific prohibition* against the threat or use of a weapon in general or in certain circumstances? (2) In the absence of a specific prohibition, is there a *general prohibition* against the threat or use of a weapon in general or in certain circumstances?

88 Tim L H McCormack, Paramdeep Mtharu and Sarah Finnin, 'Report on States Parties' Responses to the Questionnaire: International Humanitarian Law and Explosive Remnants of War' (Asia Pacific Centre for Military Law 2006) 35: 'the weapon should be assessed for its compliance with the terms of any treaty to which the [reviewing] State is a party, taking into account any reservations that the State may have entered upon ratification.'

89 ibid 36.

90 For example, Chief of the Australian Defence Force, *Australian Defence Doctrine Publication (ADDP) 06.4: Law of Armed Conflict* (11 May 2006) ch 4 provides a list and description of specifically prohibited or restricted weapons.

91 For the history of, and current information on, discussions on lethal autonomous weapon systems, see United Nations, 'Background on Lethal Autonomous Weapons Systems in the CCW' <https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?

their own State's treaty obligations (including non-binding arrangements such as the Wassenaar Arrangement) to determine if a particular ACC was affected.[92]

PART 5 GENERAL LAW

Requirement. The second test recommended by the *Nuclear Weapons* Advisory Opinion requires a reviewer to determine whether there is a general prohibition against the use of a weapon, whether generally or in certain circumstances (hence focussing on the effect of weapons or the means of warfare). While evidence of State practice is limited,[93] this would usually entail reviewing the technology against core LOAC principles — unnecessary suffering, indiscriminacy, and environmental harm — to determine whether a capability or operation is unlawful *per se* or in certain circumstances.[94]

Analysis. ACC pose novel challenges to conventional weapons review practice in the application of the second test. Traditional weapons reviews primarily focus on the design of the means of warfare to identify compliance with the core LOAC requirements (effectively an 'instrument' review). While the design and structure of an ACC will extend the requirements of the 'instrument review',[95] the inclusion of autonomy algorithms (whether some level of AI or coded step-by-step automation) creates a significant additional review element,[96] namely assessing the decision-implementation capability of the ACC (effectively a 'use' review). The

OpenDocument> accessed 31 December 2020.

92 While there is no specific treaty obligation for ACC at present, this does not mean that specific types of ACC are not captured by extant treaty obligations or customary international law. For example, a States interpretation of its obligations under the Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as amended on 3 May 1996 (Amended Protocol II) (adopted 3 May 1996, entered into force 3 December 1998) 2048 UNTS 93 may result in certain ACC being limited by that State due to their similarity with 'booby traps'. Alternatively, the 'normal or expected' target device of an ACC (i.e. the operating system controlling a dangerous force such as chemical storage facility) may enliven specific treaty obligations.

93 Lawand (n 48) 925–30.

94 For AP I States this would involve the articulation of the AP I and customary rules; for non-AP I States only the customary international law positions. In particular for AP I, see art 35(2), covering 'weapons, projectiles and materials and methods of warfare that cause superfluous injury or unnecessary suffering'; art 51(4), covering indiscriminate attacks; and art 35(3), covering environmental modification. See Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law* (Cambridge University Press 2005) vol 2, rules 43–45, 70 and 71 for the status of these requirements.

95 It is expected that an ACC would combine a delivery device (potentially including propagation method, and exploit), sensors (for data gathering), a harming mechanism(s), and a decision-implementation capability.

96 There is a limited history of systems that employ an automated target identification (and indeed even selection and engagement). Examples include the Phalanx CIWS, missiles with 'lock-on-after-launch capacity', and autonomous data fusion and target identification systems.

less deterministic (or more intricate) the AI used to undertake LOAC decision-implementation, the greater the potential for the ACC to pose interesting issues regarding assessment, and the greater the likelihood for expansion and complexity in the weapons review. Some of the more obvious issues specific to ACC are addressed below.

1 *Specific Issues for ACC*

(a) No Presumption of Lawful Use

In an 'instrument' review, the focus is on the legality of the capability, i.e. is it legal to field and use in some, or all, circumstances of its 'normal and expected use'? It requires the capability to be assessed against each weapons law criterion independently, and while it will consider capability to be employed legitimately,[97] this occurs within the context of a presumption of lawful use by the military. An ACC contains autonomy algorithms that provide it a level of autonomous functionality to undertake or enable combat functionality against an adversary (whether using its own harming mechanism or by controlling an adversary device or network). This gives effect to command intent by executing pre-approved decision-implementation functionality, which previously would have been undertaken at the time of targeting by humans. By allowing ACC to participate in decision-implementation, a level of targeting autonomy effectively negates the presumption of lawful use.

Absent a legitimate presumption the weapons review would be required to consider the legality of use ('use' review) in specific LOAC scenarios: Can the ACC apply targeting law in combat in accordance with its 'normal and expected use'? Providing autonomy to a cyber capability therefore has two significant consequences for this Part of a weapons review. First, it expands the focus of the weapons review for those elements that are subject to a 'use' review (of every exercise of combat functionality)[98] when compared to an 'instrument' review. Second, it requires consideration of an expanded range of elements, in particular targeting law, and use compliance standards.

97 Lawand (n 48) 928.
98 Performance would be evaluated against relevant standards. It does not require evaluation of misuse.

i Targeting Law

An 'instrument' review will consider the ability of the means to discriminate, and its potential to cause unnecessary suffering. Incorporating a decision-implementation component requires the reviewer to set aside the presumption of lawful use and determine if the means or method can exercise combat functionality.[99] In particular an ACC weapons review will also require assessment of the 'normal or expected use' of the capability in replacement of, or integrated with human decision-making, with regards to some or all aspects of proportionality and the AP I precautions in attack (including specific protections), individually and in combination.[100]

ii Standards

An 'instrument' review typically focusses on performance reliability. Performance reliability is effectively an objective measure or measures of the success of a weapon in performing as intended.[101] It is usually identified by the manufacturer/developer during the test and evaluation phase to assess the weapon against design specifications.[102] Traditionally, performance reliability is quantified as a percentage of the number of successful tests to the total number of tests, within given statistical confidence bounds.[103] It is not however a static standard, continually evolving as means of warfare and technological precision develop.[104]

99 A contextual art 36 review is assessing legality of the ACC (i.e. is it a legal weapon through compliance with the pertinent aspects of targeting law relevant to the normal or expected use of the ACC), but is not assessing the legality of actual use of an ACC (which can only occur in the context of an actual use).

100 An ACC designed to autonomously identify and attack specific cyber systems must be capable of complying with the IHL principle of distinction. The ability to do this relies on programming and target analysis. This may be achieved by targeting unique cyber infrastructure. However, if the harming mechanism attacks software or hardware common to both the target and civilian cyber systems, the weapons review must confirm the ACC's ability to distinguish between its intended target and other civilian cyber systems. This will require the reviewing State to determine the standard of certainty it will require the ACC to achieve.

101 Defense Science Board, 'Defence Science Board Task Force Report on Munitions Systems Reliability' (US Department of Defense, September 2005) 14, acknowledging that 'reliability means different things to different groups', but settling on the view that '[r]eliability describes the expectation of a munition's ability to perform its intended function over successive trials'.

102 See Alan Backstrom and Ian Henderson, 'New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Reviews' (2012) 94(886) International Review of the Red Cross 483, 508–13 for discussion on reliability. See also Defense Science Board (n 101) ch 1 for a detailed explanation of munitions system reliability rates.

103 Backstrom and Henderson (n 102) 508. Quantifying reliability is not a binary proposition, but rather requires statistical confidence bounding based on the extent and nature of testing undertaken.

104 For instance, early focus of performance reliability was on the success (or failure) rate of the operation and effects of munitions (i.e. a bomb detonating). As weapons (and methods of warfare) have become more advanced, including the incorporation of fuzing and guidance mechanisms, the importance of 'accuracy' has become more prominent in weapon function and thus performance reliability has expanded to incorporate it.

For ACC, performance reliability will primarily be framed in terms of predictability. Unlike traditional success/failure conceptions of reliability, predictability of effect will focus on the extent to which the effects are expected.[105] The predictability of effect will vary according to the type of ACC, the function, and the environment.[106] It will be for States to determine how they quantify this. One reasonable approach is to quantify the assessment of the ACC's ability to perform a task in accordance with testing,[107] and the extent to which the performance of the autonomous elements can be anticipated when employed.[108] A secondary predictability measure that States will need to consider is the qualitative predictability of foreseeable but unintended effect.[109] This measure recognises that the interaction between the ACC, the mission and operating environment requires consideration of the risks of foreseeable consequential effects that are not intended.[110] The required level of statistical confidence for predictability measures will vary according to States' acceptance of risk, and will be potentially affected by limitations on testing cyber capabilities.[111]

Autonomous functionality generates the requirement for supplementary assessable standards. The primary example of this are performance (and acceptance) standards, established in the design specifications of the new capability to address the required legal thresholds for executing LOAC decision-implementation and risk control autonomously. Performance (and acceptance) standards will be quantified into an objective

105 Holland Michel (n 64) 6. This is 'the degree to which the outcomes or effects of a system's use can be anticipated'.

106 ibid 6–7. Predictability of effect is determined by consideration of a 'broad range of overlapping factors' including qualities of the ACC (type of system, system development and training, testing, capability of self-learning, data and training diet); function (type of task and function, scale and length of deployment); and employment (complexity of environment, number of interacting systems).

107 ibid 5. Effectively a performance reliability measure: 'In the technical sense, "predictability" generally refers to a system's ability to execute a task with the same performance that it exhibited in testing …'.

108 ibid. This 'refers to the degree to which an autonomous system's individual actions can be anticipated.' It is a 'function of the characteristics of the environment and mission for which the system is deployed.'

109 With algorithms as they become opaquer and more complex the type and nature of consequential effects that are not intended or expected become harder to identify. The predictability of serious foreseeable but unintended effects therefore becomes relevant to weapons review when trying to determine the risks that these unexpected effects represent.

110 This is not intended to be a quantitative measure; rather it is intended to consider the nature and seriousness of foreseeable effects which are not intended so that where required measures to prevent or limit the adverse component of these foreseeable effects can be implemented.

111 The required level of statistical confidence (that the reliability/predictability percentage is accurate) will vary according to risk (which increases the lower sample size or the lower the number of tests), in recognition of the limits on testing capabilities (finite and imprecise), or complexity of the cyber technologies autonomy algorithms (the higher the complexity the more exhaustive the testing required). Complexity is multi-faceted and includes software complexity, complexity of normal or expected use, complexity of the adaptive system, complexity of the combat environment, complexity of machine-human interaction, mix of AI functionality (machine learning, machine reasoning, sensor integration and access etc.), and complexity/variability from software patches, upgrades or improvements.

measurement of probability, which would be set at or above the legal threshold (where quantifiable).[112] For the immediate future, they will likely be somewhat higher (as a risk and perception management tool).[113] Performance standards, in the form of levels of certainty required for executing LOAC decision-implementation, while not currently an aspect of performance reliability for traditional weapons, will inevitably become an integral part of performance reliability for ACC as a representation of predictability of the autonomy algorithms.[114]

In short, for an ACC, the more traditional assessment of the reliability of a capability to employ the relevant harming mechanism will be complemented by also evaluating the performance reliability of the algorithms in meeting designed performance standards. Consequently, ACC performance reliability depend on some or all of the following:

- accuracy of specification regarding LOAC decisions and the underpinning law applicable to a State (including performance standards);
- reliability to perform those LOAC decisions repeatedly (a factor to be considered here will be cognitive adaptive capacity of the algorithms);
- accuracy of the coding to allow the autonomy algorithms to operationalise LOAC decisions;[115]

112 For instance, 98–99% certainty that an object is a military objective.

113 There are a couple of reasons for this. First, algorithmic representation of a required performance standard requires quantification in the form of an objective measurement. Objective measures provide a simple method to assess performance to specification for weapons reviews. Second, the performance standard represents the minimum acceptable accuracy of the means or methods' autonomy algorithms in performing a function at the test and evaluation phase. If the performance standard relates to the normal or expected use of the cyber capability, then it forms part of the intended function of the autonomous capability, for which reliability will be important to pass the evaluation and testing phase. Performance standards therefore naturally fit as a measure(s) of accuracy to specification and thus an extra element of performance reliability.

114 For States, the degree of accuracy acceptable is weapon type and effect dependant. The required degree of confidence (represented as a statistical probability) increases as new weapons are developed. The US Defence Science Board claimed that 'typical munition reliability rates for conventional munitions fall in the 95 to 99 percent range' but noted that 'reliability was subject to a statistical confidence bound', 'no class of munition will be faultless' and there is a 'lack of information regarding actual combat reliability rates'. Defense Science Board (n 101) 15, 17–18. See also Tetyana Krupiy, 'Of Souls and Ghosts: Transposing the Application of the Rules of Targeting to Lethal Autonomous Robots' (2015) 16(1) Melbourne Journal of International Law 6, suggesting, from a limited sample, that a reliability rate of above 99% is appropriate. While this rate is not consistent with the legal position or actual State practice it would be fair to say that there is an upwards spiral. Reliability rates are increasingly expected to be above 90% for most weapons and over time some weapons at the completion of test and evaluation stage are approaching 100%.

115 This includes the reliability and accuracy of the sensors used by the cyber capability, and the reliability of integration of the autonomy algorithms with sensors.

- accuracy of autonomy algorithms in undertaking LOAC decisions to specified performance standards; and
- appropriately weighted levels of statistical confidence for LOAC decisions — based on test limitations, such as limited learning experience or no live experience, that is limited to computer modelling and simulation limitations.

A further standard to be considered is the legal standard: the technical standard of legal compliance with LOAC rules. For the purposes of coding, the standard would need to be converted into an 'objective and measurable quantity' — or numerical representation — for a cyber capability.[116] Identifying technical standards of legal compliance is not a straightforward task, as many rules are written with 'terminological imprecision,'[117] are designed for broad application across a range of conflict situations subjectively by humans,[118] and will require determination of a range of compliance standards for different operating environments or scenarios.[119] Such standards invariably are a question of reasonableness, albeit with varying levels of context[120] and subjective judgement by the appropriate decision-maker.[121]

While some argue that 'code' cannot satisfy this 'human' standard — and there is a live question regarding whether code technically will ever be able to satisfy the standards' requirements — the reality is that autonomous capabilities implement pre-defined human decision-making. This means the minimum threshold for LOAC actions executed

116 Backstrom and Henderson (n 102) 497. Consideration should be given to the accuracy of the sensors in gathering and interpreting data in the relevant situation or context to assess whether the statistical probability exceeds the 'objective and measurable' quantification of the required subjective legal standard.

117 Michael Schmitt, 'The Principle of Distinction in 21st Century Warfare' (1999) 2 Yale Human Rights & Development Law Journal 144, 150.

118 For instance, the proportionality rule requires a subjective assessment of dissimilar values.

119 For example, compliance with the rule of distinction for an autonomous weapon designed for operation in closed environments such as isolated systems separated from the internet may require a lower standard of compliance than for an autonomous weapon designed for operation in an online system.

120 For instance, under the precautions in attack requirement of AP I art 57(2)(a)(i), there is an obligation to 'do everything feasible'. What is feasible is that which is 'practicable or practically possible, taking into account all circumstances ruling at the time'. This is context specific requiring consideration of military, humanitarian, political and any other relevant to situation factors. Another example is the proportionality requirement in AP I arts 51(5)(b) and 57(2)(a)(iii), directing the decision-maker to assess the qualitative elements of military advantage.

121 For instance, AP I arts 50(1) and 52(3) (civilian status of persons and objects — reasonable doubt), arts 51(5)(b) and 57(2)(a)(iii) (proportionality — reasonable commander making reasonable use of the available information), art 57(3) (choice between military objectives — reasonable expectation). See Michael N Schmitt, 'Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics' (2013) Harvard National Security Journal Features, 21: 'Neither the human nor the machine is held to a standard of perfection; on the contrary, in international humanitarian law the standard is always one of reasonableness.' See also *Prosecutor v Galić* (Judgment) ICTY-98-29-T (5 December 2003) [58]: 'reasonably well-informed person in the actual circumstances of the perpetrator, making reasonable use of the information available'.

by ACC is the same as that required of human decision-makers.[122] This position is contentious. It is therefore important to recognise that while legal standards are often misrepresented as the only standard acceptable for legal compliance, they are in fact an absolute minimum threshold for a spectrum of conduct in armed conflict.[123] That is, military operations can be undertaken at the legal standard or at any higher standard (whether relating to the level of certainty, or contextual knowledge or some other specified requirement).[124]

As a matter of policy (risk management), or due to current technical deficiencies in code, permitting execution of LOAC decision-making at or near the absolute thresholds is unlikely. As such, the assessable standard for the algorithms of the cyber capability will, at least for the foreseeable future, be a higher or more restrictive performance standard — even if a transparency measure is included.[125] Weapons reviews would therefore involve assessing whether the ACC's coded performance standard meets or surpasses the requirements of the legal standard. This would require considering whether the ACC's claimed performance standard (in some form of statistical probability quantifying the accuracy of the cyber capability's sensors in gathering, interpreting[126] and fusing[127] data) and the ability of the cyber capability's algorithms to apply the rules of LOAC (to

122 Most States introducing new weapons that are replacing or supplementing another means or methods compare them to the replaced weapon. In this instance, the ACC's autonomy algorithms are implementing human decisions — it is in effect bringing forward the human decision-making to the point of release of the means or method with autonomous functionality vice the traditional point where force is applied to an adversary or military objective. As such, under current State review methodology, the minimum standard utilised would be that required by human decision-makers.

123 Restrictions on the operating parameters of the ACC (restrictive target sets, collateral damage limitations, narrow geographical boundaries etc) could ensure compliance by avoiding placing the ACC in a position to execute decisions where the absolute minimum legal standards would apply. For instance, an ACC could be restricted to conducting attacks only against a discrete category of objects, or when it is certain there is no or little risk of collateral damage, effectively obviating the need for a proportionality assessment to be made.

124 The main provision is AP I art 57 but other articles (arts 51, 52 etc) may be relevant. AP I provides standards for the precautions in attack, for instance: art 57(2)(a)(i) (everything reasonable with respect to initial Distinction); art 57(2)(a)(ii) (feasible precautions with respect to avoiding or minimising collateral damage); art 57(2)(a)(ii) ([reasonable] expectation with respect to proportionality); and art 57(4) ([reasonable] expectation with respect to choice of objectives for similar military advantage). In a situation where a standard is not provided by Treaty, is not clear, or there is no relevant treaty applicable to a State, relevant courts have identified applicable standards. For instance, with respect to attacking objects that, although normally dedicated to civilian purposes, are being used for military purposes, the ICTY Trial Chamber in *Galić* (n 121) [51] identified that a 'reasonable belief' standard applies.

125 Transparency (also referred to as explainability, understandabilty) measures are not a legal requirement or a performance measure under LOAC. Explaining 'how' (in varying degrees) a decision was implemented by an algorithm offers potential benefits of accountability, trust and improved performance.

126 Backstrom and Henderson (n 102) 495, calling this the 'uncertainty of measurement' and noting that it is a 'distinctly separate acceptance criterion' from the standard required by LOAC.

127 Mohammed Hossny and others, 'Towards Autonomous Image Fusion' (Paper presented at 11th International Conference on Control, Automation, Robotics and Vision, Singapore, 7–10 December 2010), noting that there are a number of errors that occur with fusing data.

the data gathered), subject to confidence bounds, exceeded the minimum legal standard.[128]

(b) Computer Meaning is Not Human Meaning

An important consideration when assessing ACC is that they are not human. This rather trite statement refers to the tendency for humans to anthropomorphise non-human entities with human qualities. Attributing human understanding to software code and algorithms that implement human decisions is one such common tendency. Software code and algorithms perceive all things as numerical representations — there is no 'meaning' in the human sense.[129] Numerical representations often rely upon mathematical generalisations: models simplified to allow an algorithm to interpret.[130] In other words they are generalisations of reality — models that support prediction. As such, while reviewing code may be important (to check, for instance, that standards have been set correctly), in the end it is the assessment of the predictability of effects from the operation of those algorithms that is key.[131]

(c) Assessment Data

Assessing the effects. Historically, weapons reviews have focussed on objective measures of performance reliability to determine if the weapon can be used lawfully.[132] For an ACC, this means an assessment of the predictability of the consequences or effects from the operation of the code as opposed to an assessment of the code itself.[133] Assessing an ACC's

128 As such, an ACC that operate with a narrower normal and expected use (in other words with constraints or risk controls on key variables such as domain, target type, duration and level of autonomy, acceptable proportionality and design safeguards including increased presumption of civilian status) is more likely to be both assessable and able to operate within the spectrum of LOAC legal standards at present.

129 Smith (n 23) 208 and 226.

130 ibid 151 and 291: an 'algorithm's representations are always models, always technical manifestations of the reference frames of the humans who created those algorithms'.

131 Peter Margulies, 'A Moment in Time: Autonomous Cyber Capabilities, Proportionality, and Precautions', this volume, ch 8, highlights the problems with autonomous agents as brittleness, bias, unexplainability, and automation bias. The core issue with unexplainability is represented by Polanyi's Paradox and the ACC knowing more than it can tell. It is believed that if the ACC can explain itself this will improve trust and predictability. No appropriate or usable standard for explainability has been identified yet although projects such as DARPA's Explainable Artificial Intelligence (XAI) are worth monitoring.

132 The assessment would normally be based on empirical testing data from the manufacturer used to support the test and evaluative stage of development. In a small number of reviews, further data (to address aspects of reliability such as accuracy, control of effects, and reliability of detonation) may have come from initial testing undertaken by Defence agencies or units collecting test data from activities such as range firing and computer modelling. While such testing might address environmental variables (rain, wind etc. — especially for Defence testing), it would highly be unusual for the underlying testing to address context of use.

133 McFarland (n 62) 27–28. See also Samuli Haataja, 'Autonomous Cyber Capabilities and the Attribution of State Responsibility — The Human Link', this volume ch 11, section II: 'it is the outcomes or effects caused by the ACCs that are most relevant'. The focus of weapons reviews on predictable effects limits concerns about code opaqueness, complexity and bias, and provides an

ability to act within legal standards will require data of a particular 'type, format and context' that permits a legal assessment of 'normal and expected use'. For the majority of ACCs, this will require them to be developed by governments (as opposed to private developers) who will have access to either the intelligence required to undertake specific cyber targeting, or the depth and extent of combat knowledge required to develop appropriate LOAC scenarios to test the code of the cyber and autonomous elements.

i Expanded Data-capture Points

The 'instrument' review approach for conventional weapons is characteristically structured around weapons procurement stages with front-ended processes for data gathering[134] and back-ended review decision-points.[135] Weapons review requirements are (intentionally) flexible, allowing for each State to determine, and adapt, how they will satisfy assessment criteria. This flexibility will be tested as the traditional procurement-review relationship may need to comprehensively adapt to address cyber capabilities with increasing autonomous functionality (especially those with adaptive capacity) to permit valid assessment.

Future cyber capabilities will more frequently be developed by government agencies or in private-public partnerships,[136] necessitating military participation in setting the design and performance specifications, along with developing and conducting relevant 'testing, evaluation, validation and verification' ('TEVV').[137] These processes in turn will drive expanded data capture requirements, and at earlier stages than data is normally

objectively assessable standard. It does not, however, address how such results would play out in the complex environment of cyberspace. Maintaining data about actual effects and comparing this against design or intended effect will be important, especially where there is doubt as to understandability of the algorithmic steps underpinning the effects.

134 Most empirical data and the aggregations of that data into objective measurements that can be assessed against specifications or capability gaps, is gathered prior to acceptance into service. While limited data is gathered after (partial) acceptance into service — the focus of these reviews remains the lawfulness of the weapon — and thus the data gathered is still focussed on the weapons design and initial specifications.

135 Whether through the normal acquisition process, or through any of the other less typical processes (operational procurement, documentation of military TTPs), the relevant stages provide important decision points for the conduct of interim and final weapons reviews. The overwhelming focus for weapons reviews is on data produced by producers in the front end of the process (that is prior to or as a part of test and evaluation). Clearly there is scope for an element of in-service testing by the State militaries, but it is not a large source of data and statistics for most States in conducting weapons reviews.

136 Traditional weapons reviews focus on the legality of the weapon under the basic weapons law criteria of unnecessary suffering or indiscriminacy. Context, where included, is usually limited to extrapolating test data to 'normal and expected use'.

137 For instance, Michèle Flournoy, Avril Haines and Gabrielle Chefitz, 'Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation and Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems' (WestExec Advisors, October 2020) <https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf>.

required by the State.[138] Unfortunately, this data will in many cases be vastly different from previously gathered empirical data. Conventional kinetic capabilities permit derivation of data from observable direct effects under controlled testing arrangements. Primarily non-kinetic capabilities, such as cyber, however, yield less observable data from primary and tertiary effects. This is due to the largely unobservable direct effect on '0s' and '1s', and the lack of interconnectedness in controlled virtual environments. The inclusion of an autonomy element further alters data requirements. For instance, it will require production of additional reviewable data, specifically contextual data,[139] to assess the operation of its algorithms against design specifications and standards as part of a 'use' review. Furthermore, where the autonomy is driven by adaptive capacity (whether internal parameters or processes), data will be required to both permit assessment of ongoing compliance and to train against.[140]

One approach historically adopted by States where a new weapon or means has insufficient empirical test and evaluation data,[141] is to undertake modelling or testing (simulated or physical) to develop necessary data to allow a weapons review to be undertaken. For autonomous technologies, the extrapolation of the traditional review approach in this or similar ways can supplement or replace instrument or effects data.[142] For instance, Copeland and Reynoldson, building upon processes for conventional and unconventional weapons TEVV, as well as exercise and operational employment cycles, have proposed an 'IHL testing and learning loop', consisting of a set of iterative but repeatable steps to assess 'nature and nurture' LOAC compliance by autonomous technologies in their 'normal and expected' use.[143] A version of this testing loop, directed to ACC, is provided in Figure 10.1.

138 For (autonomous) cyber capabilities developed by private industry the TEVV requirements (including predictability and performance standards) could be established within tender and contractual documentation. To ensure that performance and confidence acceptance standards are set appropriately (for compliance with State legal obligations) it will still require extensive interaction in the early development stages, and potentially development of sufficiently large and representative (hygienic) data sets.

139 The functionality of autonomy algorithms requires greater contextual data to address performance standards relating to LOAC decision-implementation functions. It is hard to gather data from testing that simulates operational context outside of actual military use in classified simulation, live fires, exercises, or operations.

140 Margulies (n 131).

141 Where a weapon is modified after acquisition for example the empirical data on accuracy may no longer be sufficient. In such situations States could look at testing the item through simulated (where relevant), blank and live fire practices to develop its own 'empirical' data.

142 Instrument data relates to how the weapon or means works. Effects data relates to the effects from the working of the instrument in normal and expected use.

143 Damian Copeland and Luke Reynoldson, 'How to Avoid "Summoning the Demon": The Legal Review of Weapons with Artificial Intelligence' (2017) Pandora's Box 97, 106–8. This permits testing to go beyond the original code design ('nature') of an ACC and assess it after it has been

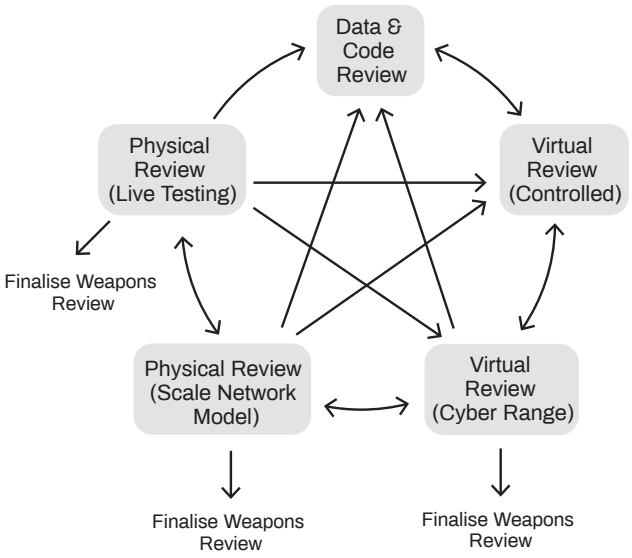

Figure 10.1 — Autonomous Cyber Capability Testing Loop

The first step of the testing loop would review the coded LOAC components of the cyber capability (a nature review),[144] and, where relevant, the data/scenario diet (a nurture review).[145] This step should not typically entail reviewing all of the code but would require confirming that a State's legal obligations can be met.[146] Code review could include software engineering processes,[147] manual confirmation of accuracy of coding of

provided extra data and scenarios ('nurture'). Specifically, the LOAC testing and learning loop would assess the ability of the weapon to comply with the relevant LOAC rules or obligations in the context of its 'normal or expected use' scenarios. These scenarios would be designed to replicate the weapon's normal or expected tasks and the LOAC decisions that it may be required to make in the anticipated operational and environmental scenarios.

144 ibid 106. This would emulate Copeland's Code review phase in which the initial programming is reviewed to ensure that it identifies the individual LOAC rules and principles and how they are to be applied in the context of the weapon's normal or expected use. The programming of this rule would require the State's position on standards of compliance for rules to be codified.

145 This would involve looking at the testing (and training) data for completeness, accuracy, provenance and hygiene (such as lack of biases and cleanliness of data). It would also consider the qualitative elements of the training such as the types of tasks, environment, interaction with other entities, training reality etc.

146 Code for autonomous technologies may run into millions of lines which would make it impossible to manually review. Similarly, where the autonomous technology is acquired from another State or a commercial developer the code may be inaccessible. In these situations there a number of options available to the State — rely solely on the effects of the autonomous technology (may require higher probability confidence bounds); seek a review and assurances of the LOAC components by the State or developer; or if have access to the code — apply an autonomous review capability.

147 Appropriate engineering processes include those to eliminate bugs, statistical accuracy checks, automated regression testing of software etc.). See, eg, Defense Science Board, 'Summer Study on Autonomy' (US Department of Defense, June 2016) <https://fas.org/irp/agency/dod/dsb/autonomy-ss.pdf>. It would behove any developer of ACC to ensure that their states LOAC obligations are not only understood but are correctly codified — this may require educating programmers on LOAC.

relevant standards, independent operation and performance assessment algorithms,[148] data assessment, and software bounding measures (coded parameter locks and human oversight capability) on the payload.[149]

The second step is the controlled testing of the ACC, or aspects of the ACC, within a contained environment such as a standalone or virtual computer or network, which replicates the targeted device.[150] Simulating the target device permits assessment of the design purpose of the ACC, focussing on the interaction of the control components (of the cyber means through autonomy and of the targeted device by cyber communications) and the predictability of the effects. An ACC would either achieve sufficient level of predictability of effects in this controlled environment and move to the next step or require remediation and regress to the code review step.[151]

The third step is controlled testing within a simulated virtual domain.[152] Using a cyber range will provide a simulation of aspects of the networks that the technology will operate through and permit the controlled introduction of variables to assess how external interaction affects operation. This broader simulation allows assessment of the predictability of the design purpose of the ACC, on the targeted device and any consequences for connected computers or networks. This step could provide States with sufficient information to make, at least, initial assessments of an ACC's ability to comply with LOAC, and identify whether it possesses sufficient levels of predictability of effects in this virtual environment to move to the next step.[153]

148 Such as the inclusion of: statistical checking; modification notification procedures or automated triggers for manual reviews; management procedures (differentiation, and discrete management learning) to allow for regular monitoring; and ongoing assessment of ACC.

149 Copeland and Reynoldson (n 143) 106; this involves effectively review of the actual code regarding LOAC decision-making to assess LOAC compliance. For more information on the issues relating to 'coding' an identification standard, see Backstrom and Henderson (n 103) 495–6. This step does not include a measure for assessing understandability/explainability/ transparency outside of predictability of effects at present. These concepts are not a legal or performance standard albeit States should consider the benefits of doing so.

150 Copeland and Reynoldson (n 143) 107: With step 2 this emulates an aspect of the 'Virtual review' phase, where the individual LOAC tasks or decisions are represented within a virtual environment that simulates the normal or expected use of the weapon in the likely operating environments across the range of possible scenarios that the weapon may be employed. Unlike Copeland and Reynoldson, the virtual phase is broken into two steps to recognise the complexity of the interaction between devices in the cyber domain and the exponentially greater levels of connectivity when compared to physical autonomous weapon systems.

151 Where a cyber device successfully navigates this step, the scenarios and responses of the ACC could be utilised as part of the library of scenarios and information that are utilised to aid any adaptive capacity in the ACC. If the ACC fails this step, then the testing simulation would have limited capacity to 'inform and refine' the ability of the ACC to implement command intent.

152 Copeland and Reynoldson (n 143) 107: With step 3 this expands the 'Virtual review' phase. The virtual environment could start with a discrete simulator and expand to a full cyber range involving other actors. The ACC would interact with the virtual environment to make decisions that can be assessed.

153 ibid: In accordance with Copeland and Reynoldson's virtual review phase this would permit assessment of ability of the autonomous element to comply with LOAC in its

The fourth step is controlled testing within a simulated virtual and physical domain.[154] Using a scale model of the physical domain integrated with a cyber range emulating the virtual domain will allow for testing response of the technology (propagation, exploit, and payload) with controlled variables. This step will not always be possible, but for important targets it permits greater assurance of the predictability of the design purpose or intended effects of the ACC, whether on the targeted device, on any connected computers or networks, or regarding the physical consequences for any connected systems.

The fifth step is testing of the technology within a 'live' environment.[155] While not always appropriate or possible, this form of testing assesses the operation of the technology in a live environment with uncontrolled/unforeseen variables. This step can be used to test aspects of the ACC in isolation (for example, propagation) or combination (propagation, exploit, and components of the payload but likely without a harming mechanism) to assess specific or general effects of the ACC in an uncontrolled environment.

Putting it simply, a weapons review would require sufficient data from testing — virtual, physical or a combination thereof — to identify whether the actual operation of the ACC replicates design specifications, but also stays within design parameters for autonomous functionality. A testing loop of a type proposed above is one way of constructing sufficient quantity and quality of data to permit a review of an ACC. It would permit the early identification of the limitations in the capabilities' LOAC decision-implementation ability, necessitating a modification of 'normal or expected use' or supporting the imposition of operational limitations

decision-implementation through: (a) capability 'of completing to the standard required by the State'; (b) capability of completing to the standard required by the State, but 'achieved through either programming restrictions or though the requirement for human input' to mitigate any identified limitations. Alternatively, the ACC could be assessed at this step of being unable to achieve the necessary standards (or effects) and is therefore incapable of passing the weapons review without alteration to the autonomy or cyber elements of the ACC.

154 ibid 107–8: This partially emulates the 'Physical review phase in a controlled environment' where the ability of the operating system to use the data is assessed against an autonomous weapon system's actual sensors and hardware. The same LOAC tasks or decisions that would be assessed in the controlled physical environment would be consistent with those likely to occur in operating environments and scenarios, although without complexities such as live adversaries. This process may reveal deficiencies with the weapon's sensors that require remediation. It may also demonstrate that, despite the results of the Virtual review phase, the weapon's sensors are unable to provide the AI operating system with the necessary data to make a lawful LOAC decision to the standard identified by the State.

155 ibid: This partially emulates the 'Physical review phase (uncontrolled environment)' which is designed to test the weapon's ability to make lawful LOAC decisions in uncontrolled circumstances that have not been foreseen. This review phase would coincide with the operational TEVV process that would be undertaken by the State's capability development organisation responsible for the procurement of the weapon. The aim of this phase is to test the LOAC decision-implementation in circumstances that will be as close to realistic operational environments as possible, including deliberate interference replicating adversary actions.

if required.[156] Alternatively, it may detect the requirement to return the ACC to earlier steps for redevelopment/reprogramming or retraining. Furthermore, it also provides additional data to enhance the technologies decision-implementation.

ii Expanded Review Points

The nature of most existing cyber capabilities — undertaking combat functionality beyond the point of direct human manipulation — involves a level of autonomy. An ACC with autonomy that only lightly touches on targeting law will, with some test and adjustment, be able to be undertaken under traditional weapons review decision-points. Cyber capabilities with greater levels of autonomy (such as ability to choose targets, assess collateral damage concerns, or adaptive capacity) will not fit as neatly within the decision-points of traditional weapons review approaches. These decision-points are derived from State practice and are not prescribed by international law, so it is open to States to determine as a matter of policy whether they conduct a series of progressive reviews, a single final review, or a combination of both.[157] While a few States that align their weapons review process with their capability development lifecycle undertake progressive reviews, most States only undertake a final weapons review.[158] Increasing autonomy will delay or add extra decision-points for States to assess lawfulness of ACCs potentially resulting in three distinct decision-points: traditional instrument and de novo use assessment against specifications;[159] ongoing assessment against testing/training performance;[160] and ongoing testing against operational performance.[161]

156 These could range from modification of its 'normal and expected' use, pre-operation limitations (such as coded blocks, higher standards for operation, narrow confidence bounds, or reductions to permitted use) or during operation limitations (such as re-review requirements, command restrictions on use, ROE restrictions, geographical boundary/network limitations, deconfliction measures).

157 For example, States Parties to AP I may interpret the wording of art 36 as describing the individual stages in the weapon procurement process as requiring a legal review at each stage. Such an interpretation would require a series of interim reviews that inform decision making as to the further procurement. Alternatively, State Parties may regard art 36 as stipulating alternative points at which the legal review obligation may be fulfilled with the ultimate objective to complete the review prior to the use of the weapon in armed conflict. Non-AP I States merely need to undertake a final weapons review prior to fielding.

158 ICRC (n 35) 23–4.

159 Addressing specification accuracy — technical specification includes correctly worded LOAC tests; design accuracy — is converted accurately into code; manufacturing accuracy — confidence acceptability levels.

160 Where accuracy improves from experience then testing would involve methods such as using the simulation in Copeland and Reynoldson (n 143), uncontrolled physical and controlled physical testing. The outcome of this phase would be a final art 36 review approving (or not) operational capability. The handling of ongoing monitoring for substantive adaptation and therefore possibly review should also be addressed.

161 Where more than de minimis adaptation of the ACC algorithms occurs then a de novo review

Application of traditional review processes. The first decision-point is concomitant with traditional review processes. The ACC would be assessed — in the form of a de novo 'instrument' and 'use' review — during and at the conclusion of capability development. While this could potentially be undertaken as a single final review, certain ACC would benefit from both inclusion of LOAC as a design criteria, as well as multiple interim reviews (whether for simplicity[162] or to avoid black-box irrevocability)[163] during capability development to ensure that legal and performance standards are correctable coded into the capability.[164] Complexity and opaqueness of algorithms driving autonomy and cyber capabilities will increasingly diminish the ability for line-by-line code review by humans. This will result in a requirement for bespoke measures (such as software tools to check for compliance or red flags),[165] to undertake weapons review assessment.

Ongoing assessment against testing/training performance. Adaptive capabilities will not only potentially create a substantive adaptation of the operation of the ACC but will also likely result in a commensurate requirement for extra data to assess capability performance.[166] It is expected that States will undertake these 'training' events at a point after weapons review decision-points traditionally occur. Consequently, a supplementary weapons review assessment may be necessary for any significant State input (data, training, simulation) required to achieve

would be required. The refinement of ACC parameters/criteria or results may be the intended reason for utilising desired to improve ACC operations through experience, or reducing unnecessary or unwanted deviation, but at what stage does it become a new weapon? Is it adding new criteria, removing new criteria or altering existing criteria? Is its structural change to the algorithms or the criteria they apply to, or is it change to the predictable effects? Policy questions would need to be addressed regarding what constitutes 'de minimis'.

162 States will need to determine whether their software development or modification permits multiple interim reviews, i.e. during research or concept development, code development which will likely be broad in application, and followed by a final review; or just undertake a final review.

163 For instance the code may be 'black-boxed' due to State producer export or security restrictions, the code may be too complex or extensive (especially for humans in the case of stochastic versus deterministic coding) to review, or the code may constantly being modified.

164 De Tomas Colatin and Väljataga (n 83) at 12: Where the new capability is being developed or acquired (noting the more straight-forward procurement of off-the-shelf capability is unlikely for cyber capabilities) States will need to ensure the contractual relationship addresses the new capability requirements.

165 Specific code review measures may include incorporating modification notification procedures, automated triggers for manual reviews, or even automated weapons review. These measures could be effected by: use of AI and statistical accuracy review software to audit/analyse algorithms — such as regression analysis; assessment of software engineering processes to review, assess and correct coding; inclusion of logging and audit system review capabilities in or with the autonomous means or capability; assessment of management review procedures including any regular monitoring capability or other ongoing assessments; and assessment of software locks or bounding where autonomous means or capability has a payload on board.

166 Traditional measures are unlikely to satisfactorily identify the 'predictable effect' from using an ACC — especially those designed for single use such as Stuxnet — where actual effect cannot be replicated. In such situations greater reliance will need to be placed upon the extrapolation of traditional approaches (such as modelling and testing processes).

operational capability beyond initial acceptance processes,[167] or for ongoing monitoring to ensure code or performance variation from testing does not result in significant modification of initial specifications.[168]

Ongoing assessment against operational performance. While most ACCs are expended after a specific use, some could potentially be multi-use. Such ACCs will require continuous monitoring to ensure modification (whether as code repair or enhancement) of the integrated autonomous functionality, does not require de novo assessment (see Part 3.2.(b)).[169]

iii Consequence of Changes to Assessment Data

A simple timeline representing the potential issues for data gathering and decision-points — see Figure 10.2 — highlights how the additional measures required to review cyber technologies with autonomous combat functionality will not only increase the complexity, duration and cost of the weapons review process,[170] but signal the distinct possibility of a continuous review process.[171] To address this, States may want to consider a formal re-review process after an ACC has been authorised for use in armed conflict, or expansion of the operational legal review ('OLR'). This would capture issues with code adaptation and predictability of operational performance, and in so doing address potential changes in methodology through new TTPs that are otherwise unforeseen and do not manifest during the traditional legal review process.[172] That is, to ensure

167 This would require development of specific training (and testing regimes) to manage the human-capability interface to address interaction between the human and the ACC.

168 This will necessitate policy discussion on whether modification requiring review is premised on changes to structure, how a device works or operates (i.e. coding and algorithms changes), or changes in results/effects.

169 For instance, modification that are more than de minimis (whether from alterations to algorithmic structure or parameters, or results from algorithmic operations) would trigger the need for a further review.

170 Admittedly, many of the positions identified above would need to be undertaken during appropriate test and evaluation (and software verification and validation), operational evaluation processes, or during typical military exercise and training cycles.

171 Eric Talbot Jensen, 'Precautions and Autonomy in the Law of Armed Conflict', this volume, ch 9, section IV; Copeland and Reynoldson (n 143) 108. For these reasons, commentators have argued that for weapons with autonomy provided by an adaptive capacity that 'legal review obligation remains during the life of the weapon system'. Furthermore, as most review data will be produced after the traditional point where initial acceptance occurs, States will need to extend the acquisition arrangement with manufacturers/developers past initial acceptance, dramatically increase its T&E capabilities, or a combination of both. Where states use commercial providers to produce cyber technologies, they will need to consider greater civilian-military integration to address classification, access to intelligence, and other issues that attach to relationships requiring greater understanding of the internal workings of the military.

172 For example, an ACC may utilise a specific exploit to gain access to a target cyber network. The ACC programming then locates its target within the network and triggers its payload to achieve the intended affect. These components would be the subject of a weapons review. If this ACC was modified, for example by updating the exploit, but it still achieves the same effect on the same target network a reviewing State may not considered it 'new' for the purpose of an art 36 weapons review. However, if the ACC's method of propagation, harming implement were modified or its intended target were changed, it would require further Weapon Review analysis.

substantive self-improvements (adaptations, changes to predictability of results) by the autonomous element are assessed in accordance with the requirements of LOAC, the weapons review will likely morph into a more fluid and informal ongoing process as opposed to a static discrete early cycle review evaluation.[173]

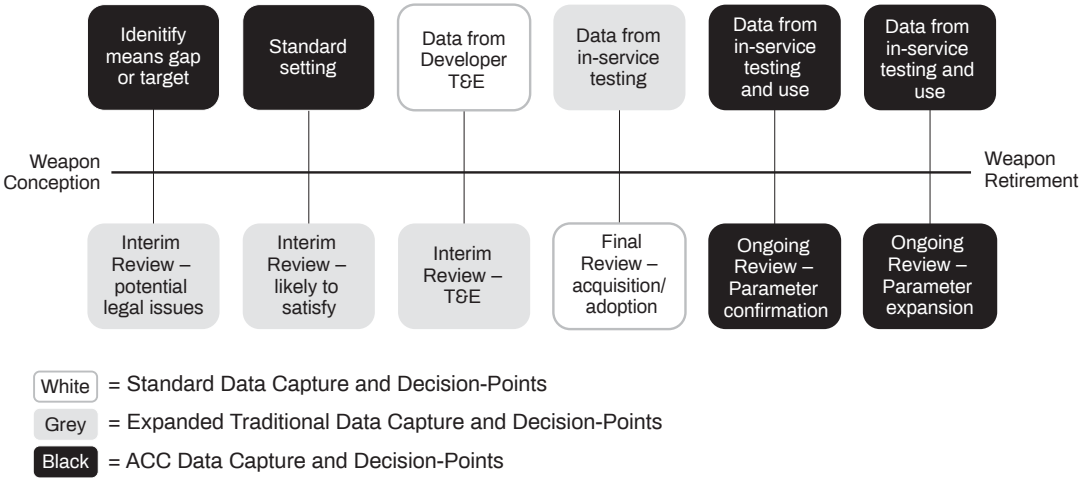

Figure 10.2 — Data Capture and Decision-Points: Traditional vs ACC

2 *Superfluous Injury or Unnecessary Suffering*

Noting the ACC-specific issues relevant to its assessment under general law were discussed in section 1 above, the 'general law component of the ACC review must also include an assessment of the basic IHL principles. The first 'instrument' review assessment under the 'general law' test for consideration in a weapons review is whether a capability', in its 'normal or expected use',[174] can be employed without causing superfluous injury or unnecessary suffering.[175] Most States apply this test as a balancing of military necessity[176] against the injury or suffering caused to

173 ICRC (n 35) 10 provides that 'an existing weapon that is modified in a way that alters its function, or a weapon that has already passed a legal review but that is subsequently modified' is covered by the material scope of art 36

174 Parks (n 12) 99 summarises this position as '[w]hile Article 36 requires States Parties 'to determine whether its employment would, in some or all circumstances, be prohibited' by the prohibition of means of a nature to cause superfluous injury or unnecessary suffering, the ICRC has acknowledged that 'some or all circumstances' refers to its 'normal or expected use'.

175 AP I art 35(2) and customary international law. Does the design, modification or employment of the capability case superfluous injury or unnecessary suffering. Unnecessary suffering means 'a harm greater than that unavoidable to achieve legitimate military objectives'.

176 See *Conference of Government Experts on the Use of Certain Conventional Weapons* (ICRC 1975) <https://www.loc.gov/rr/frd/Military_Law/pdf/RC-conf-experts-1974.pdf> 9, for confirmation on the use of military necessity. For most States, military necessity does not negate a specific treaty prohibition.

combatants by the capability.[177] That is, when analysing a weapon, means or method, the reviewer must assess the military utility of the weapon balanced against the harm (wounding and incidental effects) that the weapon is likely to cause to combatants.[178] Additionally, States will give appropriate consideration to:

- existing (lawful) capabilities that provide the same or a similar military advantage;[179]
- whether the capability has any secondary effects; and
- where the new capability injures by non-traditional means, or injures with qualitative and quantitative differences to existing lawful capabilities — including relevant medical criteria relating to the impact upon the target (i.e. foreseeable effects, mechanism of injury, expected mortality rates, expected permanent impairment, and treatability).[180]

There are a couple of unique considerations in the application of this principle to ACC. First, the nature of current cyber activities indicate that humans are not normally the target (even indirectly). Consequently, even though there is a requirement to review the ACC's 'normal or expected use' in accordance with the conventional instrument review requirements regarding unnecessary suffering, it would be unusual for this principle to be engaged because there is unlikely to be harm to combatants let alone unnecessary harm.[181]

177 While it is noted that the art 35(2) language ('weapons, projectiles and material and methods of warfare') is arguably narrower in application then that of art 36 ('new weapon, means or method of warfare'), most AP I States who undertake weapons reviews apply the broader requirements of art 36 when determining compliance with the prohibition against superfluous injury and unnecessary suffering.

178 The United Kingdom unsurprisingly comes to a very similar position. UK Ministry of Defence, *The Manual of the Law of Armed Conflict* (Oxford University Press 2004) 103, [6.1.2].

179 Parks (n 12) 125 and 133: '[A]n increase [in] injury as such to combatants may not necessarily lead to the conclusion that it constitutes superfluous injury. However, it is unlikely increased suffering without some legitimate military necessity, such as increased range or improved accuracy, would be legally defensible.'

180 Nicholas Tsagourias and Giacoma Biggio, 'The Regulation of Cyber Weapons' in Eric Myer and Thilo Marauhn (eds), *Research Handbook on Arms Control Law* (Elgar, forthcoming in 2021) 8, which identifies that a weapons review needs to consider objective factors such as science, medicine and health as well as subjective factors such as military advantage.

181 It is entirely possible to imagine cyber operations that physically (or psychologically) harm combatants, the reality is that nearly every publicly recorded cyber 'attack' does not involve physical or mental harm to humans. That being said, the potential demonstrated by some 'attacks' (such as the 2015 Ukrenergo attack that resulted in extensive loss of power in Ukraine; or the NotPetya attack that affected amongst other things hospital schedules and data) highlights the real risk from cyber activities.

Second, where the autonomous component includes some form of discretion regarding decision-implementation, a 'use' review will need to be undertaken, including potential consideration of the ability to apply or comply with a greater range of LOAC requirements. For the principle of unnecessary suffering this 'use' review would entail assessing every 'normal or expected' exercise of combat functionality by the autonomy algorithms to determine whether they can apply the principle correctly, or otherwise avoid implementing decisions that are contrary to it.

Third, where this principle is engaged for an ACC, the use of the ACC with decision-implementation capability will create a novel review requirement to evaluate the ability of ACC to:[182]

- avoid disproportionate suffering (by including the ability to stop harming combatants once military necessity to target them ceases);[183] and

- avoid rendering death or permanent impairment inevitable (by either not commencing targeting individuals whose status has changed between when the ACC was given targeting approval and when the combatant is located or stopping an attack because the circumstances of the attack have changed).[184]

Under the principle of unnecessary suffering, the ability of an ACC to stop[185] applying force when military necessity culminates is just as important as traditional design criteria. This overlaps with, but is distinct from, the protection of persons *hors de combat* and the principle of distinction. As such, the weapon review will need to assess whether the ACC's algorithms can determine whether the military necessity to use force has reduced or stopped because the target no longer poses a threat, and any further use of force merely causes superfluous injury or unnecessary suffering.[186]

182 In this instance, the ACC would include decision-making capability and weapon, although typically you would expect a weapons platform, weapon and the decision-making capability to be combined.

183 Chengeta (n 73) 88–91 believes that an ACC that cannot make this decision could cause disproportionate suffering and hence breach the rule in art 35(2).

184 ibid 92–4 identifies that an AWS that cannot identify surrender or other changes of combatant status may result in the death of combatants being rendered inevitable. Clearly, this rule overlaps with distinction. Chengeta also argues that lack of predictability may lead to situations where permanent impairment is rendered inevitable (albeit it is not clear how he links the concepts).

185 ibid 89.

186 Consideration of superfluous injury or unnecessary suffering may be relevant to the Weapon Review of Stuxnet styled ACC designed to attack an enemy computer network controlling dangerous materials such as a munitions in a storage facility. An attack may target the facility's power supply resulting in the physical destruction of munitions and injury to or death

3 *Indiscriminacy*

The second assessment for an 'instrument' review under the 'general law' test is whether the capability, in its 'normal or expected use', is capable of being used discriminately.[187] That is, the capability, and its effects, in light of its 'normal or expected use', must be capable of being controlled and directed at a distinct legitimate military target. Two elements must therefore be assessed: whether the capability can be directed against a specific military objective — in other words accuracy in target identification;[188] and whether the effects[189] of a capability can be controlled, or otherwise limited to legitimate military targets.[190] Factors considered in reviewing this principle will include accuracy, type, duration, and extent of effects.

The inclusion of a decision-implementation capability creates new and additional challenges to reviewing the prohibition on indiscriminate weapons. For the purposes of determining capability to control and direct the harming mechanism, and any reasonably expected effects as per the rule discussed above, the platform and associated weapons would be reviewed in accordance with extant weapons review practices. The incorporated decision-implementation component might require the reviewer to consider the ability of the cyber technology to direct or control the harming mechanism it utilises. As such the ACC's autonomy algorithms would need to comply with AP I Article 51(4)(a)–(c), as well as the various distinction requirements contained within AP I and customary international law. The 'use' review required here — the extent to which the ACC's exercise of combat functionality engages the principle of distinction — would be dependent upon its 'normal and expected use'.[191]

There are examples of systems already in use that apply the principle of distinction (acting as a decision-implementation instrument). Such systems, to varying degrees, operate within defined and restrictive

of personnel located at the facility. A Weapon Review will consider whether the munition destruction will cause suffering and, if so, whether that suffering is disproportionate to the military advantage to be gained.

187 Under weapons law — art 51(4) of AP I and customary international law — weapons, means and methods that are indiscriminate are prohibited.

188 AP I art 57(4)(b).

189 The effects of the weapon on civilians and civilian objects must be considered not only at the time of an attack, but also for a period after the attack. Effects such as explosive remnants of war may not of themselves indicate that a weapon is indiscriminate, but they can create other obligations for a State.

190 AP I art 51(4)(b) and (c).

191 For example, an ACC designed to exploit a software flaw, common to all computers operating a certain operating system, should be capable of distinguishing between targeted computers and civilian computers for payload execution, i.e., whether exploiting a common software flaw to gain access to a computer system would be indiscriminate if the ACC payload only effects those systems identified as lawful targets.

parameters (such as target set, domain, controls on automated/autonomous options) resulting in a requirement to assess compliance with a limited aspect of the principle of distinction.

4 *Proportionality*

A conventional weapons review process does not usually assess the legality of the capability against the principle of proportionality as this, for all intents and purposes, is a context-specific assessment applied by humans as part of the targeting or attack process.[192] Where proportionality has been included in the weapons review process, the principle has been restricted to acting as a compliance measure to identify an indiscriminate weapon rather than a 'feasible precaution'.[193] That is, a capability is assessed to determine whether it is indiscriminate because it is disproportionate (expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated). In other words, it is assessing if the effects of the use of the capability is not unreasonable or excessive,[194] and thus complies with the principle of proportionality.

Depending upon the 'normal and expected use' of the ACC, it is possible that the weapons review will need to assess the legality of the ACC's autonomy algorithms to implement human decisions in compliance with the principle of proportionality. There are currently no algorithms that can undertake proportionality analysis like humans (and thus to the full extent of LOAC). That is not to say that aspects of the proportionality assessment are not already being undertaken. For instance, the Digital Precision Strike Suite Collateral Damage Estimation tool ('DCiDE'), similar applications within the Joint Automated Deep Control System ('JADOCS'),[195] and a variety of other targeting tools, can calculate the

192 US Department of Defense (n 21) 330. 'The law of war rules on conducting attacks (such as the rules relating to discrimination and proportionality) impose obligations on persons. These rules do not impose obligations on the weapons themselves; of course, an inanimate object could not assume an "obligation" in any event… The law of war does not require weapons to make legal determinations, even if the weapon (e.g., through computers, software, and sensors) may be characterized as capable of making factual determinations, such as whether to fire the weapon or to select and engage a target.'

193 Margulies (n 131) section III.C.

194 US Department of Defense (n 21) 60–1, 240–8, and 1004–5.

195 Ratheon, 'Joint Automated Deep Operations Coordination System (JADOCS)' <https://www.raytheon.com/capabilities/products/joint-automated-deep-operations-coordination-system-jadocs> accessed 2 January 2020. JADOCS (or rather applications within JADOCS) provides detailed information about each target, including capabilities to ensure fire support personnel positively identify targets, estimate collateral damage, avoid fratricide and assess battle damage post-strike. Raytheon description of JADOCS is as an extant system that provides amongst other capabilities 'situational awareness, targeting and fires coordination tool, interfaces with intelligence and fire direction systems, providing the "glue" for decision-making'.

effects of conventional munitions, undertake basic identification and differentiation between military and civilian objects, as well as create, on the basis of objective defined parameters, collateral damage estimates. Such systems usually support command decision-making, and are primarily focussed on collateral damage to humans,[196] but there is no reason a non-kinetic targeting tool could not be developed to support ACC in undertaking aspects of the proportionality estimate.[197]

5 *Environment*

An Article 36 weapons review process will assess whether a capability is prohibited because it is designed or expected to cause widespread, long-term and severe damage to the natural environment.[198] This will require consideration of the ACC's 'normal or expected use', choice of harming mechanism, and its method of operation to determine whether any or all of these elements will engage this prohibition.[199] An ACC with the potential to engage the prohibition may require specific programming of the prohibition into its autonomy algorithms. That is, because it is a specific prohibition, which is not necessarily addressed by any of the general principles of LOAC (military necessity, unnecessary suffering, discrimination and proportionality), the prohibited technique or practice must be coded into the autonomy algorithms as a red line, to ensure that the ACC does not breach this prohibition.[200]

6 *Other LOAC Considerations (Including Precautions in Attack)*

A weapons review of a human-operated capability would not usually consider precautions in attack and other LOAC considerations relating to the use of force as such reviews typically address the legality of that use

196 Traditional collateral damage tools — appropriately — focus on the risk to civilians. States (see, e.g., US Department of Defense (n 21) 1004–5) and academics have raised the potential for collateral effects estimates or tools to assess the risk to civilian infrastructure from cyber capabilities. Given the complexity and secrecy of such a tool it is unlikely to be made public.

197 The Weapon Review of an ACC designed to attack an enemy computer system by a DDoS attack may raise proportionality concerns if the ACC's autonomy algorithms allow the DDoS to cause damage to civilian computer systems that is disproportionate to the anticipated military advantage. In such a case, the Weapon Review could require modifications to the ACC programming or restrict its use to circumstances where the proportionality risk is reduced (e.g., by targeting a closed computer system).

198 API arts 35(3) and 55(1). Convention on the Prohibition of Military and any other Hostile Use of Environmental Modification Techniques (adopted 18 May 1977, entered into force 5 October 1978) 1108 UNTS 151 art 1 prohibits parties from the hostile use of environmental modification techniques having widespread, long-lasting or severe effects as the means of destruction, damage or injury. See also McCormack (n 88) 36.

199 Tsagourias and Biggio (n 180) 11 make the intriguing argument that a cyber capability that uses a targeted system to cause the damage or harm may not be caught by AP I art 35(3).

200 For example, an ACC designed to attack enemy infrastructure such as oil reserves or nuclear power facilities must be assessed to ensure the risk of causing damage to the natural environment from damaging releases is eliminated.

of force (focussing on the actions of the decision-maker) rather than the lawfulness of a capability. These are typically considered as part of OLRs or tactical 'use' reviews.

The inclusion of the autonomy algorithms, to affect components of what would traditionally be human decision-making on the use of force, requires weapons reviews to consider the legality of the ACC in undertaking use of force decision-implementation.[201] For the purposes of the weapons review it is irrelevant whether the autonomy algorithms are deterministic or non-deterministic. What is relevant is that the ACC is performing a method of warfare — determining how a weapon will be employed or making decisions on attacks and other activities designed to adversely affect the enemy's military operations or capacity (not just selecting or engaging targets). To assess the legality of the instrument, its ability to undertake this function in accordance with LOAC must therefore be assessed.[202] The starting point in conducting this assessment would be AP I, and in particular Article 57 ('Precautions in attack').

The principles relevant to 'precautions in attack' guide planners, commanders and end-users of capabilities in how they can utilise them in conducting attacks.[203] For the purposes of a weapons review, it will need to be determined whether the autonomy algorithms in their 'normal or expected use' implement decisions that are covered by these rules on attack (and the constant care obligation).[204] If so, the reviewer will need to analyse the ability of the autonomy algorithms to participate in the conduct of attacks within the requirements of Article 57.

Included within the precautions in attack is reference to 'special protection'. AP I and II (as well as other LOAC treaties and customary international law) provide rules that give certain objects and people special protection, that restrict them from being attacked or require certain processes to be undertaken prior to an attack (or military operation) being conducted against that object or person. Where these specific protections are not covered by prohibitions or restrictions under the general principles of LOAC (military necessity, unnecessary suffering, discrimination and proportionality), the specifically prohibited or restricted technique, practice or target may need to be coded into the autonomy algorithms

201 For a discussion on autonomy and the precautions in attack, see Jensen (n 171).
202 For a list of non-binding sources discussing the customary international law status of LOAC rules, see Michael Schmitt and Jeffrey Thurnher, '"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict' (2013) 4 Harvard National Security Journal Features, 231 fn 9.
203 Attack is defined in AP I art 49(1). AP I art 57(1) provides guidance on conducting military operations more generally, noting there is a requirement to exercise constant care to spare the civilian population, civilians and civilian objects.
204 Jensen (n 171) section III.A.

(especially those with adaptive capacity) as a 'red line' prohibiting certain actions. This reduces the risk that the ACC breaches the specific prohibitions or restriction under LOAC, and enhances its ability to comply with Article 57 of AP I.

Where relevant, these rules would need to be converted into code, and then be able to be applied in accordance with AP I or II, or customary law.[205] Of course, as noted above, the extent to which each rule needs to be addressed would be dependent upon the 'normal and expected use' of the ACC being assessed and linked to the environment in which is intended to be deployed.

PART 6 OTHER RELEVANT INTERNATIONAL LAW

Requirement. The Article 36 weapons review will evaluate new capabilities against any other applicable international law (relevant to conduct in armed conflict) binding upon a State, or that a State may choose to apply for policy reasons.

Analysis. This review would be undertaken on a case-by-case assessment for each ACC, however as noted at the start there is currently no specific treaty law dealing with the autonomous functionality of an ACC. While it is unusual for a weapons review to address other aspects of international law, the use of ACC raises some interesting international law factors for consideration. Three key considerations are residual code, neutrality and application of international human rights law (IHRL).

1 *Residual Code*

Stuxnet was highly discriminating in the effects it achieved — the alleged damage was limited to specific centrifuges at one geographical location (Natanz, Iran).[206] Stuxnet was however a worm that propagated around the world, like WannaCry and its successor Notpetya, resulting in its code infecting thousands of untargeted computers. While Stuxnet generally did not hamper the operation of the infected computers, the code continued to reside within them.[207] Residual code, whether in one or many computers, carries ongoing risks from re-use or adaptation for new uses. States will need to determine if the spread or remains of

205 This includes the safeguard responsibilities required under Article 57(2)(a)(i) or Article 57(2)(b) with respect to special protections.

206 Ziolkowski (n 61) 4.

207 Nicolas Falliere, Liam O Murchu and Eric Chien, 'W32.Stuxnet Dossier' (v1.4, Symantec Security Response, February 2011) 3 <https://archive.org/details/w32_stuxnet_dossier>.

an ACC are an effect that needs to be assessed in the absence of other collateral effects related to the ACC spreading to non-targeted devices or networks. If States identify that this is an effect, consideration will need to be given to how this affects legal compliance; in particular, the principles of distinction (based upon potential indiscriminate propagation) and proportionality (as a collateral effect). Potentially, this may result in a requirement to stop the spread of the ACC (irrespective of the intended effect(s)), or for the ACC to 'erase' itself or the damaging mechanism from non-targeted devices.[208] With the latter, while there is no general requirement at international law to remove remnants of attacks, the desire to negate specifically dangerous remnants has been demonstrated by the development of a specific Protocol to ensure States remove explosive remnants of war at the end of a conflict.[209]

2 *Law of Neutrality*

A conventional weapons review is unlikely to consider neutrality, but it may become an obligation for the weapons review of ACCs by States depending upon whether they classify a cyber capability as a means (weapon) or method (capability) of warfare. A cyber capability classified as a method of warfare does not engage the prohibition on moving munitions or supplies across neutral territory found in Article 2 of the 1907 Hague Convention V, but arguably falls under Article 8 of the same Convention, which permits communications to travel across neutral territory.[210] If, however, a cyber capability is classified as a means of warfare, consideration would need to be given to the implications of the 1907 Hague Convention V, if the 'normal or expected' use results in it moving (whether from infiltration or exfiltration) across neutral territory.

3 *International Human Rights Law*

LOAC as a *lex specialis* has traditionally precluded the application of IHRL,[211] but a more nuanced dialogue on the extent to which IHRL applies

208 Tsagourias and Biggio (n 180) 12 ask an interesting question as to whether there is a requirement that 'reviews should take into account the post use-life of a cyber weapon'. While the current position is likely no, there are a number of strong policy (i.e. not wanting your adversary to have your — used — code or capabilities) and legal (i.e. as an analogy of residual code or capabilities to explosive remnants of war or foreseeable harm) arguments in support of doing so. The strength of the argument will only increase as ACC become increasingly autonomous and their capability to cause damage or harm increases.

209 Protocol on Explosive Remnants of War (Protocol V) (adopted 28 November 2003, entered into force 12 November 2006) 2399 UNTS 100.

210 Hague Convention (V) Respecting the Rights and Duties of Neutral Powers and Persons in Case of War on Land (adopted 18 October 1907, entered into force 26 January 1910) 205 CTS 299.

211 Dale Stephens, 'Human Rights and Armed Conflict: The Advisory Opinion of the International Court of Justice in the Nuclear Weapons Case' (2001) 4 Yale Human Rights and Development Law

has displaced this absolute view.[212] Given the long duration of some cyber operations, and their close and sometimes inseparable relationship with espionage, it has been argued that weapons reviews should consider IHRL obligations.[213] Indeed ACCs will usually be prepared by civilians (noting that the skill and expertise to build ACCs will rarely reside in conventional combatants), operate on primarily civilian infrastructure, spread through civilian networks, and impact civilian computers, even when employed legitimately in armed conflict. That is, ACCs will operate primarily outside of the traditional reach of LOAC, and the conflict framework for which LOAC was originally contemplated to address, even while their effects may not. Conversely, the application of IHRL to cyberspace is contentious.[214] IHRL may therefore potentially apply to the consequential physical effects of ACC, but not their creation, use, or virtual effects. Unsettled as this may be, to the extent that States determine that it does, then it will need to be considered during this part of the weapons review to ascertain ACC compliance.

PART 7 PUBLIC INTEREST AND THE MARTENS CLAUSE

Requirement. The penultimate consideration in a weapons review is of the concepts of 'public interest' and the 'Martens Clause'.[215]

Analysis. 'Public interest', a concept which is often grouped and confused with the Martens Clause, relates to whether it is in the interests of a State (on behalf of the public) to study, develop, acquire or adopt a weapon, means or method of warfare. This is not a legal requirement *per se* and will often be addressed externally to the weapons review under

Journal 2.

212 *Nuclear Weapons Advisory Opinion* (n 16) [25], which identified that 'the protection of the International Covenant on Civil and Political Rights [ICCPR] does not cease in times of war, except by operation of Article 4 of the Covenant'. See also *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory* (Advisory Opinion) [2004] ICJ Rep 136, [102]–[106], which stated that that IHRL may be directly applied in situations of armed conflict; and *Armed Activities in the Territory of the Congo (Democratic Republic of Congo v Uganda)* [2005] ICJ Rep 168, where the court found IHRL must be adhered to.

213 De Tomas Colatin and Väljataga (n 83) 13. For a more complete discussion see Stuart Casey-Maslen, Neil Corney and Abi Dymond-Bass, 'The Review of Weapons under International Humanitarian Law and Human Rights Law' in Stuart Casey-Maslen (ed), *Weapons under International Human Rights Law* (Cambridge University Press 2014).

214 'NATO CCDCOE and INSCT joint workshop on "Human Rights in Cyberspace", 1st–2nd of October 2015, Tallinn, Estonia: Workshop Report (NATO CCDCOE, October 2015) 3.

215 While this requirement is probably more firmly established for AP I States, the concept has customary status. See, e.g., *Nuclear Weapons Advisory Opinion* (n 16) [87]; where the ICJ reflected upon both the customary nature of the Marten's Clause, and its effectiveness in 'addressing rapid evolution of military technology'.

policy or economic considerations. The focus of the public interest concept in the context of a weapons review is primarily 'future legal trends'. Effectively this means that future trends or developments in the law which are likely to affect the capability should be considered to avoid wasting scarce resources on an instrument or practice that will be made illegal or rendered militarily unnecessary.[216]

The 'Martens Clause' originates from the 1899 Hague Conventions,[217] with its modern codification found in Article 1(2) of AP I:

> In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from principles of humanity and from the dictates of public conscience.[218]

The Martens Clause is recognized by the ICJ as 'an effective means of addressing rapid evolution of military technology',[219] however its status is the subject of academic debate.[220] For the purposes of the clause, the 'principles of humanity' (or elementary considerations of humanity) are the 'fundamental general principles of humanitarian law'[221] which include the customary principles of distinction, proportionality, prohibition against unnecessary suffering, that the means and methods of warfare are not unlimited, and also extend to the protections contained under Common Article 3 to the Geneva Conventions.[222] The meaning of 'dictates of public conscience', for the purposes of the Martens Clause is a little less clear. Proponents have indicated that public opinion,[223]

216 Vincent Boulanin, 'Implementing Article 36 weapons reviews in the Light of Increasing Autonomy in Weapon Systems' (Stockholm International Peace Research Institute 2015) 6, referring to Swedish and UK practices. See also McCormack (n 88) 36 regarding future legal issues.

217 Hague Convention (II) with Respect to the Laws and Customs of War on Land of 1899 (entered into force 4 September 1899) 32 Stat 1803, TS No 403, preamble. The clause was created, in a large part, as a positive law political expedient to mollify the humanitarian concerns of small military powers against the interests of large military powers. See Antonio Cassesse, 'The Martens Clause: Half a Loaf or Simply Pie in the Sky?' (2000) 11(2) European Journal of International Law 187.

218 AP I art 1(2).

219 *Nuclear Weapons Advisory Opinion* (n 16) [78] and [84].

220 Christopher Greenwood, 'The Law of Weaponry at the Start of the New Millennium' in Michael N Schmitt, Leslie C Green (eds), *The Law of Armed Conflict: Into the next Millennium* (US Naval War College 1988) 206; Yoram Dinstein, *Conduct of Hostilities Under the Law of International Armed Conflict* (Cambridge University 2007) 57, citing Paul A Robblee, 'The Legitimacy of Modern Conventional Weaponry' (1976) 71 Military Law Review 95, 125.

221 *Corfu Channel Case (UK v Albania)* (Merits) [1949] ICJ Rep 4, 22.

222 Rupert Ticehurst, 'The Martens Clause and the Laws of Armed Conflict' (1997) 317 International Review of the Red Cross 125.

223 Theodor Meron, 'The Martens Clause, Principles of Humanity, and the Dictates of Public Conscience' (2000) 94(1) American Journal of International Law 78, 83 identifies that one perspective is to look at 'public opinion that shapes the conduct of parties to a conflict and promotes the development of international humanitarian law, including customary law'. This

legal communications from learned persons or organisations,[224] 'sources which speak with authority',[225] and *opinio juris*[226] are evidence of the public conscience.

Unfortunately, there is no unanimously accepted interpretation of the Martens Clause.[227] Within the plethora of interpretations there appear to be a spectrum of three main interpretive positions in practice.[228] The narrow position, supported by many States, is to restrict its application to preserving customary international law.[229] The centre position is that the clause is 'used to confirm or bolster the interpretation of other international rules of humanitarian law'.[230] The broad (and perhaps most controversial) position is that it elevates the principles of humanity and the dictates of public conscience to the level of independent sources of international law.[231]

The application of the Martens Clause and consideration of the public interest to reviews of ACC will vary according to a State's interpretation of the Clause. It would however be good practice as a matter of policy for any weapons review to identify if there is any information suggesting the:

presumably includes academic discourse, reports and commentary from the humanitarian community, international conferences, and the media.

224 Ticehurst (n 222) 2, referring to page 56 of the Nauru written submission to the *Nuclear Weapons Advisory Opinion* (n 16).

225 Ticehurst (n 222) 2, referring to the Judgement of Judge Shahabuddeen in the *Nuclear Weapons Advisory Opinion* (n 16).

226 Meron (n 223) 83.

227 Ticehurst (n 222). See also Cassesse (n 217) 189.

228 The clause has attracted a wide and varied range of interpretations and uses — many are logically weak and have few adherents. Cassesse (n 217) 189 notes that there are three main 'trends' in thinking re the application of the Martens Clause ('a contrario', creates two new sources of international law, the clause 'expresses notions that have motivated and inspired the development of international humanitarian law'. Ticehurst (n 222) 1 also suggests three interpretations: preservation of custom; an 'a contrario' position; and two new sources of law; Cassesse (n 217) 202. See also Tyler D Evans, 'At War with the Robots: Autonomous Weapon Systems and the Martens Clause' (2013) 41(3) Hofstra Law Review 723–5, art 8.

229 This position could range from merely recognising that customary international law continues to apply even after the adoption of a treaty, or extend to refuting the argument that what is not restricted by a treaty is permitted (that is, the clause acts only to reinforce the interpretive position at international law, that if a matter is not covered by one set of laws (for instance AP I), then it can still be covered by another set of laws (for instance or custom)). Cassesse (n 217) 189, 192–3: The rationale for including a clause adopting this latter interpretive position is to prevent 'a contrario' positions being taken. According to Schmitt and Thurnher (n 202) 27, it is not an 'overarching principle that must be considered in every case', rather it is safeguard against a 'lacunae in the law'. For the purposes of ACC there are arguments both for and against there being a lacunae in the law. With AP I addressing for international armed conflicts both the prohibition on weapons that are illegal per se, and also the covering the field on the use of weapons, in addition to the growing range of treaties addressing bans or restrictions on weaponry support, including the CCW process, the argument is that the Martens Clause is rapidly becoming, if not already, irrelevant to ACC. A contrary position on the facts has also been argued, that is the treaty law insufficiently recognises the 'novel' issues raised by ACC, such that the Martens Clause is applicable.

230 Cassesse (n 217): As such, under this position both the 'principles of humanity' and the 'dictates of public conscience' act to provide additional evidence to existing humanitarian laws, thus while not reaching the level of independent sources they do more than merely preserve customary international law.

231 That is, a new weapon, means or method of warfare would be illegal if the principles of humanity or the dictates of public conscience were breached, even if the instrument satisfied the relevant weapons and targeting law criteria. See Cassesse (n 217) 193; and also Evans (n 228).

- capability is abhorrent to the public conscience or it offends the principles of humanity; [232] and
- use of the capability is not in the public interest.

If such information is apparent then the weapons review should consider both criteria to determine if either will affect acceptance of the capability.[233] It is important to note that both of these matters require (to some extent) a non-legal policy position to be adopted by the State conducting the review.[234] Depending upon the nature and type of capability being assessed this position could be sourced from within the State's Defence Department or from a relevant government agency with portfolio responsibility for weapons law or legal/policy positions. For extremely complex or contentious matters, which could be relevant to future conceptions of particular ACC, a whole of government position would likely be sought.[235]

PART 8 **DOMESTIC LAW**

During a review, issues may be raised in relation to the application of a State's domestic law to the capability. The weapons review is not normally utilised to analyse use of the capability under domestic law with one significant exception. Domestic law prohibitions or limitations on the ability of the military to have the capability in its inventory, or on its subsequent use, must be addressed within the weapons review.[236]

232 This is the position taken by Human Rights Watch and International Human Rights Clinic, 'Losing Humanity: The Case Against Killer Robots' (19 November 2012) 36 and Christof Heyns, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions' (Human Rights Council, 9 April 2013) UN Doc A/HRC/23/47, in calling for a ban on LAWS. The argument can be flipped. For instance, Galliott and Scholz claim there is a moral imperative to develop weapons that are more compliant with LOAC. For example, weapons that are capable of averting attacks on protected symbols, protected sites and signals to surrender. See Jai Galliott and Jason Scholz, 'AI in Weapons: The Moral Imperative for Minimally-Just Autonomy' (2018) 1(2) Journal of Indo-Pacific Affairs 57.

233 Many States adopt a policy position that is closer to the broad position in practice, albeit this is not to be taken as acceptance of this as being its legal position on their interpretation of the Martens Clause.

234 See Parks (n 12) 130 fn 251 regarding weapons review being 'cognizant of trends in the law of war or arms control law'.

235 This could also include creating a public forum to allow the public and other interested parties an opportunity to comment on the issue.

236 Where the capability is likely to be used domestically by the military (to have an effect on an armed conflict), then the relevant domestic law should be considered.

PART 9 CONCLUSION

This part contains the weapons review conclusions and the associated legal advice decision. A weapons review decision will normally be one of three types of decision:

- provide legal advice of review clearance;
- provide legal advice of review clearance with conditions or limitations;[237] or
- provide legal advice that review clearance cannot be given.

Where the weapons review provides a conditional clearance (or no clearance), appropriate guidance can be provided to indicate what is necessary for the capability to achieve clearance.[238]

It is the authors' recommendation that a further aspect be added to address re-review requirements for any ACC that is reviewed. That is, the weapons review decision should clearly articulate:

- if an ACC requires re-review, or some form of certification that no re-review is required; and
- where a re-review is required a description of the anticipated requirements of that re-review including nature (formal weapons review or OLR), timing, triggers, duration, and authority.

237 Conditions can include bounding, re-training, or re-coding. Bounding covers external and internal measures that restrict the autonomy which permits the weapon to achieve intended effects that were not envisioned when adopting/acquiring/developing the cyber technology. Bounding options include restrictions on operating parameters, limiting operational use, deployment restrictions, human control requirements, geo-blockers etc. Re-training requires controlling data and scenario input to re-train the autonomous element. Re-coding means re-writing the code.

238 The conclusions of weapons reviews will inform the development and use of an ACC. Ultimately, the review report will not be a static assessment but will act as a marker for ongoing use and development of an ACC. It will provide a basis upon which to assess the requirement for a re-review and ongoing legality of the ACC as factors that affect its use or effect change.

III
CONCLUSION

A cyber capability, which is capable of undertaking tasks to cause damage to military objectives or harm to combatants without human interaction, is a novel technological capability that will challenge the weapons review processes of States. Cyber capabilities possess unique propagation, access/exploit, and variable payload capabilities not usually addressed in weapons reviews. Weapon autonomy permits a capability to algorithmically execute action within pre-determined (albeit potentially broad) parameters without (and potentially beyond) human intervention. In combination, ACCs reflect two lenses of control — control of the means or method (autonomy) and control of a target device (cyber), that warrant consideration of legitimacy prior to fielding in combat, rather than waiting until actual use. While existing practices can be lent upon to evolve existing weapons review approaches to accommodate ACCs, these reviews are unlikely to be simple. In fact, there is the prospect of significant complexity, and at times a certain amount of contention.

The unique elements of an ACC would require an individual weapons review to include a conventional 'instrument' review combined with a 'use' review. This is not the only change that will be required. The nature of the traditional review will need to be refined or adapted to address: the rise in in-house intelligence driven single use capabilities; expanded review timelines and review points (both earlier and later than traditional acquisition cycle review points); and variable review obligation entry and exit points. weapons reviews of ACC will also potentially require new standards (i.e. predictability, performance, and legal), review focussed software design, and the development and application of custom testing loops. Finally, given the potential difficulty in identifying the reviewable components (harming mechanisms) of ACC, especially where the autonomy algorithms permit adaptability or even re-writing, States may be required to develop ACC-specific testing and training regimes combined with increased (internal) transparency on data provenance and software standards to clearly demonstrate predictable effects stay within designed or expected parameters. Failure to do so may risk an ACC being limited in its use, or not being cleared for employment by a State.

International Legal Responsibility

Chapter 11

Autonomous Cyber Capabilities and Attribution in the Law of State Responsibility

Samuli Haataja

I
INTRODUCTION

This chapter considers attribution in the law of State responsibility for conduct involving the use of autonomous cyber capabilities (ACCs). It consist of four sections. The first section provides an overview of ACCs and the concerns they raise when used by States in cyber operations. The second section outlines the general rules on State responsibility as detailed in the in the International Law Commission's Articles on the Responsibility of States for Internationally Wrongful Acts, and in the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Section three then examines the rules surrounding attribution and their application to conduct involving the use of ACCs, and the extent to which autonomy problematizes these rules. Finally, section four considers the application of these rules in a possible future scenario where States have given software entities a degree of legal personality, and it examines the extent to

which this status can be used to avoid responsibility for the actions of ACCs. This chapter concludes that, while there are challenges to attributing cyber operations generally, the technical autonomy of ACCs ultimately does not change how the law applies and imposes obligations on the State and its human agents in their use of ACCs. Further, even where software agents were given a degree of legal personality, the link between these entities and the human beings responsible for their creation is sufficient to establish attribution under the law on State responsibility.

II
AUTONOMOUS CYBER CAPABILITIES

Discovered in 2010, the malicious software named 'Stuxnet' illustrated how offensive cyber capabilities with a significant degree of autonomy can be designed and used by States in pursuit of their interests.[1] Stuxnet infected non-networked computers within Iran's Natanz enrichment facility and adjusted the frequency setting that determines the speed at which nuclear centrifuges are spun.[2] It used a number of features to prevent anti-virus and other security mechanisms from detecting it, and also made it appear to the human operators of the facility that the infected computers were operating normally.[3] Ultimately, it is believed to have been responsible for causing physical damage to approximately 1000 centrifuges in the Natanz facility.[4] While Stuxnet's creators had the technical ability to control Stuxnet through its command-and-control servers, given the non-networked nature of the Natanz facility it operated

1 Josh Halliday and Julian Borger, 'Nuclear Plants Likely Target of Foiled Cyber Sabotage' (*The Guardian*, 25 September 2010) <https://www.theguardian.com/world/2010/sep/25/iran-cyber-hacking-nuclear-plants>. In 2012 it was reported that Stuxnet had been part of operation 'Olympic Games'– an operation started during the George W Bush administration and continued during the Barack Obama administration whereby the US and Israel sought to undermine Iran's ability to enrich uranium. David Sanger, 'Obama Order Sped Up Wave Of Cyberattacks Against Iran' (*The New York Times*, 1 June 2012) <http://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html>. Little is known about how exactly it was created, but one suggestion is that Stuxnet was created 'in a modular fashion' by teams from various organisations involved in malicious cyber activities, and that these teams potentially had no idea about the overall project they were working on. Alexander Klimburg, 'Mobilising Cyber Power' (2011) 53 Survival 41, 43.

2 Nicolas Falliere, Liam Murchu and Eric Chien, 'W32.Stuxnet Dossier' (ver 1.4, Symantec 2011) 41–3 <https://archive.org/details/w32_stuxnet_dossier>.

3 ibid 14, 48–9.

4 Joby Warrick, 'Iran's Natanz Nuclear Facility Recovered Quickly from Stuxnet Cyberattack' (*The Washington Post*, 16 February 2011) <http://www.washingtonpost.com/wp-dyn/content/article/2011/02/15/AR2011021506501.html>.

in, all of its functions were embedded within its code enabling it to operate autonomously.[5] Therefore, once activated, Stuxnet was capable of propagating, identifying the particular systems it targeted, and delivering its payload without any direct or real-time human control.

An illustration of the possible ways in which ACCs can be used comes from the 'Cyber Grand Challenge' which was organised by the United States (US) Defense Advanced Research Projects Agency (DARPA) in 2016. The Cyber Grand Challenge involved the use of 'cyber reasoning systems' (CRS) to perform cybersecurity functions without any real-time human intervention. For participating teams, the objective was to score points (and avoid losing them) by protecting the team's software from adversaries by finding and patching vulnerabilities, keeping their own software available, functional and efficient, and exploiting vulnerabilities in adversary software.[6] All of this needed to be done autonomously by the CRS in that, once activated, humans could not intervene in their functioning.[7] Mayhem, the system that won the competition, was able to autonomously discover and patch software vulnerabilities, as well as to discover and exploit vulnerabilities in its adversaries' software. It and the other CRSs involved in the competition needed to make strategic decisions around which vulnerabilities to patch (or leave unpatched), which patches to use, which teams to attack and with what exploits, and how to allocate their resources in performing these functions.[8] As such, Mayhem was intelligent (in the computer science sense of the term) as it could adapt to the behaviour and strategies of its adversaries in a changing and unknown environment. For example, in deciding whether to deploy a patch against a vulnerability that was being exploited, instead of using a fixed probability cut off, Mayhem adjusted the threshold dynamically which allowed it to adapt to various strategic situations.[9]

While Stuxnet behaved in a way that was pre-determined and known to its developers, Mayhem in turn displayed behaviours in terms of the decisions it made about its strategies that were not necessarily foreseeable (its developers were 'often surprised' by the strategic decisions that it made about whether or not to patch certain vulnerabilities).[10]

5 Falliere, O'Murchu and Chien (n 2) 3.
6 Thanassis Avgerinos and others, 'The Mayhem Cyber Reasoning System' (2018) 16 IEEE Security & Privacy 52, 53. See also Steve Lohr, 'Stepping Up Security for an Internet-of-Things World' (*The New York Times*, 16 October 2016) <https://www.nytimes.com/2016/10/17/technology/security-internet.html>.
7 Avgerinos and others (n 6) 58.
8 ibid 56.
9 ibid 57–8.
10 ibid 58.

As such, while its overall goal of defending its own server while attacking other servers, and the techniques it had available were programmed into Mayhem by its developers in advance, the exact ways in which it decided to do so (for example, which capabilities to deploy and when) were not.

Accordingly, Stuxnet and Mayhem demonstrate the real and possible ways in which ACCs can be used by States for offensive and/or defensive purposes.[11] Particularly where malware is equipped with capabilities that allows it to operate for extended periods of time in complex and uncertain environments (in ways not always known to the developers), there is a risk of unintended and unpredictable effects that the cyber operation may cause. When used by States in their international relations, this raises questions about State responsibility for conduct involving the use of these capabilities.

For the purposes of this chapter, autonomy will be defined in technical terms as 'the ability of a system to behave in a desired manner, or achieve the goals previously imparted to it by its operator, without needing to receive the necessary instructions from outside itself on an ongoing basis.'[12] Essentially, autonomy in cyber capabilities involves software with the ability to act within an environment in pursuit of its goal without direct or real-time human control. This definition accounts for ACCs where the high level goal or purpose of the entity has been defined by human programmers in advance, even if every specific low level step that the entity must take to achieve that goal is unforeseeable.[13] As such, this chapter adopts the approach that ACCs are ultimately tools which are implemented or used by human individuals to achieve a goal.[14] Even though ACCs, once activated, operate without direct or real-time human control, their behaviour is nonetheless pre-determined in advance by human beings and enforced by their code.[15] Similarly, even if some of the specific steps taken by the ACC in pursuit of its goal may not be foreseen, the overall goal which it seeks to achieve has been pre-determined by its programming. Therefore, for the purposes of legal analysis, the control link between humans and software is not severed but only modified by the autonomous capabilities of the system. Further, it is the outcomes or effects caused by the ACCs that are most relevant for the law, and not the fact that humans are not directly completing the tasks causing these outcomes or effects.[16]

11 See also Lily Hay Newman, 'AI Can Help Cybersecurity — If It Can Fight Through the Hype' (*Wired*, 29 April 2018) <https://www.wired.com/story/ai-machine-learning-cybersecurity>.
12 Tim McFarland, 'The Concept of Autonomy', this volume, ch 2, [21].
13 ibid 17.
14 ibid 20.
15 ibid.
16 ibid 27–28.

III
INTERNATIONAL LAW ON STATE RESPONSIBILITY

International law on State responsibility is detailed in the International Law Commission's (ILC) Articles on the Responsibility of States for Internationally Wrongful Acts (ILC Articles).[17] Further, the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (*Tallinn Manual*) provides a detailed account about how the rules on State responsibility are considered to apply to State activities in cyberspace specifically.[18] While neither of these are legally binding documents, much of the ILC Articles are widely recognised as an authoritative statement of customary international law,[19] and the *Tallinn Manual* rules in relation to State responsibility largely align with the ILC Articles in terms of the substantive content of the law.

The law on State responsibility consists of secondary rules that determine the circumstances in which a State is responsible for an internationally wrongful act. This means there must first be conduct (either an act or omission) by one State (the responsible State) which amounts to a violation of a primary rule of international law (the internationally wrongful act) against an obligation owed to another State (the injured State). If these conditions are met, and provided no circumstances precluding wrongfulness exist (including, for example, necessity or self-defence), then the injured State will have the right to a remedy.

In the cyber context, an international wrongfully act is a cyber operation that is in violation of a primary rule of international law. For example in relation to Stuxnet, it is generally considered to have constituted a use of force in violation of article 2(4) of the United Nations Charter because it caused physical damage to the centrifuges.[20] Assuming that there were no relevant circumstances precluding wrongfulness (such as self-defence),

17 International Law Commission, *Report of the International Law Commission on the Work of Its Fifty-Third Session*, *Draft Articles on Responsibility of States for Internationally Wrongful Acts* (UN GAOR, 56th sess, Supp No 10, UN Doc A/56/10, 2001) ('ILC Articles').

18 Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017).

19 See UN General Assembly Secretary-General, *Responsibility of States for Internationally Wrongful Acts: Comments and Information Received from Governments* (21 April 2016) UN Doc A/71/79; UN General Assembly Secretary-General, *Responsibility of States for Internationally Wrongful Acts: Compilation of Decisions of International Courts, Tribunals and Other Bodies* (21 April 2016) UN Doc A/71/80.

20 See Samuli Haataja and Afshin Akhtar-Khavari, 'Stuxnet and International Law on the Use of Force: An Informational Approach' (2018) 7(1) Cambridge International Law Journal 99, 109–11.

Stuxnet's use would constitute an internationally wrongful act by the responsible State(s). It would also mean that Iran as the injured State would have the right to a remedy.[21]

For a State to be responsible for an internationally wrongful act, however, the conduct in question must be attributed to the State. As abstract legal entities, States can only act through human agents and representatives.[22] Thus the rules on attribution provide the process through which the conduct of a natural person becomes an 'act of State' for which the State is responsible.[23] Generally this is clear where, for example, a State organ such as its armed forces engages in an activity amounting to a violation of international law. However, where the conduct is by other entities, such as private individuals or corporations, the rules of attribution are more complex. Therefore, the actor in question and its connection to the State will determine whether the conduct can be attributed to the State under the law of State responsibility. In the cyber context, attribution can be particularly problematic on a factual or evidentiary level given the frequent use of proxy actors or technical means to obfuscate the true geographical origin of the operation (for instance by spoofing). However, provided this can be achieved on a technical level, then the rules on State responsibility establish the legal criteria in relation to whether the conduct can be attributed to a State.

A common issue raised in relation to autonomous systems is the question of fault or intention and how or to what degree the autonomy of a system impacts on this for the purposes of responsibility. Under the law on State responsibility, there has been some debate about the relevance of intention or fault in relation to whether a State can be held responsible for a violation of international law,[24] and classically a degree of fault was required for States to be responsible for their conduct.[25] As a generalisation, there have been two competing theories or approaches on this issue.[26]

21 There is, however, debate around other primary rules of international law and the threshold at which cyber operations amount to violations of international law (especially sovereignty and non-intervention). See Michael Schmitt and Liis Vihul, 'Respect for Sovereignty in Cyberspace' (2017) 95 Texas Law Review 1639; Gary P Corn and Robert Taylor, 'Sovereignty in the Age of Cyber' (2017) 111 AJIL Unbound 207.

22 See James Crawford, *State Responsibility: The General Part* (Cambridge University Press 2013) 113.

23 ibid.

24 James Crawford and Simon Olleson, 'The Nature and Forms of International Responsibility' in Malcolm D Evans (ed), *International Law* (2nd edn, Oxford University Press 2010) 464–5.

25 See Robert Kolb, The International Law of State Responsibility: An Introduction (Edward Elgar 2017) 22.

26 Crawford and Olleson (n 24) 465; Sandra Szurek, 'The Notion of Circumstances Precluding Wrongfulness' in James Crawford and others (eds), *The Law of International Responsibility* (Oxford University Press 2010) 433; Martti Koskenniemi, 'Doctrines of State Responsibility' in James Crawford and others (eds), *The Law of International Responsibility* (Oxford University Press 2010) 49–51; Brigitte Stern, 'The Elements of An Internationally Wrongful Act' in James Crawford and others (eds), *The Law of International Responsibility* (Oxford University Press 2010) 209–10.

According to the subjective fault theory, attributing an internationally wrongful act to a State requires some degree of culpability or negligence on the part of its organs.[27] In contrast, pursuant to the objective fault theory, there is no requirement of fault meaning that a State will be responsible for violations of international law regardless of their intention.[28] Based on this approach, the State's subjective intention is irrelevant and the notion of fault is not necessary to establish State responsibility.[29]

The approach taken by the ILC Articles largely aligns with the objective responsibility approach, however, the notion of fault is not entirely abandoned. The ILC Articles provide that '[i]n the absence of any specific requirement of a mental element in terms of the primary obligation, it is only the act of a State that matters, independently of any intention.'[30] Essentially under this approach, the intention of the State is only relevant to the extent that it is an element required to establish a violation of an international legal obligation.[31] For example, if a State uses a cyber operation that amounts to a violation of another State's sovereignty, the intention of the responsible State will not be relevant because there is no specific element of intention required for establishing a violation sovereignty.[32]

IV
ATTRIBUTION

As such, whether the use of an ACC amounts to an internationally wrongful act will depend on the primary rule of international law in question. However, the conduct in question must also be attributed to the State for it to be responsible, and there various grounds on which this can be

27 Szurek (n 26) 433. For example, in the Corfu Channel case, Albania was held liable based on its presumed knowledge of the fact that mines had been laid in its territorial waters and the subsequent failure to notify the British to prevent its warships from being damaged. *Corfu Channel (UK v Albania)* (Merits) [1949] ICJ Rep 4, 22.

28 ibid.

29 Stern (n 26) 209.

30 James Crawford, *The International Law Commission's Articles on State Responsibility: Introduction, Text and Commentaries* (Cambridge University Press 2002) 84.

31 Despite this, Crawford and Olleson note that 'it is illusory to seek for a single dominant rule' as the factual circumstances on the case will vary. They distinguish between whether the wrongful conduct involves an act or an omission — fault is generally relevant in relation to the latter whereas if a 'State deliberately carries out some specific act, there is less room to argue that the harmful consequences were unintended and should be disregarded.' See Crawford and Olleson (n 24) 465.

32 Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) 19 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>; Michael Schmitt, 'Autonomous Cyber Capabilities and the International Law of Sovereignty and Intervention', this volume, ch 7, section VI.

done.[33] These include where the conduct is by a State organ; where the conduct is by a private entity that has been empowered to exercise inherently governmental functions; where the conduct is by private persons or groups that are acting under the instructions of, or under the direction or control of a State; and where a State acknowledges and adopts the conduct as its own. This section will explain each of these grounds and examine the issues raised by ACCs in relation to attribution in these contexts. It will demonstrate that the challenges in the application of these rules to wrongful conduct involving the use of ACCs are not unique due to the technical autonomy of these systems but instead common to those involving cyber operations generally. This is given the approach taken by the ILC Articles in relation to fault, and the fact that ACCs are tools which are programmed and used by human beings (whether by agents of State organs or those acting under the authority or control of the State) who international law imposes legal obligations on.

A STATE ORGANS

According to article 4 of the ILC Articles, States are responsible for the wrongful conduct of their organs.[34] This is echoed in rule 15 of the *Tallinn Manual* which provides that cyber operations conducted by State organs are attributable to the State.[35] According to the ILC, a State's organ includes 'any person or entity which has that status in accordance with the internal law of the State.'[36] For the purposes of State responsibility, no distinction is made between organs exercising executive, legislative or judicial functions (or some combination of these public powers). Similarly the position of the person within the organisation and how the organisation is characterised within the State's domestic law does not matter.[37] Thus, as the commentary to the ILC Articles notes, the term

33 On the attribution of State responsibility for cyber operations more generally, see, eg, Scott Shackelford and Richard Andres, 'State Responsibility for Cyber Attacks: Competing Standards for a Growing Problem' (2011) 42 Georgetown Journal of International Law 971, 986–93; Nicholas Tsagourias, 'Cyber Attacks, Self-Defence and the Problem of Attribution' (2012) 17 Journal of Conflict & Security Law 229, 236–40; Peter Margulies, 'Sovereignty and Cyber Attacks: Technology's Challenge to the Law of State Responsibility' (2013) 14 Melbourne Journal of International Law 496, 506–16; William Banks, 'State Responsibility and Attribution of Cyber Intrusions After Tallinn 2.0' 95 Texas Law Review 29. See also Christian Payne and Lorraine Finlay, 'Addressing Obstacles to Cyber-Attribution: A Model Based on State Response to Cyber-Attack' 49 George Washington International Law Review 35.

34 ILC Articles (n 17) art 4; Crawford (n 30) 94.

35 Schmitt (n 18) 87.

36 ILC Articles (n 17) art 4.

37 ILC Articles (n 17) art 4.

organ is used 'in its most general sense' and 'extends to organs of government of whatever kind or classification, exercising whatever functions, and at whatever level in the hierarchy' of the State.[38] Even where an entity is not officially a State organ, an entity can be considered a *de facto* organ of the State where it acts in complete dependence of the State.[39] Consistent with the ILC Articles, the *Tallinn Manual* authors adopt a broad construction of the term organ and maintain that it includes all persons or entities with that status in the State's domestic laws.[40] They note that attribution is clearest where the conduct is by State organs — for example, where a State's military or intelligence agency engages in cyber activities that constitute an internationally wrongful act.[41]

1 *Ultra vires*

Pursuant to article 7 of the ILC Articles, the conduct of a State organ is attributable to a State 'even it if exceeds its authority or contravenes instructions' provided that 'the organ, person or entity acts in that capacity'.[42] Accordingly, a distinction is made between official acts and those in a person's private capacity. To illustrate this, in the *Caire* case members of Mexico's armed forces unsuccessfully attempted to extort money from Caire (a French national operating a boarding house in Mexico City), and later took him to their barracks to be stripped and eventually killed him in a separate location. Because the officers were in uniform and made use of army barracks, even if they had been acting contrary to orders, they appeared to be acting on behalf of the State. As such, while their actions were *ultra vires*, the conduct was attributable to Mexico as their actions were not private.[43] Similarly, where an agent of a State organ gives the appearance of State authority and, for example, seizes money under official customs powers, then the State will be responsible even though the actions are *ultra vires*.[44] Therefore, what matters according to the ILC is that the person was 'purportedly or apparently carrying out their official functions' — that is, they were 'cloaked' with State authority and thus acting with apparent authority.[45]

38 Crawford (n 30) 95.
39 Crawford (n 22) 124–5; *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v Serbia and Montenegro)* ('*Bosnian Genocide*') (Judgment) [2007] ICJ Rep 43, 205.
40 Schmitt (n 18) 87.
41 ibid.
42 ILC Articles (n 17) art 7.
43 Crawford (n 22) 137; *Estate of Jean-Baptiste Caire (France) v United Mexican States* (1929) 5 RIAA 516, 231.
44 In contrast, where there are bribes obtained for personal profit without any appearance of making use of State authority to do so, then those acts will not be attributable. See ibid 138. *Yeager v Iran* (1987) 17 Iran–US CTR 92.
45 Crawford (n 30) 108.

In the cyber context, the *Tallinn Manual* gives the example of where the member of a State's cyber unit 'conducts unlawful cyber operations in defiance of orders to the contrary', then this will nonetheless be attributable to the State.[46] However, the conduct in question must occur in the official capacity of the person with apparent authority (and not in their private capacity).[47] An example given by the *Tallinn Manual* of where acts would have been conducted in a purely private capacity opposed to under an apparent official capacity is where an individual exploits 'access to cyber infrastructure for criminal activity leading to private gain.'[48]

2 *ACCs*

Where a State organ's internationally wrongful conduct involves an ACC, there is no difference in the application of these provisions that arises specifically due to the autonomous capabilities of the means used. The conduct of State officials responsible for procuring the ACC, as well as other personnel such as those within the State organ responsible for developing, designing, or programming the ACC, will be considered conduct by the State under the law on State responsibility.[49] Similarly it would not matter whether a military or intelligence agency, law enforcement agency, or any other agency made use of the ACC as it would nonetheless be attributable to the State. It would also not matter whether the ACC was used in a defensive or offensive capacity — provided it was used by a State organ, that conduct would be attributable to the State.

In relation to *ultra vires* acts, even in situations in which ACCs operate in ways unknown to their programmers or result in unpredictable consequences these acts would be attributable. Ultimately the focus of this inquiry is whether the State official acted with apparent authority in using the ACC even if they acted contrary to instructions, and not on whether the ACC did so. The only situation in which the conduct will not be attributable to the State is where a State official makes use of an ACC for purely private purposes without the appearance of State authority. For example, consider where an official from the State responsible for Stuxnet acted contrary to instructions and altered Stuxnet so that it would deliver its payload to a nuclear facility in a third State, and also so it would disrupt the operation of systems within a private

46 Schmitt (n 18) 89.
47 Crawford (n 30) 108.
48 Schmitt (n 18) 89.
49 Liivoja, Naagel and Väljataga (n 32) 33.

Iranian company that is a direct competitor to the official's family business. While the conduct in both of these situations would be *ultra vires*, whether the official in question acted with the apparent authority of the responsible State or not would be difficult to determine. Unlike the case law involving human officials wearing uniforms or making use of their appearance of authority and army barracks, the use of Stuxnet in this hypothetical scenario would lack any of these factors indicating a physical appearance of authority. However, in the cyber context the appearance of authority can be evident from the use of particular cyber infrastructure (for example, the communication structures used to deliver Stuxnet or maintain its command-and-control capabilities) which are known to be associated with a particular State actor.[50] Where these kind of indicators are evident to the injured State, it would be difficult to determine whether the misuse of a cyber capability such as Stuxnet would have been in a purely private capacity. Instead, it is more likely to appear to have been conducted under the apparent authority of the State.[51]

Accordingly, while there are challenges in attributing cyber operations on a technical level to determine who was in fact responsible for them,[52] these are not issues limited to the use of ACCs. As such, provided the cyber operation can be attributed to a State on a technical level, this is likely to be sufficient to give the appearance of State authority in the use of the ACC as it would be difficult to distinguish between official and private acts conducted using the capability. This also extends to situations where the use of the ACC causes unpredictable or unintended effects.

50 Cyber infrastructure is among the indicators used for technical attribution of cyber operations. See Office of the Director of National Intelligence, 'A Guide to Cyber Attribution' (12 September 2018) 3 <https://www.dni.gov/files/CTIIC/documents/ODNI_A_Guide_to_Cyber_Attribution.pdf>.

51 For example, in the non-cyber context, Trapp highlights the difficulty of attributing terrorist acts carried out by State organs where such acts are 'carried out by secret service agents who do not display any outward manifestation of the authority under which they act. The State organs will appear to be private citizens, engaging in private conduct. As a result, such acts of terrorism could not be said to have been carried out under colour of authority.' Kimberley Trapp, *State Responsibility for International Terrorism* (Oxford University Press 2011) 35.

52 See Nicholas Tsagourias and Michael Farrell, 'Cyber Attribution: Technical and Legal Approaches and Challenges' (2020) 31(3) European Journal of International Law 941.

B PRIVATE ACTORS EMPOWERED TO EXERCISE GOVERNMENTAL AUTHORITY

Pursuant to article 5 of the ILC Articles, the conduct of persons or entities that are not State organs but have been empowered to exercise elements of governmental authority can also be attributed to the State.[53] This is echoed in rule 15 of the *Tallinn Manual* which provides that, in addition to the conduct of State organs, the conduct of persons or entities empowered by domestic law to exercise elements of governmental authority are attributable to the State.[54] According to the ILC, this rule is designed to capture entities that 'exercise elements of governmental authority in place of State organs, as well as situations where former State corporations have been privatized but retain certain public or regulatory functions.'[55] This may include, for example, private security companies in charge of a State's prisons or detention centres, or airlines with powers over immigration.[56] Regardless of how the entity is characterised domestically (whether public or private), as well as any financial links it may have with the State or a degree of executive control exercised by the State, the core concern under article 5 is whether the entities in question 'are empowered, if only to a limited extent or in a specific context, to exercise specified elements of governmental authority.'[57] For example, in the cyber context, the *Tallinn Manual* gives the examples of a private company authorised by law to engage in an offensive cyber operation against another State, and a private entity legally empowered to engage in espionage by cyber means.[58]

1 *Ultra vires*

Under article 7 of the ILC Articles, the conduct of these entities can also be attributed to the State where they act outside the scope of their authority or in contravention of instructions.[59] As mentioned above, the question here is whether, even if they acted *ultra vires*, the conduct was cloaked with State authority or conducted in a private capacity. The *Tallinn Manual* provides the example of a State that lacks the technical

53 ILC Articles (n 17) art 5.
54 Schmitt (n 18) 87.
55 Crawford (n 30) 100.
56 ibid.
57 ibid. In this context, the commentary to the ILC Articles gives the example that 'the conduct of a railway company to which certain police powers have been granted will be regarded as an act of the State under international law if it concerns the exercise of those powers, but not if it concerns other activities (e.g. the sale of tickets or the purchase of rolling stock).' ibid.
58 Schmitt (n 18) 89.
59 ILC Articles (n 17) art 7; see also Schmitt (n 18) 90–1.

capacity to defend its governmental cyber infrastructure and thus provides regulatory authority for a private company to do so through passive defence measures.[60] If this company were to use active defence measures (hacking back) in this situation, even though this was not authorised by the State (and thus *ultra vires*) it would be attributable to it because it was incidental to defending the government's networks (which it was empowered to do).[61] However, where for example this company conducts private criminal activities using cyber means, then those activities would not be attributable.[62]

2 *ACCs*

In relation to the attribution of conduct involving the use of ACCs under article 5, as with article 4, there is little difference in the application of these rules due to the technical autonomy of these capabilities. Consider, for example, a private company that has been authorised to develop and deploy an ACC for an offensive cyber operation. The ACC has a capability allowing it to learn from previous attack scenarios and choose the most appropriate vector to compromise a particular network and disrupt the operation of computers therein. However, the ACC operates in an unpredictable way causing more damage than intended and spreads to networks outside the target State causing similar effects there. In this scenario, provided the company had been authorised to exercise governmental functions and appeared to act in that capacity (and assuming engaging in an offensive cyber operation on behalf of the State amounted to such), the conduct would be attributable even if *ultra vires*. But where the company uses the ACC in pursuit of its own goals — for example, by adding a ransomware feature that allows the software to spread to civilian computers within the target State and encrypt the users' data for ransom — this would not be attributable. As discussed previously, while on a technical level it may be difficult to distinguish whether the conduct involving the use of an ACC was undertaken with State authority or in a private capacity, ultimately the focus for the legal analysis is on whether the company was doing so in its use of an ACC (and not whether the ACC acted *ultra vires*). Determining this is a challenge common with cyber operations generally and is not problematized by ACCs specifically.

60 Schmitt (n 18) 90.
61 ibid.
62 ibid 91.

C PERSONS OR GROUPS INSTRUCTED, DIRECTED OR CONTROLLED BY A STATE

1 *Article 8*

As such, the conduct of State organs and other entities that have been empowered to exercise governmental authority are attributable to a State. Normally, however, States are not responsible for the conduct of private persons or entities. An exception to this is contained in article 8 of the ILC Articles which provides that a person's or group's conduct will be considered an act of State if they are 'in fact acting on the instructions of, or under the direction or control of, that State in carrying out the conduct.'[63] Consistent with the ILC Articles, the *Tallinn Manual* provides in rule 17 that '[c]yber operations by a non-State actors are attributable to a State when engaged in pursuant to its instructions or under its directions or control'.[64]

According to the ILC, article 8 provides two different types of situations where the factual relationship between the private individual or entity attracts State responsibility. First, where the person or entity acts 'on the instructions of the State in carrying out the wrongful conduct', and second, where the person or entity under the State's direction or control more generally.[65] Despite these two general categories, the terms instructions, directions or control are disjunctive and operate autonomously.[66] As such, instead of exercising public power under article 4 or having been authorised to exercise it under article 5, article 8 concerns situations where the factual circumstances demonstrate a 'real link' between the person or entity in question and the State machinery.[67] The *Tallinn Manual* gives various examples of actors whose conduct could potentially be attributed to the State on this basis, including 'individual hackers; informal groups like Anonymous; criminal organisations engaged in cyber crime; legal entities such as commercial IT services, software, and hardware companies; and cyber terrorists or insurgents.'[68] What is essential in each case is that the private actor in question is acting under the instructions of, or under the direction and control of a State.

63 ILC Articles (n 17) art 8.
64 Schmitt (n 18) 94.
65 Crawford (n 30) 110. While the term 'directions' is generally conflated with term 'instructions' (including by the ILC), some argue that directions can be understood to refer to an ongoing period of instructions. See Kubo Mačák, 'Decoding Article 8 of the International Law Commission's Articles on State Responsibility: Attribution of Cyber Operations by Non-State Actors' (2016) 21 Journal of Conflict & Security Law 405, 417–19.
66 Crawford (n 30) 113.
67 ibid 110.
68 Schmitt (n 18) 95.

(a) Instructions

Acting under the instructions of the State means that the conduct of the non-State actors has been authorised by the State and, distinct from article 5, in this context it does not matter whether the non-State actors are engaging in a 'governmental activity' or not.[69] It involves a situation in which the non-State actors is in a subordinate position to the State at the time when the decision to engage in unlawful conduct is made.[70] Examples given by the ILC include where States recruit individuals or groups outside of their official structures such as where members 'not forming part of their police or armed forces, are employed as auxiliaries or are sent as "volunteers" to neighbouring countries, or who are instructed to carry out particular missions abroad.'[71] Elsewhere James Crawford gives the example of where a company is engaged to conduct certain activities on behalf of a State (for instance, private military or security) — 'the State may incorporate instructions into the terms of the company's contract or issue instructions in the field, or both.'[72] An example given by the *Tallinn Manual* in this context is where a State without a defensive cyber organisation recruits private individuals or volunteers to respond to large unanticipated cyber operations against it.[73] The actions of those individuals in this scenario would be attributable on the basis that they are acting as an auxiliary of the State and an instrument of it, and acting on its behalf.[74] Another example given is where the armed forces of a State requests a private company to conduct a particular type of cyber operation in support of an ongoing kinetic/non-cyber operation — here the cyber operations within the scope of the request would be attributable as they are acting under specific the instructions of the State.[75]

However, there is a degree of uncertainty about the specificity of instructions needed for the wrongful conduct. According to the ICJ in the *Bosnian Genocide* case, instructions must be given by the State 'in respect of each operation in which the alleged violations occur, not generally in respect of the overall actions taken by the persons or groups of persons having committed the violations'.[76] The uncertainty arises as to the scope of the meaning of 'operations' in this context and whether

69 Crawford (n 30) 110.
70 Lindsey Cameron and Vincent Chetail, *Privatizing War: Private Military and Security Companies under Public International Law* (Cambridge University Press 2013) 205.
71 Crawford (n 30) 110.
72 Crawford (n 22) 145.
73 Schmitt (n 18) 95.
74 ibid.
75 ibid 95–6.
76 *Bosnian Genocide* (n 39) 208; see also Crawford (n 22) 145.

general instructions are sufficient or whether specific instructions must be given for every particular instance of wrongful conduct.[77] The ILC commentary on this is vague, though it has been interpreted to lean towards the position that general instructions are sufficient.[78] Crawford also adopts the position that instructions can be general leaving the method of doing so open and it does not need to refer to specific acts. He writes that 'where ambiguous or open-ended instructions are given, acts which are considered incidental to the task in question or conceivably within its expressed ambit may be considered attributable to the State.'[79] It is unclear from the *Tallinn Manual* which position its authors adopted on this, but the examples provided in the *Tallinn Manual* involve situations in which non-State actors are instructed to respond to a particular incident (for example, an unanticipated massive cyber operation against the State, or a company instructed to conduct cyber operations in support of ongoing kinetic operations). In each of these situations, the general instructions given to conduct the specific operations would be sufficient to attribute any wrongful conduct occurring in that operation. Assuming general instructions to conduct a cyber operation would be for achieving a particular effect or outcome, then the specific cyber means used in the operation would not matter for the purposes of attribution. Thus whether the company makes use of DDoS attacks or sophisticated ACCs, the conduct would be attributable as it has been instructed to do so.

(b) Direction and control

As to 'direction or control', conduct in this context 'will be attributable to the State only if it directed or controlled the specific operation and the conduct complained of was an integral part of that operation.'[80] While the ILC maintained that these terms are disjunctive and thus operate autonomously, according to Crawford international courts have tended to treat them together as a single basis for attribution.[81] But he does suggest that direction 'implies a continuing period of instruction'.[82] Others suggest that direction requires 'that the State leads the steps to be taken in the commission of the unlawful conduct; it must show how the operation is to be conducted.'[83]

77 Hannah Tonkin, *State Control over Private Military and Security Companies in Armed Conflict* (Cambridge University Press 2011) 115; see also Crawford (n 22) 145.
78 Trapp (n 51) 38.
79 Crawford (n 22) 145.
80 Crawford (n 30) 110.
81 Crawford (n 22) 145.
82 ibid 145 fn 28.
83 Cameron and Chetail (n 70) 209.

The most contentious basis for attribution under article 8, however, is the notion of control. Despite some debate in light of the jurisprudence of international criminal law tribunals, the generally accepted standard of control is one of 'effective control.' This requires more than a 'general situation of dependence and support' between the State and the non-State actor and instead, based on the *Nicaragua* case, requires an actual exercise of control to a degree to justify that the non-State actor is acting on behalf of the State.[84] The ICJ in the *Bosnian Genocide* case confirmed this standard of control.[85] In the *Nicaragua* case, the ICJ maintained that for a State to be in effective control of a non-State actor, there needed to be more than a general level of control (in that case, the provision of training, and financial, logistical and intelligence support to the non-State actor). Instead, what was needed was control over particular operations in which the conduct amounting to internationally wrongful acts was committed. In that case, the conduct of the non-State actors in question could have been committed without the control of the US.[86]

According to the *Tallinn Manual* (and with reference to the ILC commentary), a State is in effective control of a cyber operation where it 'determines the execution and course of the specific operation and the cyber activity engaged in by the non-State actor is an "integral part of that operation".'[87] The *Tallinn Manual* authors note that the standard of effective control 'includes both the ability to cause constituent activities of the operation to occur, as well as the ability to order the cessation of those that are underway.'[88] They provide the example of a State that contracts with a software company and is involved in directing the process of creating software embedded with exploits in order to use a software update feature as the vector to conduct a cyber operation against a State using software from that company.[89] In contrast however, where a State simply provides 'general support or encouragement' to a non-State actor, including providing malware that is subsequently used by the non-State actor to conduct a cyber operation, the State will not be in effective control of them.[90]

84 Crawford (n 30) 111.
85 *Bosnian Genocide* (n 39) 208.
86 Trapp (n 51) 39–40; Crawford (n 22) 149.
87 Schmitt (n 18) 96 citing Crawford (n 30) 110.
88 Schmitt (n 18) 96.
89 ibid.
90 ibid 97.

2 *Ultra vires*

Article 7 of the ILC Articles provides, as discussed above, that the actions of State organs that are *ultra vires* are attributable to the State (unless performed in a private capacity). But this provision does not apply to conduct attributed under article 8. However, the commentary to article 8 provides that where a non-State actor acts in contravention of instructions or directions given by a State, if the 'unlawful or unauthorized conduct was really incidental to the mission' and not 'clearly beyond it', then the State will be responsible.[91] Some suggest that this is involves 'weighing whether or not the unlawful act was done to assist in the accomplishment of the mission' in order to determine 'whether the instructing State had accepted the likelihood of its occurrence.'[92]

According to the *Tallinn Manual*, an example of a situation where the conduct would not be *ultra vires* and thus attributable to the State is where a State authorises a company to conduct a cyber operation against another State's industrial control systems but the malware used in the operation spreads to a third State causing damage to systems there.[93] However, where a company is instructed to produce malware to use against one State, but instead it misappropriates it and uses it against another State, then the conduct against the third State would not be attributable because it was outside the scope of its instructions.[94]

In relation to wrongful conduct committed under the effective control of a State, according to the ILC the conduct in question must be attributable to the State under article 8 for the State to be responsible (that is, whether *ultra vires* or not depends on the control link).[95] This effectively means that, the main factor differentiating these circumstances involving the factual link between the State and non-State actor (that is, whether under instructions or directions on one hand, or control on the other) is temporal — 'in one case a factual link at a particular point, while in the other, "control" constitutes a continuous factual link.'[96] The State must be in effective control over the particular operation in which the internationally wrongful act by the non-State actor occurred.

The *Tallinn Manual* adopts the same approach and highlights how the application of this basis for attribution is a complex issue that must

91 Crawford (n 30) 113.
92 Cameron and Chetail (n 70) 207.
93 Schmitt (n 18) 98.
94 ibid.
95 Crawford (n 30) 113.
96 Olivier de Frouville, 'Attribution of Conduct to the State: Private Individuals' in James Crawford and others (eds), *The Law of International Responsibility* (Oxford University Press 2010) 271.

be considered on a case-by-case.[97] Its authors write that 'if the cyber operations are extraneous or unrelated to the purpose of the operation over which the State exercises 'effective control', they are not attributable to the controlling State.'[98] As such, the State will only be responsible for conduct for cyber operations conducted outside the scope of the non-State actor's authority where they are integral to the mission 'in the sense that they are an essential part of the operation over which the State exercises "effective control."'[99] Further, in this context, it does not matter if the non-State actor 'ignores or disobeys the directions' that they have received from the State for the particular operation.[100]

For example, in relation to attributing the conduct of private military companies to a State on this basis, Hannah Tonkin highlights relevant factors to be considered including a detailed contract with the State having a 'preponderant or decisive role in selecting, financing, organising and planning the particular … operation to be performed under the contract' and potentially where 'the State will also supply and equip the contractors for the operation' and 'set[s] out the specific goals of the operation' — then this will contribute to fulfilling the effective control requirement.[101] The *Tallinn Manual* gives the example of where an IT company that is under the effective control of a State and conducts a cyber operation in a way that violates an obligation owed to a third State (using a server located there when the originally planned server/location has become impossible to use during the operation). While using this server and violating an obligation owed to the third State was not authorised by the State, it would be attributable as it was incidental to the cyber operation.[102] In contrast, where the company also gathered data unlawfully from the server about a business competitor, that part of the operation would be *ultra vires* and not attributable.[103]

3 *ACCs*

As to the article 8 bases for attribution in relation to conduct involving the use of ACCs, ultimately there is little difference that arises from the autonomous capabilities of these systems in relation to the outcome

97 Schmitt (n 18) 97.
98 ibid 98.
99 ibid.
100 ibid.
101 Tonkin (n 77) 120. In contrast, 'where the contract of hire is relatively broad in scope and/or gives the company a high degree of discretion in planning, organising and performing its activities, it will be necessary to focus on the other mechanisms available to the hiring State to control PMSC [private military security company] conduct in the field.'
102 Schmitt (n 18) 98.
103 ibid 98–9.

of the legal analysis. Instead, the problems are common to those involved in attributing cyber operations generally on these bases.

In relation to instructions and directions, where this factual link exists so that the non-State actor is instructed or directed by the State, any conduct involving the use of ACCs in this context will be attributable. For example, consider a scenario where a non-State actor was instructed to develop and deploy an ACC in an offensive cyber operation to undermine a specific uranium enrichment facility in a particular State. Here the initial instructions and directions provided to the non-State actor would be sufficient to create the factual link between the State and the non-State actor's conduct. The autonomous capabilities in question do not change this as the State's instructions or directions are, in effect, translated into the programming of the ACC. This conclusion would be the same even if the ACC was capable of learning and thus where the exact ways in which it propagated or undermined the operation of centrifuge machines in that facility were not entirely foreseeable as it would nonetheless have operated in pursuit of its overall high level goal that it was programmed for.

In relation to *ultra vires* acts in this context, consider where the ACC operated in an unforeseeable way which resulted in it also undermining the operation of an enrichment facility in a third State that was not the intended target. Whether this conduct is attributable would depend on whether the conduct of the non-State actor in developing and deploying the ACC was incidental to the mission that it received instructions or directions for. In this example, given that the wrongful conduct resulted from the way in which the ACC was programmed (that is, it was not programmed with sufficiently specific parameters, or it was not properly equipped to deal with uncertainty in the environment it operated in), it would be difficult to suggest this was not incidental to the mission (as it was done in order to serve the mission). Only where the non-State actor programmed the ACC to pursue a goal unrelated to the mission that it received instructions or directions for, would the wrongful conduct arising from that not be attributable. Again, the focus is on the conduct of the non-State actor in the given context, and the technical autonomy of the means used in this context does not impact on the legal analysis.

In relation to effective control, a more continuous factual link must be established to demonstrate the State was in control of the entire operation involving the use of an ACC by the non-State actor. This requires that the State was in control over every specific wrongful act caused by the use of the ACC. Here the question arises as to whether, for example,

an operation involving deploying an ACC that operates for an extended period of time in a closed network environment without real-time human control remains within the 'effective control' of the State. In other words, whether the control link is severed by the technical autonomy of the system given that it is no longer within the direct control of human beings. Given the approach to autonomy adopted here, any effects caused by the ACC would be attributable provided the State was in effective control of the non-State actor when the cyber capability was programmed and deployed. The effects caused by the use of the ACC in these circumstances would not be *ultra vires*. This is because even where an ACC operates in a closed network environment or results in unforeseeable effects, it is only capable of operating according to its programming and in pursuit of its overall high level goal or purpose provided by human beings. Therefore, the control link is not severed simply by the technical autonomy of the ACC (that is, by the fact that a human being is not in real-time control of it).

D ACKNOWLEDGEMENT AND ADOPTION OF CONDUCT

Another basis for the attribution of conduct of non-State actors comes from article 11 of the ILC Articles. Under this article, even if conduct cannot be attributed to a State under circumstances such as those in articles 4, 5 and 8, it can still be considered an act of the State 'if and to the extent that the State acknowledges and adopts the conduct as its own.'[104] The *Tallinn Manual* echoes this in rule 17 which provides that cyber operations by non-State actors can be attributed to a State where 'the State acknowledges and adopts the operation as its own.'[105] However, this is a narrow basis for attribution as it requires that the State does more than merely supports or endorses the conduct, as the conduct needs to be acknowledged and adopted as if it was the State's own conduct.[106] For example, according to the *Tallinn Manual*, a State merely expressing its support or approval of a cyber operation conducted by non-State actors would not be sufficient to attribute it to that State.[107] But where, for example, a State 'intentionally employ[s] its cyber capabilities to protect the non-State actor against counter-cyber operations so as

104 ILC Articles (n 17) art 11. This is akin to domestic law on agency wherein unauthorised acts of an agent can be subsequently ratified by the principal. See Crawford (n 22) 181.
105 Schmitt (n 18) 94.
106 Crawford (n 30) 122–3.
107 Schmitt (n 18) 99.

to facilitate their continuance as acts of that State',[108] then that would amount to acknowledgement and adoption. Despite this, article 11 does not require an 'all or nothing' approach as it is possible for a State to selectively adopt the conduct of non-State actors.[109] For example, in the cyber context, a State may acknowledge and adopt cyber operations by a non-State actor against a certain target but not others.[110]

1 *ACCs*

As to the attribution of conduct involving the use of ACCs under this basis of attribution, there is no difference in the way that article 11 applies to cyber operations generally. If a non-State actor that has not been empowered to exercise governmental functions, or is not acting under the instructions, directions or control of a State and engages in conduct involving the use of an ACC, this can be attributed to the State where it acknowledges and adopts the conduct as its own. As mentioned, this requires that the State does so explicitly — merely approving of the conduct, tolerating it, or refraining to disown conduct is not sufficient to establish acknowledgement and adoption, and instead it must be definitive.[111] For example, Nicholas Tsagourias and Michael Farrell note in relation to Stuxnet, that even though US officials in media reports acknowledged that the US was behind Stuxnet, these kind of statements would not be sufficient for attribution under article 11.[112] They also highlight how the law requires clear and explicit adoption of the conduct which needs to come from 'the highest levels of government'.[113] As such, even if a State's officials made comments in support of non-State actors engaging in cyber operations from their territory, or a State refrained from taking action to prevent a cyber operation being conducted from its territory, this in itself would not be sufficient to attribute that conduct to the State on the basis of acknowledgement and adoption.[114] Whether or not an ACC is used in these contexts does not matter for the purposes of the legal analysis as in any case the State needs to acknowledge and adopt the conduct of the non-State actor using that capability. Accordingly,

108 ibid.
109 Crawford (n 22) 187; see also Crawford (n 30) 123.
110 Schmitt (n 18) 99–100.
111 Crawford (n 22) 187–8.
112 Tsagourias and Farrell (n 52) 15. Also, for attribution under art 11 Stuxnet would have needed to be conducted by a third party or a non-State actor whose conduct could not otherwise be attributed to the US.
113 ibid.
114 However, in relation to the latter, the principle of due diligence would be relevant. See, eg, Michael Schmitt, 'In Defense of Due Diligence in Cyberspace' (2015) 125 Yale Law Journal Forum 68; Luke Chircop, 'A Due Diligence Standard of Attribution in Cyberspace' (2018) 67 International & Comparative Law Quarterly 643.

when it comes to the attribution of conduct involving the use of ACCs, the technical capabilities of the means used to engage in the cyber operation are not relevant for this basis for attribution.

V
ARTIFICIAL AGENTS AND LEGAL PERSONALITY

The preceding analysis considered ACCs as mere tools used by human agents, and the extent to which existing rules on attribution of conduct apply to their use. This final section now considers the attribution of conduct by ACCs in a hypothetical but possible scenario in which a State has domestically given software agents a degree of separate legal personality. Given growing discussion of new forms of legal personhood like this, the possibility exists that States could use this legal mechanism in an effort to avoid or obfuscate responsibility for their conduct. While the law on State responsibility adopts the approach that the legal status (if any) of an entity within a State does not matter as the entity's factual connection to the State is what is relevant, this hypothetical scenario nonetheless allows for the examination of the application of existing rules of international law on attribution for the conduct of ACCs as if they were separate legal agents from the human beings using them.

As such, this section shifts the focus of the inquiry to the conduct of ACCs as agents of the State and considers the extent to which their conduct can be attributable to the State. However, as will be demonstrated, even though the rules of State responsibility were developed with human actors and agents of the State in mind, on a conceptual level there is little difficulty in extending the rules on attribution to software agents. Further, given the approach adopted by the ILC Articles, the legal status of these entities does not matter if a factual connection can be established to the State. Given that software agents need to be programmed and possibly registered as legal entities in some way by human beings, that link will ultimately provide the basis for the attribution of their conduct.

A LEGAL PERSONALITY?

Given the increased technological sophistication of autonomous systems, concerns are often raised about accountability and responsibility for their actions. In this context, many have discussed the possibility that States will create a special legal status for artificial agents. For example, in its 2016 Report on Robotics, the EU Parliament outlined a range of issues relating to robots that need to be considered in the future when legislating around the use of autonomous systems such as autonomous vehicles and health care robots.[115] The Report highlighted that robots with larger amounts of autonomy will require us to consider whether existing rules around liability are sufficient or whether new principles are required to provide clarity on the legal liability for the actions of robots causing harm.[116] According to the Report, the impact assessment for any such future legal instrument should consider, among other things, the implications of

> creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently.[117]

This notion of robots having a special legal status as 'electronic persons' has been met with scepticism and criticism,[118] and many argue that there is not yet a practical need for such a status.[119] However, given how different legal systems have granted a degree of personhood to various non-human entities (including animals, ships, temples and idols), there is nothing preventing States from granting artificial agents with a degree of legal personality.[120]

115 Mady Delvaux, 'Report with recommendations to the Commission on Civil Law Rules on Robotics' (European Parliament Committee on Legal Affairs Report, 2016) <https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html>.

116 ibid 6–7.

117 ibid 18.

118 See, eg, Robotics Open Letter.EU, 'Open Letter to the European Commission Artificial Intelligence and Robotics' (2018) <http://www.robotics-openletter.eu>. See also Joanna J Bryson, Mihailis E Diamantis and Thomas D Grant, 'Of, For, and By the People: The Legal Lacuna of Synthetic Persons' (2017) 25 Artificial Intelligence & Law 273.

119 See, eg, Nathalie Nevejans, 'European Civil Law: Rules in Robotics' (European Parliament, 2016) 14–6 <https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf>.

120 See Samir Chopra and Laurence White, *A Legal Theory for Autonomous Artificial Agents* (University of Michigan Press 2011) 160. See also Visa AJ Kurki, *A Theory of Legal Personhood* (Oxford University Press 2019); Simon Chesterman, 'Artificial Intelligence and the Limits of Legal Personality' (2020) 69 International & Comparative Law Quarterly 819.

In addition to autonomous systems embodied into a clearly defined mechanical chassis, the question of the legal status of software agents operating in virtual environments has also emerged, including whether these entities could or should have some form of legal personality. Samir Chopra and Laurence White argue that, while at this stage the law of agency is the most appropriate doctrinal mechanism to deal with the issues relating to artificial agents (both virtual and mechanical),[121] a degree of legal personality is possible in the future.[122] Similarly, others argue that there are practical reasons why some form of personhood is possible as a way of ensuring there is liability for the actions of these agents, particularly where they make unpredictable decisions.[123] This has also been highlighted by the United Nations Commission on International Trade Law's in relation to the use of electronic agents in contract formation. In an explanatory note to a provision relating to the use of automated systems for contract formation, it is noted that while currently the actions of these agents are attributed to natural or legal persons on the basis that they are only capable of performing within their pre-programmed technical structures, where they begin to act more intelligently and modify their own instructions or develop new instructions, then the law's existing paradigm may no longer be applicable.[124] Therefore, given the increasing social, political, and economic interactions with these agents, and their growing technical capacity to perform legally significant actions, a form of legal personality may be needed in the future.[125]

121 Chopra and White (n 120) 22–3.

122 ibid 186–8. They argue that, as a relational concept, legal personality of artificial agents will depend on the scope and extent of the social, political and economic interactions with these agents. Further, that the decision to accord or refuse this status will ultimately be based on whether there is a pragmatic need to do so based on the results of this status (opposed to being a decision made based on conceptual claims about this status).

123 Tomasz Pietrzykowski maintains that the most compelling reason why autonomous artificial agents should have some form of personhood (instead of being regarded as mere machines) is given the benefits of this for people who could avoid being liable for the effects of the decisions made by these agents. This is the case particularly where these agents are capable of learning and adapting their behaviour in dynamic environments and thus where all of the decisions they make are not predictable to the people who programmed and created or used them. Therefore, he notes that in this context, a serious legal issue arises in relation liability for damages or unwanted consequences. See Tomasz Pietrzykowski, 'The Idea of Non-Personal Subjects of Law' in Visa AJ Kurki and Tomasz Pietrzykowski (eds), *Legal Personhood: Animals, Artificial Intelligence and the Unborn* (Springer 2017) 64. Rafał Michalczak notes that even in relation to the current capabilities of intelligent software, the unpredictability of actions taken by the software 'create a gap in the causal link between the actions of the user and the resulting consequences' thus creating conceptual complications for the law. See Rafał Michalczak, 'Animals Race Against the Machines' in Visa AJ Kurki and Tomasz Pietrzykowski (eds), *Legal Personhood: Animals, Artificial Intelligence and the Unborn* (Springer 2017) 98. See also Jaap Hage, 'Theoretical Foundations for the Responsibility of Autonomous Agents' (2017) 25 Artificial Intelligence & Law 255.

124 See United Nations Commission on International Trade Law, *United Nations Convention on the Use of Electronic Communications in International Contracts* (2007) <https://www.uncitral.org/pdf/english/texts/electcom/06-57452_Ebook.pdf>.

125 See Chopra and White (n 120) 186–8. While the exact form and the extent of rights and liabilities associated with this possible status would likely differ across States, ultimately it is most likely

B STATE RESPONSIBILITY AND NEW LEGAL ENTITIES

In light of this, the remainder of this section will consider a potential future in which software agents are given a degree of legal personality within a State's domestic legal system.[126] Similar to privatising a State owned company or creating a company as a separate legal entity, there is the potential for this to be done in an effort to limit or avoid responsibility for the actions of the software entity. The extent of legal personality (the rights and obligations the entity has, as well who has the ability to enforce those rights) of these entities can vary among States but those entities would nonetheless constitute 'legal persons' within those domestic legal systems. As mentioned above, while the law on State responsibility does not require an entity to have a separate legal status within a domestic legal system for the purposes of attribution, this scenario allows for an analysis of the rules on attribution for the conduct of ACCs.

1 *Entities and organs*

On a conceptual level, it is not difficult to view software agents as entities for the purpose of the State responsibility analysis. Articles 4 and 5 of the ILC Articles make explicit reference to 'entities' and, while article 8 only refers directly to 'persons and groups', its commentary also makes reference to 'persons or entities'.[127] An entity is generally understood as 'a thing with distinct and independent existence'.[128] In law, the notion of an entity has a more specific meaning generally referring to either an individual, an association, or other kind of legal or administrative arrangement. However, the ILC Articles use the term 'entity' in a more general sense meaning that the entity in question (be it an individual or group) does not need to have any distinct legal status under a State's domestic law.[129] As such, the term can be used to refer to various

to involve a form of dependent personality where human actors are needed to enforce any rights and obligations held by these entities (similar to a corporation). On the distinction between dependent and independent legal personality, see ibid 159–61. See also Jiahong Chen and Paul Burgess, 'The Boundaries of Legal Personhood: How Spontaneous Intelligence Can Problematise Differences between Humans, Artificial Intelligence, Companies and Animals' (2019) 27 Artificial Intelligence and Law 73, 81.

126 On how software entities could obtain a degree of legal personality within existing legal frameworks and without a legislative act, see Shawn Bayern, 'The Implications of Modern Business–Entity Law for the Regulation of Autonomous Systems' (2016) European Journal of Risk Regulation 297. See also Thomas Burri, 'Free Movement of Algorithms: Artificially Intelligent Persons Conquer the European Union's Internal Market' in Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018).

127 Crawford (n 30) 94, 100, 110.

128 Angus Stevenson (ed), *Oxford Dictionary of English* (online, 3rd edn, Oxford University Press 2010).

129 The commentary notes that 'person or entity' 'is used in a broad sense to include any natural

non-human entities both in descriptive terms (as things with a separate existence) and in legal terms (as entities with some form of legal agency, such as a corporation). Similarly, the term 'organ' is used in the ILC Articles 'in its most general sense' and covers any organ regardless of its legal status, hierarchy or how it is characterised within the State's domestic legal structures.[130] This is echoed in the *Tallinn Manual*.[131] Accordingly, there would be no conceptual difficulty in considering a software agent's actions as those of a State organ for the purposes of State responsibility.

Even where an entity does not form part of or have the status of a State organ under the State's internal law, it is possible for it to be equated to one under international law where it acts in 'complete dependence' of the State of which it merely constitutes an instrument of.[132] As software agents need to be programmed by human beings and would likely require some form of registration as legal entities, where a State is involved in creating this entity and in fact exercises complete control over it, the entity could be deemed a *de facto* organ of the State.[133] Thus, even where a software agent were not considered an agent of a State organ under a State's domestic law, where it can be seen to clearly function as one and merely act as an instrument of the State that has a 'great degree of control' over it, then its conduct can be attributed to the State.[134] In this scenario, regardless of its characterisation or legal status as an agent under a State's domestic law, where a software agent is in complete dependence on the State to the extent that it is simply acting as an instrument of it, then it is possible to equate that entity to a State organ for the purposes of attribution.

Conversely, where a software agent is not considered an agent of a State organ, it is possible to attribute its conduct to the State where it has been empowered to exercise governmental functions pursuant to article 5 of the ILC Articles.[135] As mentioned above, the ILC takes a broad

or legal person, including an individual office holder, a department, commission or other body exercising public authority, etc.' See Crawford (n 30) 98. Also, in relation to a draft article on attribution of conduct by private individuals, the term 'individual' was replaced with 'person' so as to cover natural and legal persons. See de Frouville (n 96) 262.

130 Crawford (n 30) 95–6. According to the ILC, the State 'is held responsible for the conduct of all of the organs, instrumentalities and officials which form part of its organization and act in that capacity, whether or not they have separate legal personality under its internal law.' ibid 93.

131 The *Tallinn Manual* authors adopt a broad construction of the term organ and maintain that it includes all persons or entities with that status in the State's domestic laws. Schmitt (n 18) 87.

132 *Bosnian Genocide* (n 39) 205 cited in Schmitt (n 18) 88.

133 Crawford (n 22) 125. Crawford writes that according to the ICJ in *Nicaragua*, this depends on '(a) whether the nonstate entity was created by the State; (b) whether State involvement exceeded the provision of training and financial assistance; (c) whether complete (as opposed to a degree of or potential for) control was exercised in fact; and (d) whether the State selected, installed or paid the political leaders of the group.'

134 *Bosnian Genocide* (n 39) 205 cited in Schmitt (n 18) 88.

135 Christopher Ford makes a similar point in relation to autonomous weapons systems. See Christopher Ford, 'Autonomous Weapons and International Law' (2017) 69 South Carolina Law Review 413, 476.

view of the notion of an entity and in this context it specifically notes that it can refer to a range of bodies, including 'public corporations, semi-public entities, public agencies' and even private companies.[136] What is important however is that this entity has been authorised to exercise what are 'quintessential government functions' such as the conduct of foreign affairs.[137] Therefore, where a software agent with a degree of legal personality is not acting in complete dependence of a State but has nonetheless been empowered to exercise certain elements of government authority — such as defence of government and military networks — it is possible for its actions to be attributed to the State on this basis.

(a) Ultra vires

Where a software agent that is a State organ (either *de jure* or *de facto*) and acts outside the scope of its instructions, the State will be responsible provided it acted with apparent State authority. While it is impossible for a software agent to act contrary to its instructions, it is possible for it to act in an unpredictable, unforeseeable or unintended way. And in each of these cases the State would be responsible for its conduct as the conduct would only be possible as a result of the capabilities programmed into the software agent. Hypothetically, if a software agent acted *ultra vires* its conduct would only be attributable if it acted with the apparent authority of the State. In relation to the signs of apparent authority, unlike in the real world where officials of State organs, such as police offices, may use badges and identification cards, as well as uniforms and official vehicles which are generally necessary to give an appearance of State authority, these kind of indicators are not present in software operating in virtual spaces. However, other indicators used to technically attribute cyber operations generally — such as tradecraft, infrastructure, and intent[138] — could indicate the software agent was acting with the apparent authority of the State, as well as possibly the place of registration of the entity.

2 *Entities directed or controlled by a State*

In addition to articles 4 and 5 of the ILC Articles discussed above, another basis on which the conduct of a software agent can be attributed to the State is pursuant to article 8. Under this provision, even where the software agent is considered a private entity and not an organ of a State or empowered to exercising elements of governmental authority, its actions

136 Crawford (n 30) 100.
137 Schmitt (n 18) 89.
138 Office of the Director of National Intelligence (n 50) 3.

can be attributed to the State provided it is found to be acting under the instructions, directions, or control of a State. Unlike the analysis above which considered whether the non-State actor (for example, a private company using the ACC) is acting under the instructions, direction or control of the State, here the focus shifts to whether the software agent as a separate entity is acting under these bases.

Any software agent, regardless of its sophistication or degree of autonomy, must at some point in time or in some way be created by human programmers. Similarly, where such an entity has a degree of separate legal personhood under a State's domestic law, there would need to be rules and processes around its registration and enforcement of its rights. Even where equipped with (unsupervised) machine learning capabilities, the entity must have been given some direction in terms of its goal or purpose, or the parameters of its functioning. As such, some degree of instructions or directions will always need to be conveyed to these agents. As a result, where the agent is programmed to perform a particular action by a human being whose actions can be attributed to the State, then it will also be clear that the software agent is acting under the instructions or directions of the State. Even where a software agent operates in a closed network environment for extensive periods of time, the initial instructions it would have received from human programmers can provide the basis on which attribution can be established.

The alternative basis for attribution under article 8 is where a software agent is under the 'effective control' of a State.[139] For example, according to the *Tallinn Manual*, a State is in effective control over the actions of a non-State actor where it 'determines the execution and course of the specific operation', where the State has 'the ability to cause constituent activities of the operation to occur', or where the State is able to order the cessation of those activities that are underway.[140] Given the definition of autonomy adopted in this chapter, the requisite degree of control over a software agent would be considered to have been exercised in advance of it being deployed or activated. Even if a software agent acted for extended periods of time in closed networks without direct or real-time human control, this does not mean the State that programmed or developed that entity could claim that it was not under their control. Instead, the control over the entity was simply exercised in advance, and this does not server the control link between the State and the software

139 This is the standard of control adopted by the ICJ and the *Tallinn Manual*. See Schmitt (n 18) 96.
140 ibid.

agent. As such, given the programming that enables and limits the ability of software agents to operate, their conduct would remain under the effective control of the State.

(a) Ultra vires

Some complexity arises in relation to *ultra vires* acts of software agents in this context. The possible suggestion here is that where, for example, a software agent acts in an unforeseeable way or in a closed network environment for an extended period of time this would be outside of its instructions or outside the effective control of the State. However, a software agent cannot act contrary to its programming, therefore it cannot act 'outside' of its instructions. Even where the low level steps the software agent decides to take to achieve its purpose are unforeseeable, it will only be capable of operating according to the capabilities it has in pursuit of its overall high-level goal. As such, any of its actions would be incidental to the mission — that is, done in pursuit of its goal which it has been programmed to do. As discussed above, programming in this context amounts to the control exercised over the software in advance of its deployment or activation, and the State's control over the software is not severed simply by technical autonomy of the system. Accordingly, even if all the decisions made by the software are not understood to human beings, or even if it acts for extended periods of time without real-time or direct human control, those decisions and its conduct would not mean the agent is acting *ultra vires*.

Therefore, even where a State provides a software agent with a degree of legal personality within its domestic legal system, under international law there is little difficulty in conceptualising these as entities or organs within existing rules on State responsibility. This is irrespective of the exact parameters and extent of that entity's legal personality, as the important factor instead is the relationship of that entity to the State and the types of functions it performs. Given the way in which the rules of State responsibility are formulated and the approach to fault, even where a software agent acts in an unpredictable, unforeseeable or unintended way the State will be responsible for its conduct. Some issues do arise in relation to what the primary rule of international law that was violated by or through the use of the ACC — particularly where the element of intention is required for a violation to exist. Otherwise, however, the issues surrounding attribution and ACCs are common with those in relation to cyber operations generally where the challenges to technically attributing these activities often makes legal attribution difficult.

But for the purposes of legal attribution, the technical sophistication or legal autonomy of these entities does not raise significant questions for this area of law different to those raised by cyber operations generally. Ultimately these entities must be created and programmed by human beings and that link will provide the basis for attribution of conduct. As such, even if a State were to give a software agent a degree of legal personality, this could not be used as a means of avoiding responsibility for its actions under international law.

VI
CONCLUSION

This chapter considered the use of ACCs and the attribution of conduct under international law on State responsibility. First, it demonstrated how the attribution of conduct involving the use of ACCs focuses on the actions of the human beings using these capabilities. For this reason, the issues surrounding the attribution of conduct involving the use of ACCs are similar to those in relation to cyber operations generally, and the autonomous capabilities of these systems do not significantly problematize this area of law. Second, it examined the hypothetical but possible future scenario in which software agents are given a degree of legal personality and the implications this has forlegal analysis. Here it was demonstrated that the rules on State responsibility are compatible to these new legal entities on a conceptual level, and that, even where software agents have a degree of legal personality, the link between these entities and the human beings responsible for their creation provides the basis on which their conduct could be attributed to the State.

Chapter 12

Autonomous Cyber Capabilities and Individual Criminal Responsibility for War Crimes

Abhimanyu George Jain[1]

I
INTRODUCTION

Weaponization of the pervasive cyber infrastructure and of nascent autonomous technologies poses difficult challenges for individual criminal responsibility for war crimes arising from the military use of these technologies. For cyber technologies, the problem is identifying the perpetrator of the conduct which corresponds to the *actus reus* of the relevant war crime. For autonomous technologies, there is the problem of the 'responsibility gap': if the impugned conduct is effectuated by an algorithm or by a human relying on an algorithm,[2] there is no human

1 This chapter has benefitted greatly from comments and suggestions from Paola Gaeta, Andrew Clapham, Dorothea Endres, Shri Singh, Alessandra Spadaro and David Stewart, and from the participants at the NATO CCDCOE conference on autonomous cyber capabilities and international law. I retain sole responsibility for any errors.

2 This chapter adopts the definition of 'autonomy' proposed in Tim McFarland, 'The Concept of Autonomy', this volume, ch. 2. However, even recognising that autonomy is a form of control rather than the absence of control (ibid 34) does not necessarily or directly address the responsibility gap in the form discussed here.

being with the *mens rea* required for the crime.[3] Autonomous cyber capabilities ('ACC') compound these two very significant and quite different challenges.

This chapter analyses the challenges posed by ACC to criminal responsibility for war crimes. It does so by considering the practicalities of war crimes prosecution and how these challenges might be addressed in practice. On this basis, it presents two arguments. First, the practical impact of the challenge of identifying perpetrators and of the responsibility gap on individual criminal responsibility may be mitigated in some cases by the practice of charging and adjudicating war crimes. Second, for the remaining cases, the impossibility of criminal responsibility should not be seen as diminishing the enforcement of international humanitarian law ('IHL') but instead, as indicating the preferability of enforcing IHL in these cases by invoking the parallel responsibility of belligerents.

This chapter does not deny the significance of the difficulty of identification or of the responsibility gap and nor does it take on the quixotic burden of resolving them. Instead, it argues for destabilising the unitary and fixed nature of these problems, and for seeing them as variable challenges which may manifest differently in different cases. This differentiated perspective allows for the recognition that in some cases these challenges do not pose insurmountable barriers to prosecution and allows for the refocussing of attention on the residual cases. In turn, the resilience of these challenges in 'residual' cases draws attention to the possibility that criminal responsibility is simply inappropriate in these cases, and that these cases are better addressed in terms of the underlying responsibility of belligerents.

3 The characterisation of this problem in terms of 'responsibility' demands consideration. Existing analyses refer variously to gaps in responsibility and accountability, and on occasion these terms seem to have been used interchangeably in the broader criminal law literature. This chapter takes the view that 'responsibility' refers to the substantive capacity to comply with legal obligations, to be bound by them and to be liable for breach. Liability for breach of the rule presumes responsibility under the rule. 'Accountability' refers to the procedures of enforcing legal obligations upon obligors. See, Renée SB Kool, '(Crime) Victims' Compensation: The Emergence of Convergence' (2014) 10 Utrecht Law Review 14, 16–20; HLA Hart, *Punishment and Responsibility: Essays in the Philosophy of Law* (2nd edn, Oxford University Press 2008) 196–7. Given that what is at issue here is not only the enforcement of the obligation but also the more fundamental question of applicability of the obligation and existence of an obligor, the term 'responsibility gap' is preferred to 'accountability gap'. However, while the problem is better articulated in terms of responsibility than of accountability, from a strictly technical perspective, the problem is narrower than responsibility. The crux of the problem is that of culpability (guilty mind or *culpa*) which, along with actionable conduct, is a necessary requirement for criminal responsibility. Notwithstanding the greater technical accuracy of 'culpability gap', the term 'responsibility gap' is preferred here because the culpability gap necessarily produces a responsibility gap. Moreover, the significance and consequences of the technical problem of culpability become much clearer when seen from the perspective of a gap in responsibility: the absence of a human who can be held criminally responsible for breaches of international humanitarian law, and the consequent gap in the criminal enforcement of international humanitarian law.

The scope of this chapter is limited in several ways. To begin with it is restricted to the prosecution of conduct of hostilities war crimes, which present significant evidentiary challenges even in kinetic contexts, and within that, for reasons of space, to war crimes corresponding to the IHL rule of distinction. However, the analysis presented here may be extended *mutatis mutandis* to other conduct of hostilities war crimes. Further, only prosecution at the International Criminal Court ('ICC') is considered here, although a fuller discussion of this issue would also take account of the prospects of prosecution at the national level. Finally, this chapter assumes the existence of an armed conflict, the applicability of IHL and that the impugned use of ACC constitutes a cyber-attack.[4]

Sections II and III discuss the challenge of identifying perpetrators and the responsibility gap respectively. Section IV concludes with a discussion of the preferability of addressing residual cases through belligerents' responsibility under IHL.

II
IDENTIFYING PERPETRATORS

The difficulties of tracing cyber-attacks and identifying the perpetrators are well-recognised, particularly in the context of State responsibility for breaches of international law rules on uses of force, sovereignty and interventions in States' internal affairs.[5] The crux of the problem lies in identifying the source of the cyber operation and, frequently, in attributing the actions of non-State actors to States in accordance with the international law of responsibility.

4 An account of the questions implicated here is set out in Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd edn, Cambridge University Press 2017) rr 80–5, 92 and accompanying commentary; Kai Ambos, 'International Criminal Responsibility in Cyberspace' in Nicholas Tsagourias and Russell Buchan (eds), *Research Handbook on International Law and Cyberspace* (Edward Elgar 2015).

5 In addition to the chapters by Michael N Schmitt, 'Autonomous Cyber Capabilities and the International Law of Sovereignty and Intervention', this volume, ch 7 and Samuli Haataja, 'Autonomous Cyber Capabilities and Attribution in the Law of State Responsibility', this volume, ch 11, recent contributions to the literature on the challenges of attribution include Nicholas Tsagourias and Michael Farrell, 'Cyber Attribution: Technical and Legal Approaches and Challenges' (2020) 31(3) European Journal of International Law 941; William Banks, 'Who Did It? Attribution of Cyber Intrusions and the Jus in Bello' in Ronald TP Alcala and Eric Talbot Jensen (eds), *The Impact of Emerging Technologies on the Law of Armed Conflict* (Oxford University Press 2019); Hans-Georg Dederer and Tassilo Singer, 'Adverse Cyber Operations: Causality, Attribution, Evidence, and Due Diligence' (2019) 95 International Law Studies 430.

In this context, there is increasing recognition that attribution is not a unique, technical problem with a definite answer: it is an art not a science.[6] It is a political process which necessarily involves subjective assessments, and its nature and results vary depending on the purposes of attribution (public or not), standard of proof, timeframes, attributing agency (political or judicial), etc.[7] In short, attribution as a technical process cannot produce absolute certainty. This rationalisation of expectations has highlighted the related but distinct technical and legal aspects of attribution,[8] and focussed attention on adaptation of legal requirements to technical limitations, including reliance on a 'preponderance of evidence' standard of proof.[9]

When cyber-attacks correspond to the *actus reus* of a war crime, the difficulties of attribution for the purposes of State responsibility are translated into the challenges of identifying the perpetrator of the attack for the purposes of criminal responsibility. In this context, recognising the limitations of technical attribution and turning to a preponderance of evidence standard of proof may not be feasible. For criminal responsibility a presumption of innocence operates until guilt is established 'beyond reasonable doubt'.[10] In the context of both criminal and State responsibility, the common problem is that of identifying the perpetrators, but in the criminal responsibility context this identification must satisfy the requirements of the criminal standard of proof.

This is the problem of identifying the perpetrators of ACC-related IHL breaches. This articulation of the challenge relies on a particular, fixed and rigid idea of 'beyond reasonable doubt' and assumes that this is a very stringent and rigorous standard. This assumption is not supported by the practice of criminal law at either the national or international levels.

Though the generally applicable[11] criminal standard of proof beyond

6 Clement Guitton, *Inside the Enemy's Computer: Identifying Cyber-Attackers* (Hurst & Company 2017). See also, Thomas Rid and Ben Buchanan, 'Attributing Cyber Attacks' (2015) 38 Journal of Strategic Studies 4. These and similar approaches are endorsed in legal analyses by Tsagourias and Farrell (n 5) 4–11; Banks (n 5) 248; Dan Efrony and Yuval Shany, 'A Rule Book on the Shelf? Tallinn Manual 2.0 on Cyberoperations and Subsequent State Practice' (2018) 112 American Journal of International Law 583, 636.

7 Guitton (n 6) 11.

8 Thus, for instance, Rid and Buchanan propose three levels of attribution: tactical (how the attack was conducted), operational (what it entailed) and strategic (who and why). None of the three levels necessarily yields certain answers, but the level of uncertainty increases from the tactical (technical) to the strategic level. See, Rid and Buchanan (n 6).

9 See, eg, Tsagourias and Farrell (n 5).

10 This is the requirement under the Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3 art 66 ('Rome Statute').

11 It is true that civil law systems prefer the '*intime conviction du juge*' standard, but the reasonable doubt standard has a long history in international criminal law, having been adopted and applied by the Yugoslavia and Rwanda tribunals: Salvatore Zappalà, 'The Rights of the Accused' in Antonio Cassese, Paola Gaeta and John RWD Jones (eds), *The Rome Statute of the International Criminal Court: A Commentary* (Oxford University Press 2002) 1346–7. Moreover, the reasonable doubt standard

reasonable doubt is frequently abbreviated to certainty or near certainty,[12] in fact it simply requires the elimination of all alternative possibilities that are reasonable or plausible.[13] There is an extensive body of literature discussing the difficulty of defining and applying the standard, pointing to significantly lower thresholds for conviction in practice in both national and international criminal law.

For instance, in the American context, interpretations of reasonable doubt vary widely,[14] to the extent that definitions deemed acceptable by one court are found by other courts to violate the constitutional rights of the defendant.[15] Empirical research has shown that juries frequently enter convictions based on a perceived probability of guilt ranging from 50–75%.[16]

In the international context, Combs has undertaken an extensive review of trial transcripts at the International Criminal Tribunal for Rwanda, the Special Court for Sierra Leone and the Special Panels in the Dili District Court in East Timor. She highlights a large number of significant infirmities in the evidence relied on for conviction, including extensive reliance on organisational affiliation as evidentiary proxy,[17] ultimately questioning these tribunals' adherence to the requirements of proof beyond reasonable doubt.[18] Her explanation of the handful of acquittals (six) at the Rwanda tribunal is particularly disturbing: 'the inclination of these Trial Chambers to conduct a more searching inquiry into testimonial deficiencies was driven primarily by their sense that the defendant did not generally support the genocide.'[19]

Another aspect of the indeterminacy of the reasonable doubt standard is reflected in the frame of assessment for evidence relating to specific

finds some recognition in international human rights law: Otto Triffterer and Kai Ambos (eds), *The Rome Statute of the International Criminal Court: A Commentary* (3rd edn, CH Beck 2016) 1643.

12 Quantifications of the standard generally characterise it as requiring 90–95% certainty. See, eg, Stephen Wilkinson, 'Standards of Proof in International Humanitarian and Human Rights Fact-Finding and Inquiry Missions' (Geneva Academy of International Humanitarian Law and Human Rights and Geneva Call) 17 <https://www.geneva-academy.ch/joomlatools-files/docman-files/Standards%20of%20Proof%20in%20Fact-Finding.pdf> accessed 10 March 2021.

13 Triffterer and Ambos (n 11) 1645; Zappalà (n 11) 1347.

14 Larry Laudan, *Truth, Error, and Criminal Law: An Essay in Legal Epistemology* (Cambridge University Press 2006) 32–47. He also notes the increasing practice of simply not instructing juries on the meaning of the standard: ibid 47–51. To similar effect, an official training document for criminal trial judges in the UK provides: 'It is unwise to elaborate on the standard of proof…although if an advocate has referred to "beyond reasonable doubt", the jury should be told that this means the same thing as being sure.' See, UK Judicial College, 'The Crown Court Compendium Part I: Jury and Trial Management and Summing Up' (July 2020) 5–1 <https://www.judiciary.uk/wp-content/uploads/2020/07/Crown-Court-Compendium-Part-I-July-2020-09.10.20.pdf> accessed 10 March 2021.

15 Laudan (n 14) 47.

16 Nancy Amoury Combs, *Fact-Finding Without Facts: The Uncertain Evidentiary Foundations of International Criminal Convictions* (Cambridge University Press 2014) 350.

17 ibid 235–72.

18 ibid 189–223.

19 ibid 254.

facts. This question is at the heart of an ongoing and unsettled debate in the case law of the ICC.[20] One side suggests that each piece of evidence relating to a fact should be individually assessed for evidentiary value (for example, reliability) and then all eligible pieces of evidence should be considered together to determine whether they establish the fact in question.[21] Contrasted with this 'atomistic' or 'fragmentary' approach is a 'holistic' approach which proposes collective assessment of all pieces of evidence pertaining to a fact to determine whether, as a whole, they establish the fact in question.[22] An illustration of the difference between these two approaches lies in their treatment of contradictory evidence. An atomistic approach might reject two pieces of evidence altogether based on their mutual inconsistency; a holistic approach might advocate reconciling the inconsistency by reference to the broader evidence adduced in relation to the fact in question.[23]

This difference has the reasonable doubt standard at its centre. Compliance with the standard is much more difficult through the atomistic approach than through the holistic approach.[24] This difference has not yet been resolved, and arguably it never will be. The inherent subjectivity of the idea of 'reasonable doubt', and divisions along the axes of public international law/criminal law and civil law/common law which

20 For academic commentary, see, Mark Klamberg, 'Epistemological Controversies and Evaluation of Evidence in International Criminal Trials' (Stockholm Faculty of Law Research Paper, 19 May 2020) 65 <https://ssrn.com/abstract=3313509> accessed 10 March 2021; Darryl Robinson, 'The Other Poisoned Chalice: Unprecedented Evidentiary Standards in the Gbagbo Case? (Part 1)' (*EJIL: Talk!*, 5 November 2019) <https://www.ejiltalk.org/the-other-poisoned-chalice-unprecedented-evidentiary-standards-in-the-gbagbo-case-part-1/> accessed 10 March 2021; Yvonne McDermott, 'Strengthening the Evaluation of Evidence in International Criminal Trials' (2017) 17 International Criminal Law Review 682. As these authors note, this issue has also been discussed at length in the jurisprudence of other international criminal tribunals.

21 As per the separate opinions referred to note 21 below, this approach has been followed in *Prosecutor v Gbagbo and Blé Goudé* (No Case to Answer Decision) ICC-02/11-01/15-1263 (16 July 2019); *Prosecutor v Ngudjolo Chui* (Appeal Judgment) ICC-01/04-02/12-271-Corr (7 April 2015); *Prosecutor v Katanga* (Trial Judgment) (Minority Opinion of Judge van den Wyngaert) ICC-01/04-01/07-3436-AnxI (7 March 2014).

22 This approach has been endorsed in *Prosecutor v Gbagbo and Blé Goudé* (No Case to Answer Decision) (Dissenting Opinion of Judge Carbuccia) ICC-02/11-01/15-1263-AnxC-Red (16 July 2019) [5], [26]–[51]; *Prosecutor v Ngudjolo Chui* (Appeal Judgment) (Joint Dissenting Opinion of Judges Trendafilova and Tarfusser) ICC-01/04-02/12-271-AnxA (7 April 2015) [31]–[51]; *Prosecutor v Katanga* (Trial Judgment) (Concurring Opinion of Judges Diarra and Cotte) ICC-01/04-01/07-3436-AnxII-tEng (7 March 2014) [4]–[5]. This approach also finds support in *Prosecutor v Lubanga Dyilo* (Appeal Judgment) ICC-01/04-01/06-3121-Red (1 December 2014) [22]: 'In the view of the Appeals Chamber, when determining whether [the reasonable doubt standard] has been met, the Trial Chamber is required to carry out a holistic evaluation and weighing of *all the evidence taken together* in relation to the fact at issue. Indeed, it would be incorrect for a finder of fact to do otherwise.' (emphasis in the original)

23 This example is drawn from *Ngudjolo Chui Appeal Judgment (Trendafilova and Tarfusser)* (n 22) [47]–[51].

24 This is particularly clear in ibid [31]–[41]. Consider, for instance, at [31]: 'The Chamber assessed in isolation individual items of evidence and failed to properly consider the evidence in its entirety. As a result of this approach, the Trial Chamber disregarded trustworthy, coherent and vital evidence which, when pieced together with other relevant and credible evidence, would have provided a solid basis for the determination of the truth.'

characterise even the narrow epistemic community of ICC judges,[25] make the resolution of this difference difficult. For present purposes, the existence of the debate is more interesting than its resolution: the persistence of these differences in the interpretation of a long-standing and well-established standard embellishes its inherent variability and context-specificity.

In sum, shorn of rhetoric and mythology, the reasonable doubt standard is indeterminate, subjective and dependent on context, including the nature of the crime charged and the decider's perception of the defendant.[26] As Lord Justice Denning has noted, 'In criminal cases the charge must be proved beyond reasonable doubt, but there may be degrees of proof within that standard.'[27] And as noted by Damaška, 'it seems psychologically naive to assume that sufficiency of proof requirements do not change in the process of decision-making.'[28]

Nor is this indeterminacy and variability, in and of itself, necessarily problematic. A standard of proof is not a historical, legal or moral necessity. It represents a socio-political decision as to the appropriate allocation of the burden of legal error: a high standard of proof beyond reasonable doubt suggests that acquitting the guilty is seen as far preferable to convicting the innocent.[29] A low standard of proof, conversely, suggests that convicting the innocent is a lesser concern than acquitting the guilty. In the socio-political context of international criminal law, there are a large number of reasons supporting a shifting of this allocation of the burden of legal error. The conviction of the innocent may be seen as less costly than the acquittal of the guilty by reference to, *inter alia*:[30] the investigatory challenges of international prosecutions;[31] the likelihood that a defendant whose case has reached this far bears some responsibility;[32] and, the accountability, deterrence and historical aspects of international criminal trials.

25 See, eg, the qualifications for ICC judges set out in Rome Statute art 36.
26 Combs (n 16) 344–50; Laudan (n 14) 32–51. A well-recognised example is that men tend to be more concerned than women about wrongful conviction for sexual crimes: Combs (n 16) 350.
27 *Bater v Bater* [1950] 2 All ER 458, cited in Combs (n 16) 348.
28 Mirjan Damaška, 'Evidentiary Barriers to Conviction and Two Models of Criminal Procedure: A Comparative Study' (1973) 121 University of Pennsylvania Law Review 506, 542.
29 Laudan (n 14) 1–2.
30 Combs (n 14) 350–9; Fergal Gaynor and others, 'Law of Evidence' in Göran Sluiter and others (eds), *International Criminal Procedure: Principles and Rules* (Oxford University Press 2013) 1148. The impracticability and undesirability of excessively fastidious application of the reasonable doubt standard in international criminal law features in criticism of the ICC's recent decision that Laurent Gbagbo and Charles Blé Goudé had no case to answer, by the dissenting judge and academics alike. See, *Gbagbo and Goudé No Case to Answer Decision (Carbuccia)* (n 22) [6]–[7]; Robinson (n 20).
31 See, eg, *Katanga Trial Judgment (Diarra and Cotte)* (n 22) [5].
32 See, eg, above (n 19) and accompanying text.

None of this implies a rejection of the reasonable doubt standard, and nor should it be interpreted to support a dilution of the standard for international crimes in general or for ACC-related war crimes in particular. The objective of the foregoing analysis is simply to draw attention to the well-recognised variability and context-specificity of the reasonable doubt standard. The mythologies of certainty surrounding the standard should not obfuscate the inherent and unavoidable contingency of any factual determination.[33]

Against this rationalised understanding of the reasonable doubt standard and its operation in practice, the difficulties of identifying perpetrators of ACC-related war crimes may not be as significant or as uniform as they seem at first sight.

III
RESPONSIBILITY GAP

This section discusses the challenge of the responsibility gap and proceeds in three sub-parts. Section III.A introduces the responsibility gap. Sections III.B and III.C discuss two features of international war crimes prosecution which may operate to mitigate the responsibility gap to varying degrees — the practicalities of proving *mens rea* (III.B) and the in-built seriousness requirement in the ICC's jurisdiction (III.C).

A INTRODUCING THE RESPONSIBILITY GAP

The responsibility gap presents a conceptual problem for criminal responsibility for ACC-related war crimes, based on the impossibility of a culpable human ie, a human who has the *mens rea* necessary for criminal responsibility.

33 The contingent nature of determinations of 'fact' is well-recognised in the rich literature on the epistemological philosophy of criminal law and fact-finding but seemingly under-recognised in the practice and promise of criminal law. For a review of this literature, see Simon de Smet, 'Justified Belief in the Unbelievable' in Morten Bergsmo and Carsten Stahn (eds), *Quality Control in Fact-Finding* (Torkel Opsahl 2020). For a persuasive account of the role of judges in 'constructing' facts in international adjudication, providing an explanation for the deficiencies identified by Combs in the fact-finding practice in international criminal tribunals (above n 16–18 and accompanying text), see, Ana Luísa Bernardino, 'The Discursive Construction of Facts in International Adjudication' (2020) 11 Journal of International Dispute Settlement 175.

Under the Rome Statute, the default *mens rea* or mental element requirement is intention and knowledge.[34] The Rome Statute war crimes relating to the IHL rule of distinction refer to '[i]ntentionally directing attacks against the civilian population as such or against individual civilians not taking direct part in hostilities' in international and non-international armed conflicts and '[i]ntentionally directing attacks against civilian objects, that is, objects which are not military objectives' in international armed conflicts.[35] These crimes of attacking civilian targets have been interpreted as crimes of conduct which do not require a specific result,[36] and as requiring the deliberate launching of an attack against a target known to be civilian in nature.[37]

In the event of ACC-related breaches of IHL the stringent *mens rea* requirement of the Rome Statute may prove a barrier to war crimes prosecutions and criminal responsibility.[38] The soldier who deploys or relies on ACC knowing that civilian targets will be attacked and intending to attack them satisfies the *mens rea* requirement and may be criminally responsible if the *actus reus* requirements are met.[39] The soldier who has no reason to doubt the prospective IHL-compliance of ACC but is implicated in a breach of IHL is not culpable and need not concern us further. It is the soldier who has reason short of certainty to doubt the deployment of or

34 Rome Statute art 30. The rather confusing structure of art 30 necessitates a brief explanation of its requirements. The requirement of 'intent and knowledge' in art 30(1) does not mean that both together constitute the default, but instead that the general mental element (*mens rea* or *dolus*) under the Rome Statute has both volitional (intent or purpose or wanting) and cognitive (knowing or awareness) elements. One or both may constitute the requirement for specific war crimes, and arts 30(2) and 30(3) go on to define what each means. Thus, art 30 envisages neither a default separation of intent and knowledge nor a default conjunction. See, International Criminal Court, 'Elements of Crimes of the International Criminal Court' (2011) General Introduction, [2] < https://www.icc-cpi.int/resourcelibrary/official-journal/elements-of-crimes.aspx#intro> accessed 10 March 2021; Triffterer and Ambos (n 11) 1117; Elies van Sliedregt, *Individual Criminal Responsibility in International Law* (Oxford University Press 2012) 46; Albin Eser, 'Mental Elements — Mistake of Fact and Mistake of Law' in Antonio Cassese, Paola Gaeta and John RWD Jones (eds), *The Rome Statute of the International Criminal Court: A Commentary* (Oxford University Press 2002) 904–8.

35 Rome Statute arts 8(2)(b)(i), 8(2)(b)(ii) and 8(2)(e)(i). See also, the elements of these crimes in Elements of Crimes of the International Criminal Court (n 34). Given the similar structure and elements of these crimes, in the following analysis they will be treated as co-extensive and grouped under the common rubric of 'war crimes of attacks against civilian targets'.

36 *Prosecutor v Ntaganda* (Trial Judgment) ICC-01/04-02/06-2359 (8 July 2019) [904]; *Prosecutor v Katanga* (Trial Judgment) ICC-01/04-01/07-3436-tENG (7 March 2014) [799].

37 *Ntaganda Trial Judgment* (n 36) [903], [917], [921]; *Katanga Trial Judgment* (n 36) [808]. Though these war crimes explicitly require 'intentionally directing attacks', the *Katanga* trial chamber has confirmed that the reference to 'intentionally' does not amount to a specific *mens rea* requirement distinct from the art 30 default: ibid [806].

38 As noted above (n 2), recognising that autonomy is a form of control rather than its absence, and that ACC are developed by humans and operate within human-defined parameters, does not necessarily mitigate the responsibility gap in relation to soldiers deploying or relying on ACC in active hostilities.

39 It is possible that in cases of deployment of ACC there may also be questions as to whether the *actus reus* component of the war crime has been satisfied. This possibility arises from the *Ntaganda* trial judgment where the ICC interpreted the requirement of 'directing attacks' as 'selecting the intended target and deciding on the attack'. See *Ntaganda Trial Judgment* (n 36) [917].

reliance on ACC, that is, the soldier who acts negligently or recklessly or with *dolus eventualis*, who is difficult to accommodate within the mental element required by the Rome Statute. It is impossible that this soldier *knew* that a civilian target would be attacked and *intended* that attack, and it is this soldier who bestrides the responsibility gap, rather like a colossus.[40]

This responsibility gap is not simply a theoretical proposition. The nature of ACC means that they act and respond to their environments independently, and human involvement is restricted to controlling or supervising very sophisticated and possibly unpredictable technologies.[41] In this context there is a transfer of agency which diminishes the possibility of culpability of the soldier.

Early articulations of the responsibility gap hypothesised weapon systems which could autonomously make and execute attack decisions.[42] More recent literature points instead to the application of autonomy at multiple stages of the targeting process in support of human decision-making.[43] Thus, for instance, autonomous technologies might be used for intelligence analysis, for generating possible targets, for target-identification, for assessing collateral damage, etc.

The shift from deploying autonomous technologies to relying on them for human decision-making changes the responsibility gap: it makes it less obvious but no less significant. In cases of deployment, the conduct which constitutes a breach of IHL (and consequently, possibly, the *actus reus* for the corresponding war crime) is effectuated by the weapon system and the only proximate human conduct is the decision of deployment. The culpability of this deployment is a necessary precondition for criminal responsibility. In cases of reliance, the conduct which constitutes a breach of IHL is effectuated by a human, but in reliance on autonomous technologies. This reliance may be unwarranted or unjustified or compromised by cognitive biases such as over or under-reliance and cognitive

40 This formulation of the problem is drawn from Shakespeare's Julius Caesar, and is deployed to reflect the vast consternation and analysis the problem has spawned. The substance of the underlying problem is drawn from Neha Jain, '*Autonomous Weapons Systems: New Frameworks for Individual Responsibility*' in Nehal Bhuta and others (eds), Autonomous Weapons Systems: Law, Ethics, Policy (Cambridge University Press 2016) 315; Jens David Ohlin, 'The Combatant's Stance: Autonomous Weapons on the Battlefield' (2016) 92 International Law Studies 1, 21–2.

41 Vincent Boulanin and others, 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control' (Stockholm International Peace Research Institute and International Committee of the Red Cross, 2 June 2020) ix <https://www.icrc.org/en/document/limits-autonomous-weapons> accessed 10 March 2021.

42 In addition to the sources above (n 40), see also, Rebecca Crootof, 'War Torts: Accountability for Autonomous Weapons' (2016) 164 University of Pennsylvania Law Review 1347.

43 Merel AC Ekelhof, 'Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control through the Lens of Military Targeting' (2018) 71 Naval War College Review 61.

overloading.[44] In these cases, criminal responsibility requires establishing the culpability of this reliance.

Thus, the shift from deployment to reliance on autonomous technologies shifts the locus of the culpability assessment.[45] But while this shifts the source of the responsibility gap, it does not change it. In both cases the key underlying premise of the responsibility gap — the questionable culpability of the soldier who is negligent or reckless in their deployment of or reliance on autonomous technologies remains the same.[46]

For present purposes, it is not necessary to assess the exact nature and scope of the change in the responsibility gap across deployment and reliance, and it suffices to note that these may constitute two different but related challenges. Indeed, in the context of ACC both versions of the responsibility gap (deployment and reliance) are relevant given the existing state of the technology and expected trajectories of development.[47] The singular exception to the emerging consensus[48] regarding the requirement of meaningful human control for autonomous weapons systems relates to defensive applications of ACC.[49] In this context, the original responsibility gap thesis, centred around the increasingly obsolete trope of deployment of 'killer robots' without a human in or on the 'loop', regains prominence and applies in parallel to the revised responsibility gap thesis relating to reliance on ACC.

44 Marta Bo, 'The Human-Weapon Relationship in the Age of Autonomous Weapons and the Attribution of Criminal Responsibility for War Crimes' (2019) <https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/Bo_Human-Weapon-Relationship.pdf> accessed 10 March 2021. The problem posed by biases should be treated with some caution. It is undeniably true that the question of biases is particularly significant in the context of human-machine interaction and teaming, and all the more so in relation to autonomous technologies where the interaction poses existential challenges for the very meaning of human agency and control. However, the fact remains that biases are an unavoidable (and in their role as heuristics, possibly necessary) aspect of human cognition which can never be eliminated but can only be accounted for and managed.

45 As to the relationship between culpability and the responsibility gap, see the discussion above (n 3).

46 The shift from deployment to reliance may change the responsibility gap in one more way, through multiplication in the instances of human-machine interaction, and proliferation of consequent questions of culpability and responsibility in relation to a single attack. However, determining whether this constitutes a change, in what way and to what extent depends on the specificities of the autonomous weapons being deployed and the autonomous technologies being relied on, and cannot be assessed further in the abstract.

47 Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) 11–13 <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/> accessed 10 March 2021.

48 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 2019, 'Report of the 2019 Session' (25 September 2015) UN Doc CCW/GGE.1/2019/3, [21]–[22].”plainCitation”:”'Report of the 2019 Session' (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 2019

49 Tanel Tammet, 'Autonomous Cyber Defence Capabilities', this volume, ch 3, section IV; Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (WW Norton & Co 2018) ch 14.

This brief introduction to the nature of the responsibility gap explains the furore it has generated. The soldier who deploys or relies on ACC with reason short of certainty to doubt the IHL compatibility of the ensuing actions exposes a gap in the criminal enforcement of IHL.[50] Though there is no conceptual solution to the responsibility gap, its practical significance may be mitigated by two features of the practice of charging and adjudicating war crimes at the ICC. It is to the first of these two features that the next sub-section turns.

B *MENS REA* IN PROBATIVE PRACTICE

The first feature of the practice of war crimes prosecution which may mitigate the responsibility gap concerns the modalities of proving mental elements or *mentes reae*. It will be argued here that the common practice of inferring intent from conduct and circumstances provides limited mitigation of the responsibility gap by shifting focus from what the soldier actually knew and intended to what they *must have* known and *therefore* intended.

Criminal law, both national and international, relies on a strict application of Cartesian dualism — the distinction between body and mind, according to which criminal responsibility requires the conjunction of *actus reus* and *mens rea*.[51] The distinction has been the subject of criticism and critique in psychology and neuroscience,[52] and in criminal theory,[53] but it endures in criminal law.[54] However, the insistence on a guilty mind in addition to proscribed conduct raises evidentiary challenges because 'substantive rules regarding the mental element require the actual occurrence of a subjective mental state, whereas the law of evidence can provide only an assumption that the required state may have occurred.'[55]

50 Academic analyses have turned instead to theories of indirect perpetration, including command responsibility. See, eg, Russell Buchan and Nicholas Tsagourias, 'Command Responsibility and Autonomous Cyber Weapons', this volume, ch. 13; Jain (n 40); Ohlin (n 40).

51 Jeroen Blomsma, *Mens Rea and Defences in European Criminal Law* (Intersentia 2012) 41. This is expressed in the Latin phrase: '*actus non facit reum nisi mens sit rea*'.

52 Dov Fox and Alex Stein, 'Dualism and Doctrine' in Dennis Patterson and Michael S Pardo (eds), *Philosophical Foundations of Law and Neuroscience* (Oxford University Press 2016).

53 Antony Duff, *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law* (Blackwell 1990).

54 On this gap between philosophy and criminal law, see George P Fletcher, *Rethinking Criminal Law* (Oxford University Press 2000) 451–2.

55 Keren Shapira-Ettinger, 'The Conundrum of Mental States: Substantive Rules and Evidence Combined' (2007) 28 Cardozo Law Review 2577, 2685. See also Fletcher (n 54) 120. Hart neatly encapsulates the resonance of this concern and its rejection by juxtaposing the 15th century dictum of Chief Justice Bryan — 'The thought of man is not triable; the devil alone knoweth the thought of man' — with that of Lord Justice Bowen in the 19th century — 'the state of a man's mind is as much a fact as the state of his digestion.' Hart (n 3) 188.

In practice, this hurdle is overcome by inferring *mens rea* from conduct and circumstances.[56] It is uncontroversial, for instance, that the intent to murder can be inferred from the act and context of stabbing the victim in the stomach.[57] This constraint posed by the law of evidence upon the ideals of the criminal law is not a rejection of the *mens rea* requirement.[58] It is simply an acknowledgment that there are acts and circumstances (eg, stabbing somebody in the stomach) for which a particular mental state (intention) is the only possible, *though still rebuttable*,[59] conclusion.[60]

This probative practice — inferring mental element from conduct and circumstances which allow for no reasonable alternative explanation — is also well-established in international war crimes prosecutions.

The International Criminal Tribunal for the Former Yugoslavia ('ICTY') has consistently held that intent for the war crime of attacking civilians:[61]

56 Jens David Ohlin, *Criminal Law: Doctrine, Application, and Practice* (Wolters Kluwer 2016) 167 (American law); AP Simester and others, *Simester and Sullivan's Criminal Law: Theory and Doctrine* (6th edn, Hart 2016) 147–8 (English law); Michael Bohlander, *Principles of German Criminal Law* (Hart 2009) 65 (German law); *State of Maharashtra v Mohd Yakub s/o Abdul Hamid & Ors* [1980] SCR (2) 1158 (Supreme Court of India) 1163–4 (Indian law); *X und Y gegen Staatsanwaltschaft des Kantons Luzern sowie Obergericht des Kantons Luzern* (Urteil des Kassationshofes) [8.4] (Swiss law); Blomsma (n 51) 54–8 (Dutch, English, German and European law); Thomas Weigend, 'Subjective Elements of Criminal Liability' in Markus D Dubber and Tatjana Hörnle (eds), *The Oxford Handbook of Criminal Law* (Oxford University Press 2014) 508 (generally). By way of example, s 8 of the UK Criminal Justice Act 1967 provides: 'Proof of criminal intent. A court or jury, in determining whether a person has committed an offence, — (a) shall not be bound in law to infer that he intended or foresaw a result of his actions by reason only of its being a natural and probable consequence of those actions; but (b) shall decide whether he did intend or foresee that result by reference to all the evidence, drawing such inferences from the evidence as appear proper in the circumstances.' On this issue more generally, see, Shapira-Ettinger (n 55).

57 In itself this practical reality represents the endorsement of Wittgenstein's observation: 'An "inner process" stands in need of outward criteria.' Ludwig Wittgenstein, Philosophical Investigations (4th edn, Wiley-Blackwell 2009) s 580. On the challenges and difficulties of defining and divining intention from the perspective of analytical philosophy, see, GEM Anscombe, *Intention* (2nd edn, Harvard University Press 2000).

58 The separate but intertwined nature of *mens rea* and *actus reus*, and the possibility of divining one from the other finds a parallel in the relationship between State practice and *opinio juris* for the determination of customary international law. For instance, though the International Law Commission's draft conclusions on the identification of customary international law emphasise the distinct but conjunctive requirements of State practice and *opinio juris*, they also recognise the possibility of inferring *opinio juris* from State practice. See, International Law Commission, 'Draft Conclusions on Identification of Customary International Law, With Commentaries' (Commentary to Conclusion 2, [1]; Commentary to Conclusion 10, [3]).

59 Compare, for instance, UK Criminal Justice Act 1967 ss 8(a) and 8(b), quoted above (n 56).

60 This is in keeping with the observation (n 12 and accompanying text) that the requirement of proof beyond reasonable doubt does not require absolute certainty, but merely the elimination of all plausible alternatives.

61 *Prosecutor v Dragomir Milošević* (Trial Judgment) ICTY-98-29/1-T (12 December 2007) [948]. See also, *Prosecutor v Galić* (Appeal Judgment) ICTY-98-29-A (16 November 2006) [132]; *Prosecutor v Blaškić* (Trial Judgment) ICTY-95-14-T (3 March 2000) [501]–[512]; *Prosecutor v Kupreškić et al* (Trial Judgment) ICTY-95-16-T (14 January 2000) [513]; Héctor Olásolo, *Unlawful Attacks in Combat Situations: From the ICTY's Case Law to the Rome Statute* (Martinus Nijhoff 2008) 76–8.

> can be inferred from many factors, including the means and method used in the course of the attack, the status and number of the victims, the nature of the crimes committed, the extent to which the attacking force may be said to have complied or attempted to comply with the precautionary requirements of the laws of war and the indiscriminate nature of the weapon used.

Consider, for instance, the *Galić* trial chamber's discussion of 'Scheduled Shelling 5', a shell-strike on Markale open-air market on 5 February 1994. It engaged at great length with a range of expert evidence to determine the source and direction of the attack, and concluded that the shell in question was fired from territory controlled by the Sarajevo Romanija Corps (a unit of the Bosnian-Serb Army commanded by General Galić) and was aimed at the market.[62] Its subsequent discussion of the legal characterisation of the attack referred simply to the absence of military targets in the vicinity,[63] and on this basis it found that civilians had been made the object of attack intentionally or recklessly.[64] In other words, from a seeming lack of military justification for attacks, the ICTY has been willing to infer the *mens rea* of the war crime of attacking civilian objectives.[65]

The 'elementary proposition'[66] that intent can be inferred from conduct and circumstances is also well-recognised at the ICC. For instance, the General Introduction to the Elements of Crimes of the International Criminal Court expressly provides: 'Existence of intent and knowledge can be inferred from relevant facts and circumstances.'[67]

62 *Prosecutor v Galić* (Trial Judgment) ICTY-98-29-T (5 December 2003) [438]–[494]. Judge Nieto-Navia disagreed with the majority on this finding: *Prosecutor v Galić* (Trial Judgment) (Separate and Partially Dissenting Opinion of Judge Nieto-Navia) ICTY-98-29-T (5 December 2003) [71]–[97]. The appeals chamber upheld the trial chamber's decision: *Galić Appeal Judgment* (n 61) [314]–[335].

63 *Galić Trial Judgment* (n 62) [495]–[496]. This was affirmed by the appeals chamber: *Galić Appeal Judgment* (n 61) [334]–335].

64 *Galić Trial Judgment* (n 62) [596].

65 To be clear, the jurisprudence of the ICTY has no formal significance before the ICC. Indeed, given the that the war crime of attacking civilians and civilian objects requires intent or recklessness in the ICTY's case law and intent under the Rome Statute, the ICTY's jurisprudence is technically irrelevant to the interpretation of the mental element of the war crimes defined in Rome Statute arts 8(2)(b)(i), 8(2)(e)(i) and 8(2)(e)(ii). Compare the ICC's interpretation of the war crimes of attacking civilian targets, *supra* notes 35–36 and accompanying text, with that of the ICTY in, eg, *Dragomir Milošević Trial Judgment* (n 61) [951]; *Galić Appeal Judgment* (n 61) [140]. Nonetheless, the ICC has referred extensively to the jurisprudence of the ICTY, including in relation to these particular war crimes and specifically in relation to proof of mental elements. See, eg, *Ntaganda Trial Judgment* (n 36) [921]; *Katanga Trial Judgment* (n 36) [807]. Consequently, the probative practice of the ICTY in relation to *mens rea* has been referred to here, first, because of its influence upon the probative practice of the ICC, and second, to demonstrate the pedigree of the specific probative practice of inferring intent from conduct and circumstances in international war crimes prosecutions.

66 Triffterer and Ambos (n 11) 1117.

67 Elements of Crimes of the International Criminal Court (n 34) [3].

Similarly, the 'Means of Proof Digest' of the ICC's Case Matrix expressly endorses the possibility of inferring intent to attack civilian targets from conduct and context.[68] For proving that the perpetrator intended to make civilians the object of attack in relation to Rome Statute articles 8(2)(b)(i) and 8(2)(e)(i), the digest refers to means of proving the 'Knowledge of the perpetrator about the civilian status of the object of the attack' and means of proving the 'Intent of the perpetrator to target civilians'. In relation to the latter, it refers to:

> (a) evidence of the absence of military objects and/or military activity in the vicinity of the attacked area; (b) evidence showing that no military objects, real or believed, in the attacked area were targeted; (c) evidence of the extensive targeting of non-military objects in and around the attacked area concerned; (d) evidence of repeated shooting on civilians; (e) evidence of the indiscriminate nature of a weapon employed; (f) evidence of failure to take all necessary precautions to avoid injury, loss or damage to the civilian population; (g) evidence disproving accident caused by stray or ricocheting bullet; and, (h) evidence showing that the market area being attacked drew large number of people.[69]

None of these elements directly establishes an intention to attack civilians but instead, infers it from conduct and circumstances. Individually or collectively, these elements operate to discard alternative explanations for attacks on civilian targets, leaving intention to do so as the only plausible remaining possibility.

This probative practice — inferring mental element from conduct and circumstances which allow for no reasonable alternative explanation — is

68 International Criminal Court, 'Means of Proof Digest of the International Criminal Court' <https://cilrap-lexsitus.org/means-proof-digest> accessed 10 March 2021. The Means of Proof digest is not a formal ICC publication, and nor does it have any authority before the ICC. However, the Means of Proof Digest and the Case Matrix Network and Legal Tools, of which it forms a part, were developed at the ICC, though the updating and maintenance of these tools have now been outsourced and the ICC disclaims any responsibility for their content. Thus, the Means of Proof Digest is not an authoritative source. However, it does provide a valuable guide to understanding how specific questions of law and procedure have been addressed in international criminal trials, at the ICC and also at other international criminal tribunals. It is in this capacity that it is referred to here.

69 Similarly, to prove that the perpetrator intended to make civilian objects the object of attack under Rome Statute art 8(2)(b)(ii), the digest refers to: '(a) evidence with regard to ability to target with precision; (b) evidence inferred from indiscriminate targeting; (c) evidence of the non-existence of military objects in the attacked area; (d) evidence of the demarcation etc being obvious to the perpetrator at the time of the attack; (e) evidence of the extensive targeting of non-military objects in and around the attacked area concerned.'

also evident in the case law of the ICC. Trial chambers have repeatedly endorsed the probative value of circumstantial evidence.[70] In relation to the war crime of attacking civilians in a non-international armed conflict, echoing the ICTY,[71] the *Katanga* trial chamber said that the intent to make civilians the object of attack:[72]

> may be inferred from various factors establishing that civilians not taking part in the hostilities were the object of the attack, such as the means and methods used during the attack, the number and status of the victims, the discriminatory nature of the attack or, as the case may be, the nature of the act constituting the attack.

Similarly, the *Ntaganda* trial chamber has held that in relation to attacks against co-located civilian and military targets, lack of discrimination or precaution in attack may constitute an attack against civilian targets.[73]

It must be emphasised that notwithstanding these broad formulations of proof of intent to attack civilian targets in the jurisprudence of the ICTY and the ICC, in practice both courts have usually been able to rely on far more specific and concrete evidence. Thus, for instance, the *Ntaganda* trial chamber relied on the use of the phrase '*kupiga na kuchaji*' by the defendant in ordering attacks, a term which the chamber interpreted as an exhortation to attack the entire Lendu community without distinction as to civilian and combatant.[74]

Most cases involving conduct of hostilities war crimes which have been tried before international courts and tribunals so far have featured inter-ethnic strife where entire communities are targeted, regardless of civilian or combatant status. In these contexts, it is unsurprising that trial chambers have frequently been able to rely on more specific evidence of intent to attack civilian targets. This should not, however, detract from the significance of broad endorsements of the possibility of inferring intent from conduct and circumstances. Indeed, the consistency of the ICC (and the ICTY) in maintaining the validity of this probative practice

70 *Ntaganda Trial Judgment* (n 36) [69]–[70]; *Prosecutor v Bemba Gombo* (Trial Judgment) ICC-01/05-01/08-3343 (21 March 2016) [239]; *Katanga Trial Judgment* (n 36) [109]; *Prosecutor v Lubanga Dyilo* (Trial Judgment) ICC-01/04-01/06-2842 (5 April 2012) [111].

71 Above (n 61) and accompanying text.

72 *Katanga Trial Judgment* (n 36) [807].

73 *Ntaganda Trial Judgment* (n 36) [921].

74 ibid [415], [484], [922], [1181]. See also, *Katanga Trial Judgment* (n 36) [850]–[855]. The extent of the *Ntaganda* trial chamber's reliance on the use of the phrase 'kupiga na kuchaji' and the questionable correctness of its interpretation is highlighted in *Prosecutor v Ntaganda* (Defence Appeal Brief — Part II) [2020] ICC-01/04-02/06-2465-Red-Corr (30 June 2020) [75]–[90].

despite its limited utility in the specific cases before them embellishes its significance and pedigree.[75]

In sum, though intent can be established through an insight into the perpetrator's mind — for example, through a confession or witness testimony — it can also be inferred from conduct and circumstances which do not admit of alternative explanation.

Once we acknowledge this probative practice it becomes evident that the responsibility gap thesis rests on a false premise. It focusses only on the subjective state of mind of the deploying or relying soldier and ignores the possibility of inferring intent from the manner and context of deployment or reliance.

The probative practice outlined here makes it possible to shift the focus of the *mens rea* analysis from the subjective state of mind of the soldier to the more objective manner and context of deployment or reliance of ACC. This shift in the focus of the inquiry produces a shift in the nature of the inquiry: from an inquiry into what the soldier knew and intended to what the soldier must have known and consequently intended. The former posits an ambitious inquiry into the actual knowledge and intent of the soldier, which is invariably impossible in the absence of a confession or verifiable declaration of intent.[76] The latter resolves this evidentiary difficulty by invoking a standard of reasonableness to infer constructive knowledge and on this basis presuming intent, subject to rebuttal.

The mechanics of this shift in the *mens rea* analysis can be seen in the example of the person who stabs another in the stomach. As discussed, it is uncontroversial that it is possible to infer intent to murder from the act of stabbing somebody in the stomach. The chain of reasoning here may be broken down as follows. First it is necessary to posit that a reasonable person would recognise the fatal consequences of stabbing another person in the stomach. On the basis of this standard of reasonableness we can assume, subject to rebuttal, that the perpetrator recognised the fatal consequences of their action. On the basis of this constructive knowledge we can assume, again subject to rebuttal, that the perpetrator intended the consequences of their action.[77] In this manner we can infer intent to murder from the act of stabbing somebody in the stomach.

75 For instance, compare [865] of the *Katanga Trial Judgment* (n 36) in which the chamber summarises its factual findings in relation to one part of the impugned attack, and the prior summary of evidence it relies on para VIII(B)(2)(b) of the judgment. The former is phrased far more generally (even allowing for the limitations of summary) than the latter.

76 In this regard, see above (n 55 and accompanying text).

77 On the well-established nature of both these assumptions — of foresight of consequences and of their intendment, see Hart (n 3) 175.

In this way the possibility of inferring intent from conduct and circumstances effectively skirts the responsibility gap. Whether the soldier knew that their deployment of or reliance on ACC would result in an attack on a civilian target and whether they intended that attack is still useful, but it is not determinative of the criminal responsibility of that soldier. Criminal responsibility can equally be based on what the soldier must have known and therefore what they can be presumed to have intended. What they must have known can be derived from the manner and circumstances of deployment of ACC, including what was known about the performance of the ACC, what its operational abilities and constraints were, what precautions were taken,[78] the context of deployment or reliance, etc.[79] If the soldier must have known that their deployment of or reliance on ACC would result in an attack on a civilian target, it may be presumed, subject to rebuttal, that the attack was intended, and would implicate the criminal responsibility of the soldier for the war crime of attacking civilian targets.

This approach to the war crimes of attacking civilian targets finds support in the reasoning of the *Ntaganda* trial chamber. The trial chamber broke the crime of attacking civilians into two requirements: directing attacks; against civilians. It defined the first requirement as 'selecting the target and deciding on the attack'.[80] Turning to the second requirement, it went on to say:[81]

> As the burden of proof lies with the Prosecution, it must be established that in the circumstances at the time, *a reasonable person could not have believed* that the individual or group he or she attacked was a fighter or directly participating in hostilities. (emphasis added)

In effect, the trial chamber is using the standard of what a reasonable person must have known to determine knowledge of civilian status, and from a deliberate and otherwise unjustified attack against this target, it is willing to infer intent to attack civilians.

78 In his contribution to this volume Eric Talbot Jensen argues that the obligation to take precautions can be fulfilled by autonomous weapons themselves, provided that 'rigorous weapons review processes [are] in place that continually examine the autonomous system's continued "learning"'. See, Eric Talbot Jensen, 'Precautions and Autonomy in the Law of Armed Conflict', this volume, ch 9, section IV. If this argument is correct, 'what must have been known' may still be derived from failure to conduct the required ongoing reviews or shortcomings in the reviews.

79 In this regard, see the dicta of chambers of the ICTY and ICC quoted above (n 61 and 71–72) and accompanying text.

80 *Ntaganda Trial Judgment* (n 36) [917].

81 ibid [921].

A similar willingness to replace knowledge with constructive knowledge is evident in the *Katanga* trial chamber's interpretation of awareness that a consequence 'will occur in the ordinary course of events' in Rome Statute article 30(2)(b) as 'virtual certainty' of occurrence.[82] It went on to describe virtual certainty in the following terms: 'it is nigh on impossible for him or her to envisage that the consequence will not occur.'[83]

In summary, absent a confession or other declaration of intent, probative limitations are *constitutive* of the *mens rea* requirement. The *mens rea* of the soldier deploying or relying on ACC may, if possible, be determined by what they actually knew and therefore intended. But it can equally be determined by reference to what they must have known and intended. It is not necessary for the prosecution to establish knowledge of civilian status; it suffices to demonstrate that it was impossible not to have known of civilian status.

The responsibility gap thesis ignores the possibility of this shift from actual to constructive knowledge, from knowing to the impossibility of not knowing. As demonstrated here, this probative practice mitigates the responsibility gap to some extent. The scope and extent of this mitigation is, however, subject to three important restrictions and clarifications.

First, presumptions of what must have been known and consequently intended, that is presumptions of recognition of consequences and thereby of intendment, are rebuttable. Thus, if what must have been known was not in fact known, the inference of intent may be rebuttable.[84]

Second, a clarification is necessary as to the role of the reasonableness standard here.

The standard of reasonableness provides a perspective for assessment, it does not determine the substance of the assessment. The role of the reasonable person is to provide a benchmark of comparison. Whether the substance of the comparison is what the reasonable person *should* have known or what the reasonable person *must* have known depends on the underlying rule. In relation to the war crimes of attacking civilian targets in the jurisprudence of the ICC, the reasonableness standard is deployed to determine what a reasonable person in similar circumstances

82 *Katanga Trial Judgment* (n 36) [776].

83 ibid [777]. The original French text of the judgment provides: '*il lui est à peu près impossible d'envisager que la conséquence ne surviendra pas.*'

84 In this regard, it is interesting to note that the particular phrasing of UK Criminal Justice Act 1967 s 8, cited above (n 56) as support for the probative possibility of inferring intent from conduct and circumstances, was specifically intended to preserve the possibility of this inference, while ensuring that it remained rebuttable. See Hart (n 3) 175.

must have known.[85] This is a significantly more stringent requirement than that of what a reasonable person should have known.[86]

It bears emphasis that the very contingent[87] idea of the 'reasonable person' in this case refers to the reasonable military commander. Compliance with the IHL rule of distinction is determined by reference to the standard of the reasonable commander,[88] and it stands to reason that the same reasonable commander would provide the benchmark of reasonableness for the purposes of the corresponding war crimes. This conclusion is also supported by the explicit connection drawn by the ICC between the war crimes of attacking civilian targets and the IHL rule of distinction.[89]

In effect then, the mitigating influence of this probative practice on the responsibility gap will be limited to particularly egregious cases. The soldier in the responsibility gap will be deemed to have the required *mens rea* only when it is impossible for a reasonable commander in their position not to have known that deployment of or reliance on ACC would result in an attack on a civilian target. This is an important and significant limitation.

Third, the foregoing analysis raises undeniable concerns regarding the conflation of intent and recklessness or *dolus eventualis*.[90]

The simple answer to this concern is to acknowledge it. Inferring intent from conduct and circumstances, in shifting focus from what the soldier knew and intended to what must have been known and intended, assimilates the most egregious cases of recklessness or *dolus eventualis* into intent. This concern is valid, but it is also not new.[91] Moreover, it is mitigated to varying degrees by, first, the inherent limitation to the most egregious cases of recklessness or *dolus eventualis*; and, second, the continuing possibility of rebutting the inference of intent. In this regard, it bears emphasis that this chapter does not propose this conflation but

85 Above (n 81–83) and accompanying text.

86 The jurisprudence of the ICTY is inconsistent on this point. Some cases have used a 'should have known' standard: *Dragomir Milošević Trial Judgment* (n 61) [952]; *Galić Trial Judgment* (n 62) [55]. Others have used an 'impossibility of not knowing' standard: *Prosecutor v Strugar (Trial Judgment)* ICTY-01-42-T (31 January 2005) [280]; *Blaškić Trial Judgment* (n 61) [180].

87 Hart (n 3) 171: 'the judgment of the reasonable man very often is a mere projected shadow, cast by the judge's own moral views or those of his own social class.'

88 Sigrid Redse Johansen, *The Military Commander's Necessity: The Law of Armed Conflict and Its Limits* (Cambridge University Press 2019) 77.

89 *Ntaganda Trial Judgment* (n 36) [916]; *Katanga Trial Judgment* (n 36) [797].

90 The difference between recklessness and *dolus eventualis* may be summarised as the difference between being culpably indifferent to risks and culpably accepting risks: Blomsma (n 51) 134.

91 The same concerns have been raised in relation to the jurisprudence of the ICTY. See, eg, Jens David Ohlin, 'Targeting and the Concept of Intent' (2013) 35 Michigan Journal of International Law 79. These concerns may implicate a broader tussle between IHL and war crimes law: whether war crimes are simply means for enforcing IHL through criminal sanction, or whether war crimes independently secure the same values as the corresponding IHL rules.

instead, draws attention to its longstanding vintage in international war crimes prosecutions.

However, another answer to this concern might question its premises. A concern as to the conflation of intent and recklessness or *dolus eventualis* seems to assume a strict distinction between them, based on stable contours of the concept of intent and a bright line difference between intent and recklessness or *dolus eventualis*. This is a questionable assumption.

Intention and recklessness or *dolus eventualis* are inherently indeterminate concepts and the boundary between them is semantic and constructed rather than natural and immutable.[92] Consider the soldier in the responsibility gap. If they are certain that their deployment of or reliance on ACC will result in an attack on a civilian target and they intend this attack, they have intent. If they are not certain of this consequence, they lack intent. But what of the soldier who is 99% certain, or 95% or 90%? It is definitely possible to deny the (conceptual) intention of this latter soldier to attack a civilian target, but only by invoking a rigidly doctrinaire conception of intent which would sit uncomfortably with the social and political objectives of criminal responsibility.[93] Acknowledging the questionable distinction between 100% and 95% in this case forces the recognition that the line between intent and recklessness or *dolus eventualis* is necessarily fluid and contingent.[94]

Stated differently, if *mens rea* is inferred from conduct and circumstances then there cannot be a clear and definite boundary between intent and recklessness or *dolus eventualis*. It ceases to matter whether the defendant was certain or only 90% confident that stabbing the victim in the

92 Hart (n 3) 117.

93 An interesting example here is the position of certain forms of wilful blindness in English law. In cases where the defendant intentionally chooses not to inquire into the truth of something because they have no doubt as to the answer, or because they don't want to know the answer, English law assumes knowledge on the part of the defendant, even while recognising the conceptual impossibility of knowledge: Simester and others (n 56) 157–9. See also, more generally, Weigend (n 56) 497–8. Incidentally, it is worth noting that the possibility of wilful blindness has featured extensively in concerns regarding the exclusion of recklessness and *dolus eventualis* from Rome Statute art 30: Eser (n 34) 931–2. It has been argued that the exclusion of wilful blindness cannot have been in the contemplation of the drafters of the Rome Statute: Knut Dörmann, Louise Doswald-Beck and Robert Kolb, *Elements of War Crimes under the Rome Statute of the International Criminal Court: Sources and Commentary* (Cambridge University Press 2003) 131–2, 137–40, 145–7.

94 The contingency of this line represents a socially rooted classification of degrees of culpability. This gives rise to the practical possibility that the line varies depending on the nature and circumstances of the crime, as is the case, discussed above (n 26), in relation to the standard of proof beyond reasonable doubt. Indeed, given the role of judicial interpretation in defining these concepts, it may be possible to point to an iterative process of judicial definition, social response and judicial redefinition. An example of this process relevant to the present context might be the revision of the *Gotovina* trial judgment by the ICTY appeals chamber following stakeholder responses. See, eg, Gary D Solis, 'The Gotovina Acquittal: A Sound Appellate Course Correction' (2013) 215 Military Law Review 78.

stomach would prove fatal; or whether the soldier was certain or only 95% confident that their deployment of or reliance on ACC would result in an attack on civilian targets. In both cases, the assessment of *mens rea* will focus on what the defendant must have known rather than what they did know.[95]

The argument that has been presented here may be summarised as follows.

The responsibility gap is concerned with the impossibility of intent to commit a war crime in the soldier who has reason short of certainty to believe that their deployment of or reliance on ACC might lead to an attack against civilian targets. This statement of the responsibility gap ignores the practicalities of proving *mens rea* in war crimes prosecutions (and criminal prosecutions more generally), where intent can be and is inferred from conduct and circumstances subject to elimination of plausible alternative explanations. This probative practice means that successful prosecution will not require establishment of what the soldier knew or intended, which may well fall short of intent to commit the war crime. Instead, it is sufficient to establish that the soldier must have known that civilian targets would be attacked, and that therefore, the soldier must have intended that attack. Consequently, the impossibility of the careless or uncertain soldier knowing that civilian targets would be attacked and of intending the attack is not an absolute bar to the criminal responsibility of that soldier. Based on available information as to the context and manner of deployment of or reliance on the ACC, constructive knowledge and consequently intent can be imputed to the soldier if it seems impossible that the soldier did not recognise the virtual certainty of attacking civilian targets. In effect, the probative practice of inferring intent from conduct and circumstances allows for the assimilation of egregious cases of recklessness or *dolus eventualis* into the category of intent, notwithstanding the conceptual impossibility of intent.

95 A further extension of this answer to the concern of conflating intention and recklessness might recognise that mental states, like emotions, are not (only) psychological states but socio-cultural practices. See, Sara Ahmed, *The Cultural Politics of Emotion* (2nd edn, Edinburgh University Press 2014) 8–9. This is not to say that a defendant is not intending 'something'. However, the meaning of 'intention' and the classification of the mental state of the defendant are contingent socio-cultural — and in relation to criminal responsibility, political — practices which themselves play a constitutive role in defining the defendant's mental state.

C THE RESTRICTED FOCUS OF THE ICC ON THE MOST SERIOUS INTERNATIONAL CRIMES

This leads neatly to the second feature of international war crimes prosecution which mitigates the responsibility gap: the restricted focus of the ICC on those most responsible for the most serious crimes. This means that the defendants who are likely to attract the attention of the ICC are precisely those whose deployment of or reliance on ACC was so egregious that they must have known of the virtual certainty of attacking civilian targets. And consequently, through the argument set out above, they may be presumed, subject to rebuttal, to have the requisite *mens rea.*

Rome Statute article 5 provides that 'The jurisdiction of the Court shall be limited to the most serious crimes of concern to the international community as a whole.' It goes on to indicate that war crimes generally are an example of such crimes. Article 8(1) then provides that 'The Court shall have jurisdiction in respect of war crimes in particular when committed as part of a plan or policy or as part of a large-scale commission of such crimes.' Neither of these provisions amounts to a concrete restriction of the ICC's war crimes jurisdiction by reference to criteria of seriousness, plans or policies, or scale, but they suggest the prioritisation of war crimes which bear these features, and the de-prioritisation of isolated instances.[96]

The prioritisation suggested by Rome Statute articles 5 and 8(1) is mandated by article 17(1)(d) which posits the case not being 'of sufficient gravity to justify further action by the Court' as a ground of inadmissibility.[97] The ICC has described the gravity requirement as a mandatory rarefaction of the already restricted (on the basis of seriousness in article 5) material jurisdiction of the court.[98] In its article 15 decision on the *Kenya* situation, the ICC described the gravity requirement in terms of restricting focus to those who bear the greatest responsibility for the gravest crimes,[99] and listed the following factors as 'useful guidance' for assessing gravity:[100]

96 Triffterer and Ambos (n 11) 321–2; Michael Bothe, 'War Crimes' in Antonio Cassese, Paola Gaeta and John RWD Jones (eds), *The Rome Statute of the International Criminal Court: A Commentary* (Oxford University Press 2002) 380–1.

97 Triffterer and Ambos (n 11) 811–16.

98 *Situation in the Republic of Kenya* (Article 15 Decision) ICC-01/09-19-Corr (31 March 2010) [56]–[57]; *Situation in the Democratic Republic of Congo* (Arrest Warrant Decision) ICC-01/04-520-Anx2 (10 February 2006) [44], [46].

99 *Kenya Article 15 Decision* (n 98) [59]. See also *Situation in the Republic of Côte d'Ivoire* (Article 15 Decision) ICC-02/11-14-Corr (15 November 2011) [204]; *Prosecutor v Abu Garda* (Confirmation of Charges Decision) ICC-02/05-02/09-243-Red (8 February 2010) [30]–[32].

100 *Kenya Article 15 Decision* (n 98) [62]. See also *Situation in Georgia* (Article 15 Decision) ICC-01/15-12 (27 January 2016) [51]–[57]; *Côte d'Ivoire Article 15 Decision* (n 99) [204].

> (i) the scale of the alleged crimes (including assessment of geographical and temporal intensity); (ii) the nature of the unlawful behaviour or of the crimes allegedly committed; (iii) the employed means for the execution of the crimes (ie, the manner of their commission); and (iv) the impact of the crimes and the harm caused to victims and their families ….

This approach to the gravity of crimes has been adopted in the policies of the Office of the Prosecutor,[101] as well as in its practice.[102] The interpretation of these criteria has been the subject of disagreement between the Prosecutor and the Court,[103] but the charging practice of the Prosecutor reflects a continued adherence to a strict interpretation of these criteria by reference, *inter alia*, to requirements of scale and systemic nature.[104] Moreover, most cases of conduct of hostilities war crimes which have been tried at the ICC (and the ICTY) featured large-scale and systematic violations of IHL.[105]

In other words, war crimes prosecutions relating to ACC at the ICC will likely involve large-scale and systematic violations. These are

101 Office of the Prosecutor of the International Criminal Court, 'Policy Paper on Case Selection and Prioritisation' (15 September 2016) 35–41 <https://www.icc-cpi.int/itemsDocuments/20160915_OTP-Policy_Case-Selection_Eng.pdf> accessed 10 March 2021.

102 Office of the Prosecutor of the International Criminal Court, 'Situation on Registered Vessels of Comoros, Greece and Cambodia: Article 53(1) Report' (6 November 2014) [133]–[148] <https://www.icc-cpi.int/iccdocs/otp/otp-com-article_53(1)-report-06nov2014eng.pdf> accessed 10 March 2021.

103 Compare ibid; *Situation on the Registered Vessels of Comoros, Greece and Cambodia* (Decision on the request of the Union of the Comoros to Review the Prosecutor's Decision not to Initiate an Investigation) ICC-01/13-34 (16 July 2005) [20]–[50]. See also Triffterer and Ambos (n 11) 816.

104 Pre-Trial Chamber I's request to the Prosecutor to reconsider the decision not to initiate an investigation into the situation referred by Comoros, etc. was challenged and finally rejected by the Prosecutor. See, *Registered Vessels of Comoros, Greece and Cambodia Review of Prosecutor's Decision not to Initiate an Investigation* (n 103); Office of the Prosecutor of the International Criminal Court, 'Report on Preliminary Examination Activities 2017 — Registered Vessels of Comoros, Greece and Cambodia' (4 December 2017) <https://www.icc-cpi.int/itemsDocuments/2017-PE-rep/2017-otp-rep-PE-Comoros_ENG.pdf> accessed 10 March 2021. In relation to the preliminary examination into the conduct of UK forces in Iraq, the Prosecutor has noted:

> In the present situation, while there is a significant body of allegations, in light of the circumstances in which some of such allegations were collected, it remains unclear whether the crimes alleged were committed on the scale alleged by communication senders. Additionally, while several failings in army leadership, planning, and training, leading to prisoners' abuses were reported especially in the early phases of Op. Telic, the Office is seeking to assess the gravity of the role of other military or civilian personnel who may bear responsibility as an accessory or as a commander/superior.

See Office of the Prosecutor of the International Criminal Court, 'Report on Preliminary Examination Activities 2018' (5 December 2018) [208] <https://www.icc-cpi.int/itemsDocuments/181205-rep-otp-PE-ENG.pdf> accessed 10 March 2021.

105 A notable exception may be the *Abu Garda* case in which charges were based on a single attack against UN peacekeeping personnel resulting in 12 deaths (and eight further attempted killings) and damage to and appropriation of UN property: *Abu Garda Confirmation Decision* (n 99) [21]–[24]. The Prosecutor has justified the gravity of the impugned conduct in this case by reference to the interests implicated — the security of peacekeeping personnel in the context of the role they play in maintaining the collective security order: *Situation on the Registered Vessels of Comoros, Greece and Cambodia Article 53(1) Report* (n 102) [145].

precisely the sort of violations where the probative practice identified in the previous sub-section could be most significant in mitigating the responsibility gap.

IV
CONCLUSION: BELLIGERENTS' RESPONSIBILITY

This chapter has examined two challenges to the prosecution of ACC-related IHL breaches as war crimes under the Rome Statute of the ICC. First, there is the difficulty of identifying the perpetrator of the conduct corresponding to the *actus reus* of the war crime, in accordance with the criminal standard of proof beyond reasonable doubt. Second, there is the difficulty of accommodating the actual mental state of the concerned human — inevitably negligence, recklessness or *dolus eventualis*, within the stringent *mens rea* requirement of intent and knowledge under the Rome Statute.

It has been argued here that the practical realities of charging and adjudicating war crimes may, in some cases, mitigate these challenges.

The challenge of identifying perpetrators relies on a fixed and rigid understanding of the reasonable doubt standard. In principle, the standard does not require absolute certainty but merely the elimination of reasonable alternatives. In practice, it is a variable and context-specific standard which, particularly in the context of international crimes, may not pose a very exacting threshold. Shorn of its mythologies of certainty, the reasonable doubt standard may not prove to be an insurmountable hurdle to identification and prosecution.

Similarly, the challenge of the responsibility gap relies on the conceptual impossibility of the actual mental state of the alleged perpetrator corresponding to the Rome Statute requirement of intent and knowledge. This framing of the problem ignores the universal probative practice of inferring intent from conduct and circumstances ie, shifting the frame of analysis from what the perpetrator actually knew and intended to what they *must have known* and *therefore intended*. This effectively means that in cases where the manner and mode of deployment of or reliance on ACC suggest that it was impossible that a reasonable commander in the position of the perpetrator would not have recognised the virtual certainty

of an attack upon a civilian target, intent to attack civilian targets may be presumed from this knowledge. These particularly egregious cases are precisely the putative ACC-related war crimes most likely to satisfy the gravity requirement and attract the attention of the ICC. In other words, the practical realities of proving *mens rea* in war crimes prosecutions mitigate some of the challenges of the responsibility gap, at least in the most egregious cases.

These mitigating effects of the practice of charging and adjudicating war crimes are subject to two important clarifications.

First, it is necessary to emphasise that this chapter has not suggested the dilution of the criminal standard of proof or the conflation of intent and recklessness or *dolus eventualis*. Instead, it has drawn attention to the inherent indeterminacy of the reasonable doubt standard and the possibilities provided thereby for the variation or dilution of the standard in practice. And it has noted the ubiquity of inferring subjective mental states from objective physical indicators and has argued that this necessarily entails the replacement of actual knowledge with constructive knowledge, by reference to a benchmark of reasonableness. Neither of these well-recognised features of war crimes prosecutions (or prosecution more generally) is endorsed here, and nor can it be denied that they raise significant concerns for a body of law that is already vulnerable to withering critique on grounds of fairness and legitimacy.[106] That said, insofar as these practices exist,[107] they do provide some amelioration for the difficulties of prosecuting ACC-related IHL breaches as war crimes.

Second, the practical realities of proof beyond reasonable doubt and of establishing *mens rea* do not eliminate the difficulty of identifying perpetrators or resolve the responsibility gap. They operate in some cases to ameliorate these challenges and facilitate prosecution, for instance, in cases where the perpetrator of an ACC-related IHL breach can be

106 See, eg, Frédéric Mégret, 'International Criminal Justice: A Critical Research Agenda' in Christine EJ Schwöbel (ed), *Critical Approaches to International Criminal Law: An Introduction* (Routledge 2014). Indeed, it may be possible to recast the problem of criminal responsibility for ACC-related war crimes as one of prosecution rather than conviction. The history of international criminal trials (and their discourse — above (n 30–32) and accompanying text) suggests that international criminal courts and tribunals usually find a way around legal barriers to the conviction of those prosecuted before them. A more significant challenge to criminal responsibility may lie in the difficulty of prosecuting members of armed forces and citizens of technologically advanced states which are leading the race to develop these technologies. If that barrier is overcome, the technical problems posed by the standard of proof the responsibility gap may prove (relatively) easier to resolve.

107 It is worth highlighting the possibility that the assessment of criminal practice presented here is simply 'moronic', in the sense that Bilbo describes in *Foucault's Pendulum* by Umberto Eco. 'Morons never do the wrong thing. They get their reasoning wrong. Like the fellow who says all dogs are pets and all dogs bark, and cats are pets, too, and therefore cats bark … Morons will occasionally say something that's right, but they say it for the wrong reason …' See Umberto Eco, *Foucault's Pendulum* (Vintage Books 2001) 65.

identified with some degree of certainty, or in cases of egregious recklessness of *dolus eventualis*. This is not an insignificant argument, because it opens the door to destabilising the unitary and fixed nature of perpetrator identification and the responsibility gap and for seeing them as difficulties which may manifest differently in different cases. However, it undeniably leaves many residual cases where these factors operate to hinder the prosecution of ACC-related IHL breaches as war crimes.

In concluding this chapter, it is useful to consider these residual cases briefly.

To begin with, it must be emphasised that while the challenges of individual criminal responsibility in these cases suggest a gap in the criminal enforcement of IHL, they do not imply a gap in the enforcement of IHL.

War crimes are serious breaches of IHL which implicate individual criminal responsibility.[108] But war crimes and the criminal responsibility they entail are only one part of the enforcement infrastructure of IHL. The broader and indeed primary part of IHL's enforcement infrastructure draws on the responsibility of belligerents for breaches of the rules of IHL. In recent years, the difficulties of enforcing the rules of IHL against States and non-State actors alike, the comparative successes of international criminal courts and tribunals, and the lure of ending impunity have combined to privilege individual criminal responsibility over belligerents' responsibility under IHL.[109] But even if subordinated, belligerents' responsibility persists in relation to IHL rules, and applies equally to ACC-related IHL breaches.[110]

Indeed, it may be easier to invoke the responsibility of belligerents for breaches of IHL than to prosecute those breaches as war crimes. Recklessness and negligence suffice for triggering responsibility under IHL

108 *Prosecutor v Tadić* (Decision on the Defence Motion for Interlocutory Appeal on Jurisdiction) [1995] ICTY-94-1-A (25 October 1995) [94]; International Committee of the Red Cross, 'Explanatory Note: What Are "Serious Violations of International Humanitarian Law"?' (2012) <https://www.icrc.org/en/doc/assets/files/2012/att-what-are-serious-violations-of-ihl-icrc.pdf> accessed 10 March 2021.

109 This privileging of individual criminal responsibility has raised systemic challenges for IHL, including the de-prioritisation of those IHL norms which are not capable of individualisation and criminalisation, and the frequent misinterpretation of IHL norms. This argument is developed in greater detail in Paola Gaeta and Abhimanyu George Jain, 'Individualisation of IHL Rules Through Criminal Responsibility for War Crimes and Some (Un)Intended Consequences' in Dapo Akande and Jennifer Welsh (eds), *The Individualisation of War* (Oxford University Press 2021). See also Paola Gaeta, 'Autonomous Weapon Systems and the Alleged Responsibility Gap' in *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons — Expert Meeting, Versoix, Switzerland, 15–16 March 2016 (International Committee of the Red Cross 2016)* < https://icrcndresourcecentre.org/wp-content/uploads/2017/11/4283_002_Autonomus-Weapon-Systems_WEB.pdf> accessed 10 March 2021.

110 United States of America, 'Working Paper on Autonomy in Weapon Systems' (Group of Governmental Experts on Lethal Autonomous Weapons Systems, 10 November 2017) UN Doc CCW/GGE.1/2017/WP.6, [24].

provided they result in unreasonable errors in attack,[111] resolving the issue of the responsibility gap. As to the difficulty of identifying perpetrators, the transition from individual criminal responsibility to belligerents' responsibility entails less stringent burdens and standards of proof,[112] as well as a shift from identifying specific perpetrators to identifying the responsible belligerent and attributing the acts to it.

In relation to the difficulty of criminal responsibility for ACC-related IHL breaches, turning to the responsibility of belligerents is not only possible and easier, it may be more appropriate.

Consider the residual responsibility gap.[113] It includes soldiers whose mental state in deploying or relying on ACC falls below the principled intent and knowledge requirements of the ICC, and also falls below the practical 'must have known and therefore intended' threshold. In other words, these soldiers fall well below the *mens rea* requirement set out in the Rome Statute for the war crimes of attacking civilian targets.

A particular *mens rea* requirement represents a socio-political determination that a particular degree of culpability is required for criminal responsibility.[114] Breach of the rule is necessary but not sufficient to trigger criminal sanction: the breach must be accompanied by a specified culpable state of mind.[115] In other words, the impossibility of accommodating the soldier in the residual responsibility gap within the principled or practical *mens rea* requirements of the Rome Statute suggests the inadequate culpability of the soldier. The soldier may not have been culpable at all — they may have been justifiably unaware of a risk of an

111 IHL conduct of hostilities rules such as the rule of distinction do not guarantee the protection of civilians, they guarantee that civilians will not be made the object of attack and tolerate the possibility of an erroneous attack on civilians. This begs the question of which errors are permissible and the answer to that question relies on the standard of the reasonable commander. Consequently, deliberate attacks on civilians constitute a breach of the rule of distinction, as do negligent and reckless attacks (assessed in accordance with the standard of the reasonable commander). This approach to the requirements of the rule of distinction is reflected in, eg, US Department of Defense, *Law of War Manual* (December 2016) [5.3]; Human Rights Council, 'Situation of Human Rights in Yemen, Including Violations and Abuses Since September 2014' (17 August 2018) UN Doc A/HRC/39/43, 7; *Partial Award: Central Front - Ethiopia's Claim 2* (2004) 26 RIAA 155 (Eritrea-Ethiopia Claims Commission) [101]–[113]; 'Final Report to the Prosecutor of the ICTY by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia' (2000) [80]–[85] <https://www.icty.org/en/press/final-report-prosecutor-committee-established-review-nato-bombing-campaign-against-federal> accessed 10 March 2021.

112 See, eg, Banks (n 5) 247–8; Dederer and Singer (n 5) 439–45.

113 The residual cases also include those where the perpetrator of the attack cannot be identified in accordance with even the variable and operational standard of proof beyond reasonable doubt. This is a 'practical' problem rather than the 'conceptual' problem posed by the responsibility gap and so it is not addressed separately here. However, the argument as to the inappropriateness of criminal responsibility for residual cases applies equally in this context: continuing and significant doubts as to the identity of the perpetrator should not be seen as an impediment to criminal responsibility but instead as an indication of the impropriety of criminal responsibility.

114 It is also possible that particular conduct may be criminalised even in the absence of a culpable state of mind, as is the case with so-called 'strict liability' crimes.

115 See, eg, Hart (n 3) 160.

attack on a civilian target. Or the soldier may not have been sufficiently culpable — they may have been unjustifiably unaware of or accepted a risk of attacking civilian targets, but in either case it cannot be said that they must have known of the virtual certainty of attacking civilian targets.[116]

By conceptualising the mental state of the solider in the residual responsibility gap in terms of absent or insufficient culpability it becomes possible to reconceptualise the residual responsibility gap. The difficulty of satisfying the *mens rea* requirement of the Rome Statute does not only indicate the impossibility of individual criminal responsibility, it also indicates its inappropriateness. This reconceptualization of the residual responsibility gap recognises that IHL breaches are less frequently the result of individual deviance and more often arise from institutional factors including systemic interpretation and implementation of IHL.[117]

Reconceptualising the residual responsibility gap in terms of the inappropriateness of individual criminal responsibility rather than its impossibility does not deny that there has been an attack against a civilian target which may constitute a breach of IHL. It does not deny the importance of criminal responsibility in the enforcement of IHL.[118] And finally, it does not deny the broader significance and uses of criminal responsibility, from the perspectives of victims, offenders and society more broadly. But it does wonder whether in cases where attacks against civilian targets result from inadequate training, flaws in the development or use of ACC, or systemic misinterpretation or disregarding of IHL, criminal responsibility is the appropriate means of IHL enforcement. In these cases, it seems evident that the responsibility of belligerents should be the focus of enforcement efforts, and individual criminal responsibility is not only inapplicable but also inappropriate.[119]

116 It is worth emphasising that the insufficiency of culpability in these cases is only by reference to the specific and contingent standard of the Rome Statute. The inadequate culpability of negligence, recklessness or *dolus eventualis* in the context of lethal force and civilian lives is a socio-political choice and not an immutable fact.

117 See, eg, Matthew Talbert and Jessica Wolfendale, *War Crimes: Causes, Excuses, and Blame* (Oxford University Press 2018). This also reflects the broader idea that international crimes necessitate difficult distinctions between individual conduct and systemic criminality: eg, Mégret (n 106) 28–30; Neha Jain, 'Individual Responsibility for Mass Atrocity: In Search of a Concept of Perpetration' (2013) 61 American Journal of Comparative Law 831, 831–2.

118 Although, as argued above, there are concerns as to the primacy of individual criminal responsibility in the enforcement of IHL and as to the exclusion of belligerents' responsibility.

119 An interesting example here is the mistaken American strike on the Chinese embassy in Belgrade during the NATO intervention in Kosovo. In its final report to the prosecutor, the committee established to review the bombing campaign noted that the strike was erroneous and that the error resulted from misidentification of the Chinese embassy and inadequacies in the targeting process which prevented discovery of the error. But though the report deemed the strike to constitute a breach of the rule of distinction, it did not consider it appropriate to invoke the criminal responsibility of the pilots and senior military leaders on account of the systemic source of the error. See ICTY NATO Report (n 111) [80]–[85].

Clearly, the appropriateness of criminal responsibility for cases in the residual responsibility gap cannot be determined in the abstract and must be considered on a case-by-case basis. But it would seem reasonable to assume that cases in the residual responsibility gap would largely arise from systemic factors rather than individual deviance, particularly given the exclusion of cases meeting the 'must have known and therefore intended' threshold.[120]

The possibility of systemic factors in IHL breaches relating to ACC (and lethal autonomy) is particularly significant given the radical changes in the very nature of armed conflict which is facilitated by these technologies. The tactical, operational and strategic possibilities created by ACC may prove difficult to accommodate within the existing IHL framework and technological change may spur legal change. Military applications of ACC may challenge the binary of IHL compliance and breach, imperilling the prospect of prosecuting IHL breaches as war crimes. Put another way, cases in the residual responsibility gap may reflect disagreement as to the military use and manner of use of ACC. Those disagreements are entirely legitimate and indeed, necessary, but their resolution through the individual criminal responsibility of the implicated soldier is inappropriate.

The early years of the drone debates provide a fitting analogue here. In that context there were similar concerns about IHL breaches resulting from drone strikes and about the possibility of establishing criminal responsibility.[121] These concerns stemmed from disagreement as to how to assimilate the new military possibilities enabled by remote warfare within the requirements of IHL. Drone strikes have not resulted in significant war crimes prosecutions or convictions. Concerns about the manner in which drone strikes are conceptualised and conducted have been discussed largely within the framework of IHL, producing a slow and incomplete but discernible process of reconciliation between the novel practice of drone strikes and IHL.

The concerns raised by drones, like the concerns raised by the residual category of ACC-related IHL breaches, implicate the interpretation and application of substantive IHL norms at the level of belligerents rather than the actions of individual soldiers. Their resolution through the prism of the criminal enforcement mechanism of IHL would be inappropriate and unfair.

120 Of course, there remains the possibility of a core residual responsibility gap comprising of cases featuring un-prosecutable individual deviance.

121 See, eg, Kevin Jon Heller, '"One Hell of a Killing Machine": Signature Strikes and International Law' (2013) 11 Journal of International Criminal Justice 89.

Chapter 13

Autonomous Cyber Weapons and Command Responsibility

Russell Buchan and Nicholas Tsagourias

I
INTRODUCTION

Parties to armed conflicts frequently deploy cyber weapons and, recognising the competitive advantages afforded by autonomy, States are developing — or perhaps have already developed — autonomous cyber weapons ('ACWs') for use in armed conflict.[1] In this context, autonomy does not mean independence from humans because ACWs are programmed, deployed and can be supervised by humans whereas their decision-making capacity can be moulded by humans.[2]

1 United Nations Institute for Disarmament Research, 'The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations' (16 November 2017) <https://www.unidir.org/publication/weaponization-increasingly-autonomous-technologies-autonomous-weapon-systems-and-cyber>; UK Ministry of Defence, 'Armed Forces Announce Launch of Cyber Regiment in Major Modernisation' (4 June 2020) <https://www.gov.uk/government/news/armed-forces-announce-launch-of-first-cyber-regiment-in-major-modernisation>; Dan Sabbagh, 'Britain has Offensive Cyberwar Capability, Top General Admits' (*The Guardian*, 15 September 2020) <https://www.theguardian.com/technology/2020/sep/25/britain-has-offensive-cyberwar-capability-top-general-admits?CMP=Share_iOSApp_Other>.

2 'It should be made clear that all autonomous systems are supervised by human operators at

As a matter of fact, autonomy exists on a continuum[3] and the degree of autonomy enjoyed by a weapon is determined by its technical specification, the functions that have been automated and its interaction with a human agent.[4] For example, cyber weapons have limited autonomy when they are controlled by pre-established algorithms and conduct targeting operations according to pre-determined scenarios or when they are authorised or supervised by humans. At the other end of the continuum, cyber weapons are highly autonomous where they utilise adaptive intelligence and self-learning capabilities in their efforts to identify and engage targets in complex and dynamic environments but also when there is no communication with a human agent once launched. Highly ACWs are thus self-governing. Although they operate within a framework of planned behaviour, they are able to make independent decisions in response to external variables and by interacting with the external environment but these decisions are made on the basis of internal programming, information, processes, conditions and constraints.

Stuxnet is a good example of a highly ACW, even if it was not deployed during an armed conflict. Stuxnet was a computer worm that was surreptitiously downloaded (probably through a compromised USB stick) onto the Intranet at the Natanz nuclear facility in Iran and was designed to frustrate Iran's efforts to enrich uranium and develop nuclear energy. Operating within a complex web of interconnected networks, Stuxnet was able to identify specific models of programmable logic controllers ('PLCs')

some level, and autonomous systems' software embodies the designed limits on the actions and decisions delegated to the computer'; US Department of Defense, 'Task Force Report: The Role of Autonomy in DoD Systems' (July 2012) 1–2 <https://fas.org/irp/agency/dod/dsb/autonomy.pdf>.

3 US Department of Defense, 'Directive 3000.09' (8 May 2017) <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf>. According to the UK Ministry of Defence: 'An autonomous system is capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be'; UK Ministry of Defence, 'Joint Doctrine Publication 0-30.2: Unmanned Aircraft Systems' (August 2017) 13 <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/673940/doctrine_uk_uas_jdp_0_30_2.pdf>. See also UK Ministry of Defence, Joint Concept Note 2/17: Future of Command and Control (September 2017) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/643245/concepts_uk_future_c2_jcn_2_17.pdf>.

4 Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>; Alan L Schuller, 'At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law' (2017) 8 Harvard National Security Journal 379; Paul Scharre and Michael C Horrowitz, 'An Introduction to Autonomy in Weapon Systems: Working Paper' (Centre for a New American Security February 2015) <https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/Ethical-Autonomy-Working-Paper_021015_v02.pdf?mtime=20160906082257&focal=none>. The human-machine interaction in the decision-making loop is schematically described as 'human in the loop', 'human on the loop' and 'human out of the loop'. Current autonomous weapons have varying degrees of autonomy.

manufactured by Siemens. These PLCs allowed the facility's computers to control the centrifuges that were being used to enrich uranium. Stuxnet altered the PLCs' programming and this caused the centrifuges to spin too quickly and for too long. These changes prevented the enrichment of uranium and caused the physical destruction of a large number of centrifuges. To remain undetected, Stuxnet recorded sensor values during the period in which the PLCs were operating normally. Once enough data had been collected, Stuxnet modified the PLCs' programming while at the same time feeding computer operators fake sensor values, leading them to believe that the PLCs were functioning normally. An interesting feature of the Stuxnet virus was that, once deployed, it could not interact with its operators since, for security purposes, Natanz is an air-gapped facility and thus not connected to the wider Internet.[5]

When ACWs engage in operations that produce violent effects they qualify as 'attacks' and must comply with the rules of international humanitarian law ('IHL').[6] Grave or serious breaches of IHL constitute war crimes.[7] The question that immediately arises and which will be considered in this chapter is whether commanders can be held criminally liable in relation to such crimes.

Commanders can be held criminally liable as perpetrators if they use an ACW to commit the *actus reus* of a crime with intent or knowledge.[8] For example, if a commander individually or jointly with others launches an ACW in order to kill protected civilians or is aware that such killings will occur in the ordinary course of events, they will be held responsible as perpetrators or co-perpetrators of the war crime of intentionally directing attacks against a civilian population.[9] Commanders can also be held criminally liable as perpetrators if they commit a war crime through another person; when, for example, they control the will of a person who goes on to commit a war crime by using ACWs.[10] Furthermore, commanders can be held criminally liable as accomplices if they intentionally or with

5 For an overview of Stuxnet see Marco de Falco, 'Stuxnet Facts Report: A Technical and Strategic Analysis' (CCDCOE, 2012), <https://ccdcoe.org/library/publications/stuxnet-facts-report-a-technical-and-strategic-analysis-2/>.

6 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (entered into force 7 December 1978) 1125 UNTS 3 ('AP I') art 49(1).

7 Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3 art 8 ('ICC Statute').

8 ibid art 25(3)(a) ('commits such a crime, whether as an individual, jointly with another'). See also ibid art 30.

9 ibid art 8(2)(b)(i).

10 ibid art 25(3) (a) ('Commits such a crime … through another person, regardless of whether that other person is criminally responsible', which covers the case of perpetration through another person).

knowledge assist in the commission of a war crime by another person through the use of ACWs.[11]

All this may be possible when ACWs are used in clearly defined and well-structured operational environments against pre-planned and stable targets. However, when an ACW is deployed into a complex and evolving operational environment and has the capacity to make dynamic targeting decisions, it cannot be said that the commission of a war crime was intended by the commander or that s/he had knowledge that it would occur or that s/he had assisted in its commission.[12]

In view of the above, in this chapter we consider an alternative form of liability attached by international criminal law to military commanders, namely, command responsibility.[13] Command responsibility is an indirect mode of liability which holds commanders criminally liable for their failure to prevent or repress crimes committed by their subordinates.[14]

In order to apply the law of command responsibility to ACWs, it is important to explain how we envisage the relationship between a commander and an ACW. This is because command responsibility is grounded on a human-to-human relationship, that is, the relationship between commanders and their subordinates i.e. soldiers. This construction of command responsibility can certainly apply to the case at hand when, for example, soldiers commit war crimes by using ACWs and the commander fails to prevent their commission or fails to repress them.

However, our focus in this chapter is different: it is on the relationship between a commander and an ACW. We contend that the relationship between a commander and an ACW resembles and replicates the relationship between commanders and their soldiers. This is because ACWs are agents operating within an organised system of command and control even if they act on their own when executing an order or have self-learning and adaptive capabilities. This system of command and control comprises resources (human and material), structures, facilities, processes (for example, training) and legal and military authority

11 ibid art 25 (3)(b)-(d).

12 This is particularly so since *dolus eventualis* (recklessness) has not been included within the ICC's *mens rea* requirements; *Prosecutor v Thomas Lubanga Dyilo (Judgment)* ICC-01/04-01/06 (14 March 2012) [1011].

13 ICC Statute (n 7) art 28.

14 These crimes include genocide, crimes against humanity, war crimes and the crime of aggression. It should be noted that command responsibility can run in parallel with direct forms of responsibility. According to existing jurisprudence, if both direct responsibility and command responsibility are established in relation to the same conduct, the former takes precedence whereas the position of the accused as commander is considered to be an aggravating circumstance when sentencing: *Prosecutor v Blaskić (Appeals Chamber Judgement)* ICTY-95-14-A (29 July 2004) [91]-[92]; *Prosecutor v Kajelijeli (Appeals Chamber Judgement)* ICTR-98-44A-A (23 May 2005) [81]; *Judgment (Kaing Guek Eav alias Duch)* Case File 001/18-07-2007/ECCC/TC, E188 (16 July 2010) [539].

according to which commanders can exercise, maintain and actualise their command.[15] It allows commanders to manage resources; collect and assess information; plan operations; make decisions; direct and execute operations; supervise, monitor and assess operations, behaviours and actions; and take corrective action. It also ensures that operations are accomplished according to their objectives and within the law in circumstances of operational uncertainty, imperfect information and time-constraints. Operating within such a command and control system also means that the decision-making capacity and independence of subordinates is bounded and conditioned, albeit in different respects and to different degrees. It is because of the existence of such a command and control system that ACWs can be considered subordinates in the same way as soldiers who equally operate within an organised system of command and control are considered subordinates even if they are able to make independent decisions.[16]

It is also important to note that in both cases (soldiers and ACWs) commanders exercise macro-level command and control to ensure that their subordinates operate within the law and within their command as well as micro-level command and control to effectuate their command in particular circumstances or in relation to particular subordinates.[17] The two levels are interconnected, interdependent and integrated within the concept of command and control which must be viewed holistically. To explain, while micro-command requires the application of the command and control tools to specific instances and persons, it can be exercised only if the framework of macro-level command and control is in place and functioning effectively. That said, the difference between exercising command and control over soldiers and over ACWs lies in the fact that in the latter case command and control is exercised, maintained and actualised through technical means whereas in the case of soldiers it is through interpersonal, physical, legal or institutional means.

Having explained the relationship between commanders and ACWs, we will now apply the law of command responsibility thereto. The chapter

15 See Frank M Snyder, *Command and Control: Readings and Commentary* (National Defense University 1993); Loren D Diedrichsen, Command and Control: Operational Requirements and System Implementation (2000) 5 Information and Security 23; US Marine Corps, *The Nature of Command and Control* (4 October 1996) 33–60, <https://www.marines.mil/Portals/1/Publications/MCDP%206%20Command%20and%20Control.pdf>.

16 See Jens David Ohlin, 'The Combatant's Stance: Autonomous Weapons on the Battlefield' (2016) 92 International Law Studies 1 and Gary S Corn, 'Autonomous Weapons Systems: Managing the Inevitability of "Taking the Man Out of the Loop"' in Nehal Bhuta and others (eds), *Autonomous Weapons Systems: Law, Ethics and Policy* (Cambridge University Press 2016) 209.

17 *Prosecutor v Halilović (Judgement)* ICTY-01-48-T (16 November 2005) [79]–[100] (where the Trial Chamber speaks of a general and specific duty of a commander to prevent).

is thus structured as follows. Section II introduces the doctrine of command responsibility and identifies its core elements, namely, the existence of a superior–subordinate relationship; the commission or prospective commission of crimes by subordinates; and the commander's knowledge or constructive knowledge of such crimes. Sections III to VII examine how these elements apply to ACWs including the scope of the element of causality introduced by Article 28 of the ICC Statute as well as the scope of responsibility of successor commanders. Section VIII offers conclusions.

II
THE LAW OF COMMAND RESPONSIBILITY

The principle of command responsibility is well established in military doctrine and international criminal law.[18] It derives from the principle of responsible command[19] according to which the commander as the placeholder for the State becomes the 'guarantor' of IHL and for this reason s/he is entrusted with powers to foster and ensure compliance with IHL. Importantly, a commander's dereliction of this duty attracts sanctions, including criminal ones.[20]

As an international criminal law principle, command responsibility was developed by war crimes tribunals particularly after the Second World War.[21] It was later codified in the Statutes of international criminal tri-

18 *Prosecutor v Delalić et al (Judgement)* ICTY-96-21-T (16 November 1998) [333]; Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law* (International Committee of the Red Cross 2005) rule 153. See generally Mijan Damaska, 'The Shadow Side of Command Responsibility' (2001) 49 American Journal of Comparative Law 455; Kai Ambos, 'Superior Responsibility' in Antonio Cassese, Paola Gaeta and John RWD Jones (eds), *The Rome Statute of the International Criminal Court: A Commentary* (Oxford University Press 2002); Guénaël Mettraux, *The Law of Command Responsibility* (Oxford University Press, 2009); Chantal Meloni, *Command Responsibility in International Criminal Law* (Springer 2010); Nicholas Tsagourias, 'Command Responsibility and the Principle of Individual Criminal Responsibility: a Critical Analysis of International Jurisprudence' in Chile Eboe-Osuji (ed), *Protecting Humanity: Essays in International Law and Policy in Honour of Navanethem Pillay* (Brill 2010) 817–37; Kai Ambos, *Treatise on International Criminal Law, Vol. I: Foundations and General Part* (Oxford University Press 2013) 197–232; Otto Triffterer and Kai Ambos (eds), *The Rome Statute of the International Criminal Court: Commentary* (Oxford University Press 2016) 1056–106.

19 See Hague Convention (II) with Respect to the Laws and Customs of War on Land (adopted 29 July 1899, entered into force 4 September 1900) 189 CTS 429 art 1; AP I art 86 and 87. See also Yves Sandoz, Christophe Swinarski and Bruno Zimmermann (eds), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (ICRC 1987) [3549]–[3550].

20 Halilović (n 17) [39] and [87]; *Prosecutor v Bemba (Judgment Pursuant to Article 74 of the Statute)* ICC-01/05-01/08 (21 March 2016) [172].

21 See, eg, *Trial of General Tomoyuki Yamashita*, Law Reports of Trials of War Criminals, vol IV, case no 21 (4 February 1946); *United States v Wilhelm von Leeb et al*, Law Reports of Trials of War

bunals as in Article 7(3) of the Statute of the International Criminal Tribunal for the former Yugoslavia ('ICTY')[22] and Article 6(3) of the Statute of the International Tribunal for Rwanda ('ICTR'),[23] which contributed to the development of the doctrine. It has also been codified in Article 28 of the Rome Statute Establishing the International Criminal Court ('ICC Statute').[24]

The three main constitutive elements of command responsibility as formulated in Article 28 of the ICC Statute and relevant jurisprudence are:

- The existence of a superior–subordinate relationship;
- The superior knew or should have known that international crimes were about to be committed or had been committed by subordinates; and
- The commander failed to take the necessary and reasonable measures to prevent or repress the commission of the crimes or to submit the matter to the competent authorities for investigation and prosecution.

Article 28 of the ICC Statute also requires a causal nexus between the crimes and the commander's failure to exercise proper command.

How these elements apply to ACWs is explored in the sections that follow but, before we do this, we need to explain the nature of command responsibility[25] because this will have a bearing on the interpretation of its elements.

Article 28 of the ICC Statute casts command responsibility as a mode of liability for the crimes of subordinates when it says that commanders are responsible 'for' the crimes of their subordinates.[26] By contrast, the *ad hoc* tribunals treat command responsibility as responsibility for the

Criminals, vol XII, case no 72 (30 December 1947 — 28 October 1948).

22 Statute of the International Criminal Tribunal for the Former Yugoslavia (ICTY), UNSC Res 827 (25 May 1993).

23 Statute of the International Criminal Tribunal for Rwanda (ICTR), UNSC Res 955 (8 November 1994).

24 In its modern formulation, command responsibility encompasses both military and civilian superiors but this chapter focuses exclusively on military commanders.

25 See Ambos (n 18); Gerhard Werle and Florian Jessberger, *Principles of International Criminal Law* (Oxford University Press 2014) 221–2; Chantal Meloni, 'Command Responsibility: Mode of Liability for the Crimes of Subordinates or Separate Offence of the Superior?' (2007) 5 Journal of International Criminal Justice 619; Darryl Robinson, 'How Command Responsibility Got So Complicated: A Culpability Contradiction, Its Obfuscation, and a Simple Solution' (2012) 13 Melbourne Journal of International Law 1. On how command responsibility applies to cyber war see Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017) rule 85.

26 *Bemba (Judgment Pursuant to Article 74)* (n 20) [171]. Cf Ambos (n 18) 851.

dereliction of an affirmative duty to prevent or repress crimes committed by subordinates. As the Trial Chamber opined in the *Halilović* case:

> The Trial Chamber finds that under Article 7(3) command responsibility is responsibility for an omission. The commander is responsible for the failure to perform an act required by international law. This omission is culpable because international law imposes an affirmative duty on superiors to prevent and punish crimes committed by their subordinates. Thus "for the acts of his subordinates" as generally referred to in the jurisprudence of the Tribunal does not mean that the commander shares the same responsibility as the subordinates who committed the crimes, but rather that because of the crimes committed by his subordinates, the commander should bear responsibility for his failure to act. The imposition of responsibility upon a commander for breach of his duty is to be weighed against the crimes of his subordinates; a commander is responsible not as though he had committed the crime himself, but his responsibility is considered in proportion to the gravity of the offences committed.[27]

In our opinion, this is a better approach because it comports with the principle of culpability in that commanders bear responsibility for their own culpable omissions with regard to their subordinates' crimes rather than being criminally liable for these crimes themselves. Also, to treat command responsibility as a form of participation in the crimes of others undermines Article 25 of the ICC Statute or makes Article 28 irrelevant.[28] This approach also comports with the rationale of command responsibility which, as explained earlier, makes commanders the guarantors of legality during an armed conflict due to their powers and also due to the special relationship that exists between commanders and subordinates.

27 *Halilović* (n 17) [54]; *Prosecutor v Enver Hadžihasanović and Amir Kubura (Judgement)* IT-01-47-T (15 March 2006) [74]–[75]; *Prosecutor v Milorad Krnojelac (Appeals Chamber Judgement)* ICTY-97-25-A (17 September 2003) [171].

28 *Halilović* (n 17) [78]. Although command responsibility is often referred to as *sui generis* mode of liability (see *Bemba (Judgment Pursuant to Article 74)* (n 20) [174]), it is not always clear to what this refers. However, according to van Sliedregt it combines aspects of a mode liability and aspects of a separate offence liability; Elies van Sliedregt, *Individual Criminal Responsibility in International Criminal Law* (Oxford University Press 2012) 196.

III
THE EXISTENCE OF A SUPERIOR–SUBORDINATE RELATIONSHIP

The first constitutive element of command responsibility is that of a superior–subordinate relationship which in effect refers to a command and control relationship.[29] This relationship can be *de jure* or *de facto*. The former refers to the vested authority of a commander over subordinates[30] with the army being the primary institution operating on the basis of formal structures of command and control. Yet, a command and control relationship does not need to be formal and can also arise from factual or other circumstances of subordination mainly due to a person's *de facto* authority and powers of control.[31] In this case, one can speak of a *de facto* commander.[32] That said, the most decisive factor in both instances is the commander's 'effective command and control or authority and control' over subordinates.[33] This suggests a 'real' or 'actual power to control',[34] which in the case of command responsibility refers to 'the

29 'The superior–subordinate relationship lies at the heart of the doctrine of a commander's liability for crimes committed by her subordinates. It is the position of command over and the power to control the acts of the perpetrator which forms the legal basis for the superior's duty to act and for his corollary liability for a failure to do so'; *Prosecutor v Limaj et al (Judgement)* ICTY-03-66-T (30 November 2005) [521].

30 Command has been defined as '[t]he authority that a commander in the armed forces lawfully exercises over subordinates by virtue of rank or assignment" whereas command and control has been defined as '[t]he exercise of authority and direction by a properly designated commander over assigned and attached forces in the accomplishment of the mission'; Office of the Chairman of the Joint Chiefs of Staff, 'DoD Dictionary of Military and Associated Terms' (US Department of Defense, June 2020) 40 <https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/dictionary.pdf>.

31 *Limaj* (n 29) [522].

32 Command responsibility can thus extend to non-military superiors effectively acting as military commanders; *Prosecutor v Delalić et al (Appeals Chamber Judgement)* ICTY-96-21-A (20 February 2001) [195]; *Prosecutor v Bagilishema (Appeals Chamber Judgement)* ICTR-95-1A-A (3 July 2002) [35]; *Bemba (Judgment Pursuant to Article 74)* (n 20) [176]–[177]. This distinguishes them from civilian superiors as defined in Article 28(b) of the ICC Statute.

33 *Bemba (Judgment Pursuant to Article 74)* (n 20) [189]. The requirement of effective command and control is explicitly stated in Article 29 of the Law on the Establishment of the Extraordinary Chambers in the Courts of Cambodia for the Prosecution of Crimes Committed During the Period of Democratic Kampuchea, 2001, as amended by NS/RKM/1004/006 (27 October 2004). Although 'command' and 'authority' are deemed to refer to the same thing, one can say that 'command' relates to *de jure* and 'authority' to *de facto* commanders. See also *Prosecutor v Bemba (Decision Pursuant to Article 61(7)(a) and (b) of the Rome Statute on the Charges of the Prosecutor against Jean-Pierre Bemba Gombo)* ICC-01/05-01/08424 (15 June 2009) [412]–[413].

34 ibid [418]. See also *Prosecutor v Kordić and Čerkez (Judgement)* ICTY-95-14/2-T (26 February 2001) [422]; *Prosecutor v Musić et al (Judgement)* ICTY-96-21-T (16 November 1998) [370]; *Delalić* (n 32) [256]. 'Effective control' for the purposes of command responsibility does not have the same meaning as in the law of State responsibility (see Article 8, International Law Commission, *Articles on State Responsibility for Internationally Wrongful Acts* 2001), which is about the State's effective control over a wrongful conduct. It is also different from the notion of control over the crime or over the will of the perpetrator which is required for co-perpetration or for perpetration through another person according to ICC Statute art 25(3)(a). Instead, effective control in the law of command responsibility is control over subordinates; *Limaj* (n 29) [522].

material ability to prevent or punish criminal conduct'.[35] Put differently, effective control refers to the capacity of commanders to effectuate their command over subordinates in general or in particular circumstances. Whether this capacity exists in the particular circumstances is a matter of evidence and assessed on a case-by-case basis whereas the existence of *de jure* command only creates a rebuttable presumption of effective control.[36] That said, what the law requires is for the commander to have in place a functional command and control system.[37]

In its examination of existing jurisprudence, the Trial Chamber in *Bemba* identified a number of factors that may indicate the existence of effective command and control. These include:

- the official position of commanders within military structures and the actual tasks they carried out;
- the power of the commander to issue orders, including his capacity to order forces or units under his command, whether under his immediate command or at lower levels, to engage in hostilities;
- the capacity to ensure compliance with orders including consideration of whether the orders were actually followed;
- the capacity to re-subordinate units or make changes to command structure;
- the power to promote, replace, remove, or discipline any member of the forces, and to initiate investigations;
- the authority to send forces to locations where hostilities take place and withdraw them at any given moment;
- independent access to, and control over, the means to wage war, such as communication equipment and weapons;
- control over finances;
- the capacity to represent the forces in negotiations or interact with external bodies or individuals on behalf of the group; and

35 ibid; *Bemba (Judgment Pursuant to Article 74)* (n 20) [183].
36 *Delalić* (n 32) [197]; *Hadžihasanović* (n 27) [845]–[846].
37 It is important to stress that *ad hoc* command and control in relation to a particular instance is not equivalent to effective command and control as required by the law.

- whether the individual represents the ideology of the movement to which the subordinates adhere and has a certain level of profile, manifested through public appearances and statements.

By contrast, a lack of effective command and control can be established by:

- the existence of a different exclusive authority over the forces in question;
- disregard or non-compliance with orders or instructions of the accused;
- a weak or malfunctioning chain of command; and
- the existence of intermediaries which prevents a commander from exercising effective command and control over subordinates.[38]

It transpires that most of the aforementioned factors apply to ACWs which, as we said, operate within a system of command and control. A commander can in principle issue orders to an ACW, direct or modify its operations as well as supervise and monitor ACWs to ensure compliance with IHL. A commander can also replace the weapon or withdraw it from the field.[39] Whether these powers amount to effective control depends on the weapon's programing and how it is designed to operate as well as the extent to which its activities can be supervised and overridden by the commander. For example, algorithms may govern the activities of cyber weapons that restrict their attack capability to specific targets within certain networks and for a limited period of time. They can also strictly limit a weapon's area of operation or require it to comply with clearly defined rules of engagement. All these design and programming choices make its operation more deterministic and predictable. Furthermore, ACWs may remain under the close supervision of a commander through a constant and real-time monitoring mechanism, which allows a commander to adjust the algorithm in real-time in order to modify instructions, assign new tasks and correct glitches. A commander may also have the ability to abort operations or deactivate a cyber weapon if it starts behaving unexpectedly or once it has successfully completed

38 *Bemba (Judgment Pursuant to Article 74)* (n 20) [184], [190]; *Nuon Chea, Khieu Samphan (Trial Chamber Judgement)* Case No 002/19-09-2007/ECCC/TC, E313 (7 August 2014) [1016]–[1022].

39 Borrowing from the jurisprudence on 'perpetration through another' due to an organised system of command and obedience, fungibility is evidence of the superior's power of control; *Prosecutor v Katanga and Ngudjolo Chui (Decision on the Confirmation of Charges)* ICC-01/04-01/07-717 (30 September 2008) [518].

its mission. Under such circumstances, it can be said that the ACW acts under the effective command and control of the commander.

ACWs can however be driven by powerful algorithms that enable them to determine which targets to engage; when they should be engaged; how they should be engaged; whether the target has been neutralised or should be re-engaged; and when to proceed to the next task. Highly ACWs take and effectuate these decisions very quickly and relay back to the commander vast quantities of complex data,[40] all of which may effectively deprive a commander of his or her ability to supervise the weapon's activities or intervene when problems occur. Also, and as with the Stuxnet virus, ACWs may be deployed into private or secure networks in which case the commander is unable to communicate with them. In such cases the degree of effectiveness of a commander's control over highly ACWs can be questioned and needs to be proven on a case by case basis. What also needs to be reminded in this respect is that the commander's duty to control is continuous and does not cease with the order or the fielding of the weapon even if the ACW has been programmed appropriately.

The above scenario is different from where a subordinate disregards the orders or instructions of a commander. According to the *Bemba* Trial Chamber, the latter scenario describes a situation where effective control is lacking due to disobedience. However, when an ACW acts differently from what it has been instructed to do, it is not because it refuses to follow orders or disregards orders but because it executes the order differently due to the fact that it processes the information differently. The order is still executed within the parameters of command and control. Even if a question arises regarding the commander's effective control over the weapon's processing function when crimes are committed, the circumstances and reasoning are different from those involved in a scenario of disobedience.

Disobedience is also different from a situation where an ACW selects and engages targets even if not specifically instructed by the commander but being consistent with the commander's instructions regarding the type of targets to be engaged and the circumstances under which they should be engaged. In this case, the targeting decisions by the ACW fall within the commander's framework of command and control but, depending on the circumstances, questions may arise about the degree of effective control over the ACW.

A scenario where effective control would be lacking is the case of a 'rogue' ACW, that is, an ACW acting on its own initiative in order to

40 See Christopher M Bishop, *Pattern Recognition and Machine Learning* (Springer 2006).

pursue a goal outside of the framework of command.[41] In this case, its actions fall outside of the commander's command and effective control and command responsibility cannot arise.

The role of intermediaries when determining the effectiveness of command and control was highlighted in the *Bemba* judgment and it is a critical issue when it comes to ACWs.[42] One can say that programmers are intermediaries because they code ACWs prior to their deployment and will often need to update the weapons while they are operational. However, whether or to what extent they interrupt the effectiveness of command and control depends to a large extent on whether they are integrated into the chain of command or not.

If they are integrated into the chain of command, the role of the programmer is to support and enhance the commander's ability to exercise effective command and control by making ACWs operational and by enabling the commander to be kept informed of their activities and to adjust them accordingly. It would require an exceptional set of circumstances for the input of programmers to be sufficiently substantial to interrupt the commander's effective command and control over an ACW. Similarly, it would be most unlikely that programmers who form part of the chain of command are able to exercise effective control over an ACW for the purpose of command responsibility because they do not plan the operation or, more specifically, they do not decide when the weapon will be deployed; which targets will be attacked and with what level of priority; what is the weapon's overall objectives; or when the weapon should be withdrawn. As international courts have consistently held, the exertion of mere influence — no matter how strong — does not equate to effective command and control.[43] It can thus be said that programmers integrated within the chain of command and operating in conformity therewith do not interrupt its effectiveness. Integrating programmers into an effective system of command and control may also be preferable because, otherwise, programming mistakes can reverberate to the command and control level and increase the risk of crimes being committed.

Programmers outside the chain of command can be treated as commanders themselves under certain circumstances. As we said, the law of command responsibility recognises *de facto* authority and control.

41 *Prosecutor v Enver Hadžihasanović and Amir Kubura (Appeals Chamber Judgement)* ICTY-01-47-T (22 April 2008) [202]–[214].

42 The Trial Chamber also distinguishes between the military principle of 'unity of command' from effective command and control with the former not precluding the existence of multiple commanders; *Bemba (Judgment Pursuant to Article 74)* (n 20) [698]–[699].

43 ibid [183].

Moreover, effective command (or authority) and control does not mean exclusive command (or authority) and control.[44] In fact, the law of command responsibility recognises parallel and multiple command structures in which case responsibility is apportioned according to the level of control each person exercised at the relevant time. Thus, if a programmer controls how ACWs operate by feeding the system with pre-determined operational scenarios whereas the commander controls when and how these weapons are fielded in particular circumstances, it can be said that both exercise effective command (or authority) and control over the weapon and can be held responsible under command responsibility providing the other requirements are met.

IV
THE COMMANDER 'KNEW' OR 'SHOULD HAVE KNOWN'

According to Article 28(a)(i) of the ICC Statute, the commander must have 'either knew or, owing to the circumstances at the time, should have known that the forces were committing or about to commit such crimes'.[45]

Knowledge primarily refers to actual knowledge established through direct or circumstantial evidence. Indices of knowledge include the number, type, and scope of illegal acts committed; the time during which they occurred; the number and type of troops involved; the logistics involved; the geographical location of the acts; their widespread occurrence; the tactical tempo of operations; the *modus operandi* of similar illegal acts; the officers and staff involved; and the location of the commander at that time.[46] Where knowledge is inferred, it must be the only reasonable inference even if it is not necessary to demonstrate that a commander had knowledge of the specific crimes.

Consider a scenario where an ACW provides a commander with real-time reports on its activities and, in particular, it informs the commander of which targets it has selected *before* it engages them. Where the cyber weapon proceeds to attack a civilian network, it can be reasonably inferred from the circumstances that the commander must have known

44 ibid [185].
45 ICC Statute (n 7) art 28(a)(i).
46 *Delalić* (n 18) [386].

that a crime was about to be committed. By contrast, knowledge cannot be inferred where the ACW's reports are incomplete or unintelligible or unmanageable because of the size and complexity of the data being fed-back. Moreover, it may be the case that an ACW is able to select and engage a target incredibly quickly, perhaps in a matter of nanoseconds. Even if the ACW issues a report to the commander prior to targeting, the speed at which the target is actually engaged may make it difficult to conclude that the commander must have known that a crime was about to be committed. However, if the ACW reports back to the commander after the target has been engaged and it transpires that crimes have been committed, it can be said that the commander is aware that crimes have occurred.

Article 28 of the ICC Statute diverges from the *ad hoc* tribunals in relation to the second head of *mens rea*. According to the jurisprudence of the *ad hoc* tribunals, the commander must have 'had reason to know' that the criminal act was about to be committed or had been committed, which means that some general 'information was available to him which would have put him on notice of offences committed by subordinates'.[47] Article 28 instead introduces a 'should have known' standard which goes beyond the existence of notice and 'requires more of an active duty on the part of the superior to take the necessary measures to secure knowledge of the conduct of his troops and to inquire, regardless of the availability of information at the time on the commission of the crime'.[48] Consequently, a superior can be deemed 'negligent in failing to acquire knowledge of his subordinates' illegal conduct'.[49] It transpires from this that Article 28 introduces a *mens rea* of negligence by holding commanders responsible where they fail to apprise themselves of the behaviour of subordinates.[50]

In relation to ACWs, it is important to stress that commanders cannot rely upon their lack of technological expertise to claim that they were not aware that crimes were about to be committed or had been committed.[51] The 'should have known' standard envisages proactive commanders who familiarise themselves with the capabilities, decision-making

47 *Delalić* (n 32) [241]; *Hadžihasanović* (n 41) [28]; *Bagilishema* (n 32) [42]; *Prosecutor v Fofana and Kondewa (Judgment)* SCSL-04-14-T (2 August 2007) [233]. See also Law on the Establishment of the Extraordinary Chambers (n 33) art 29 and Henckaerts and Doswald-Beck (n 18) rule 153.

48 *Bemba (Decision Pursuant to Article 61(7)(a) and (b))* (n 33) [433]. The Trial Chamber did not deal with this issue (*Bemba (Judgment Pursuant to Article 74)* (n 20) [196]).

49 ibid [432].

50 Customary law as determined by the *ad hoc* tribunals takes a different approach. See *Delalić* (n 32) [226]; *Bagilishema* (n 32) [32]–[37]; *Fofana* (n 47) [245]; *Prosecutor v Sesay, Kallon, and Gbao (Judgement)* SCSL-04-15-T (2 March 2009) [312].

51 '[T]he fact that cyber operations may be technically complicated does not alone relieve commanders or other superiors of the responsibility for exercising control over their subordinates. Willful or negligent failure to acquire an understanding of such operations is never justification for lack of knowledge'; Schmitt (n 25) 400.

cycle, self-learning cycle as well as the limitations and constraints of their weapons and this is actually what the notion of responsible command requires. At the same time, the 'should have known' standard requires commanders to scrutinise the information presented to them by ACWs rather than to treat it as immediately reliable and actionable. This addresses the risk of automation bias but, more than that, it does not invert the relationship between a commander and an ACW and does not effectively abolish the role of the commander.

That having been said, commanders are only expected to know what reasonable commanders should have known in their position.[52] Thus, where a cyber weapon is highly autonomous and is capable of sensing, learning and adapting to new environments, it may not be reasonable to expect commanders to have known that the weapon was about to commit a crime because it would be unreasonable to expect them to have known all the situational variables and all the different ways that they can be processed. Likewise, where the weapon operates in a private network it may be difficult for a commander to predict in advance how the weapon will react to this unknown environment or to discover that a crime has been committed. This being said, it may be the case that reasonable commanders would not deploy a weapon into an environment where they have no control over it or if they cannot reliably predict how it will act or without testing it.

A reasonable commander is also expected to keep up to date with the latest developments in technology. Imagine, for example, that new software becomes available which enables an ACW to more accurately distinguish between military and civilian networks or to more accurately estimate the extent to which civilian networks will be collaterally damaged during an attack on military networks. Where commanders have this new technology available to them but they fail to implement software upgrades, if a cyber weapon goes on to attack civilian networks or causes excessive collateral damage it could be said that a reasonable commander should have known that the out-dated cyber weapon was prone (or at least more prone) to engage in criminal acts. Alternatively, commanders may give prior authorisation for software updates to be installed automatically. If this is the case, commanders must place limits on which types of upgrades can be automatically installed. In short, commanders must ensure that their authorisation does not extend to updates that may compromise the weapon's ability to comply with IHL.

52 ibid.

It may also transpire that an ACW has misdiagnosed a civilian network as a military network during a training exercise or when operating in the field. Technical reports may also surface which reveal that the software used by the cyber weapon is defective. Similarly, newspapers or cyber security companies may independently report a weapon's defects, erratic behaviour, vulnerabilities or unauthorised conduct. Given that the information available to a commander need not be specific and must be viewed as a whole,[53] it can be said that the commander in these circumstances 'should have known' that crimes had been or were about to be committed.

We said above that a commander under the 'should have known' standard has a positive duty to seek information and to scrutinise information including information presented by the ACW. What happens when, despite this, the commander comes to the wrong conclusion and crimes are committed? If the commander's assessment of the information was reasonable under the circumstances, the commander is exculpated. While the 'should have known' criterion may impose a proactive duty upon the commander to inquire and find information, it is assessed on a case-by-case basis against material, temporal and other related factors.

V
THE DUTY TO PREVENT OR REPRESS

Article 28 of the ICC Statute imposes three distinct duties upon a commander, namely, to prevent the commission of crimes; to repress crimes; or to submit the matter to the competent authorities for investigation and prosecution. Existing jurisprudence as developed mainly by the *ad hoc* tribunals instead imposes on commanders a duty to prevent or punish. The duty to repress included in Article 28 can include the duty to punish and the duty to submit the matter to a competent authority.[54] It is for this reason that we will concentrate on the duty to prevent and the duty to repress. That said, it is important to make two points. The first point is that these duties must be viewed on a continuum and as a spectrum of particular duties rather than as alternatives.[55] The second point is

53 *Prosecutor v Krnojelac (Public Version of the Confidential Decision on Prosecution's Motion to Admit Additional Evidence Pursuant to Rule 115 of the Rules of Procedure and Evidence Filed on 11 September 2003)* ICTT-97-25-A (16 September 2003) [169].
54 *Bemba (Judgment Pursuant to Article 74)* (n 20) [205]-[209].
55 *Bemba (Decision Pursuant to Article 61(7)(a) and (b))* (n 33) [439]-[441].

that these duties are conditional; they are assessed against the criteria of necessity and reasonableness which in turn are assessed against the criterion of effective command and control.[56]

Necessary are those measures that are appropriate to discharge the duty to prevent or repress and reasonable are those that fall within the commander's effective control.[57] The necessity and reasonableness of the measures is also assessed against the scope of the underlying crimes, the reliability of the available evidence[58] and the limitations presented when a commander is located some distance from where the crimes take place.[59] Moreover, whether the measures are necessary and reasonable does not necessarily depend on whether they were limited in 'mandate, execution, and/or results'[60] but, instead, on whether any shortcomings in this regard were sufficiently serious; the commander was aware of the shortcomings; it was materially possible to correct the shortcomings; and the shortcomings fell within the commander's authority to remedy.[61]

The duty to prevent spans from before the commission of crimes to their actual commission.[62] This duty comports with the general obligation to ensure respect for IHL[63] and with the special position and powers of the commander as a steward of IHL. To fulfil this duty, a commander must ensure that 'forces are adequately trained in IHL; secure reports that military actions were carried out in accordance with IHL; issue orders aimed at bringing the relevant practices into accord with IHL; take disciplinary measures to prevent the commission of atrocities by the troops under the superior's command'.[64] The commander must also: '(i) issue orders specifically meant to prevent the crimes, as opposed to merely issuing routine

56 *Bemba (Judgment Pursuant to Article 74)* (n 20) [197]–[200]. See further Harmen van der Wilt and Maria Nybondas, 'The Control Requirement of Command Responsibility: New Insights and Lingering Questions Offered by the *Bemba* Appeals Chamber Case' in Rogier Bartels, Jeroen C van den Boogaard, Paul AL Ducheine, Eric Pouw and Joop Voetelink (eds), *Military Operations and the Notion of Control under International Law* (Springer 2020) 327.

57 *Bemba (Judgment Pursuant to Article 74)* (n 20) [197]–[199]; *Hadžihasanović* (n 27) [121]–[128]; *Halilović* (n 17) [79]–[100].

58 *Prosecutor v Bemba (Judgment on the Appeal of Mr Jean-Pierre Bemba Gombo against Trial Chamber III's "Judgment pursuant to Article 74 of the Statute")* ICC-01/05-01/08 A (8 June 2018) [183].

59 ibid [189].

60 *Bemba (Judgment Pursuant to Article 74)* (n 20) [720].

61 *Bemba (Judgment on the Appeal of Mr Jean-Pierre Bemba Gombo)* (n 58) [180]–[181].

62 *Bemba (Decision Pursuant to Article 61(7)(a) and (b))* (n 33) [437].

63 Common Article 1 to the Four Geneva Conventions 1949: Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (entered into force 21 October 1950) 75 UNTS 31, art 1; Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea (entered into force 21 October 1950) 75 UNTS 85; Geneva Convention (III) relative to the Treatment of Prisoners of War (entered into force 21 October 1950) 75 UNTS 135, art 1; Convention (IV) relative to the Protection of Civilian Persons in Time of War (entered into force 21 October 1950) 75 UNTS 287 art 1.

64 *Bemba (Decision Pursuant to Article 61(7)(a) and (b))* (n 33) [438].

orders; (ii) protest against or criticising criminal conduct; (iii) insist before a superior authority that immediate action be taken; (iv) postpone military operations; (v) suspend, exclude, or redeploy violent subordinates; and (vi) conduct military operations in such a way as to lower the risk of specific crimes or to remove opportunities for their commission'.[65]

The above can very well apply to ACWs. In the first place, a commander must take all reasonable steps to ensure that an ACW is programmed in such a way as to enable it to comply with IHL and with the objectives of the operation. This requires a commander to ensure among others that the weapon has been tested and verified to ensure that it operates as anticipated; is functional and reliable; is secure and protected from interference through anti-tamper mechanisms; has the required connectivity; is supported by a robust and resilient communication and information exchange system; is subject to spatial and temporal limitations; operates a 'capture first' command mechanism; can be programmed to recognise a list of protected targets; gives warnings before launching attacks; and is subject to real-time supervision. This does not mean that commanders must test each and every aspect of the cyber weapon for defects, which would be a near impossible task given how many lines of code are written into computer algorithms.[66] Rather, commanders must take all necessary and reasonable measures to identify and resolve any defects prior to deployment. Moreover, a commander must be trained to use the weapon and be aware of its operational capabilities and limitations, which means that the commander must be able to respond to problems if they occur. Again, this does not mean that a commander must be able to predict every aspect of the weapon's behaviour, which would be tall order in the context of an autonomous weapon operating in a virtual domain.[67] Instead, what is required is that the commander has a sufficiently sound understanding of the weapon so as to be confident that its activities will conform to IHL and that s/he will be able to intervene if required.

Once deployed, if commanders become aware that ACWs are about to commit crimes, they must again take all necessary and reasonable measures to prevent that activity or repress it where it is ongoing. This may require the cyber weapon's algorithm to be recoded or, where this would be insufficient to prevent crimes, the commander would be expected to

65 *Bemba (Judgment Pursuant to Article 74)* (n 20) [204].
66 Interview with Alan C Schultz, *Laboratory for Autonomous Systems Research* (US Naval Research Lab, 28 January 2016).
67 Interview with Leslie Pack Kaelbling, *Learning and Intelligent Systems Group, Computer Science and Artificial Intelligence Laboratory* (Massachusetts Institute of Technology, 16 September 2016).

deactivate the weapon. Time is an important factor here because cyber weapons — especially when operating autonomously — are able to process and effectuate decisions very quickly. The upshot is that commanders may not have the time to prevent the cyber weapon from acting. In addition, it may be the case that a commander cannot interact with an ACW where it is operating in a closed network. This would mean that there is no opportunity for a commander to adjust, override or deactivate the weapon if problems arise. In these circumstances, it may be necessary and reasonable for the commander to inform the opposing party that a cyber weapon is operating on one of its networks and at risk of engaging in criminal behaviour.

Moving now to the duty to repress, it arises during ongoing crimes and after the commission of crimes.[68] It is also a broad duty; it can include criminal, disciplinary or administrative measures or punishment, criminal prosecution, investigations, reporting to relevant authorities or any other measure that can repress criminal activity. The Trial Chamber in *Bemba* defined repress as to 'put down', 'subdue, 'restrain' and 'keep or hold back',[69] which is important when this duty applies to non-human agents where some of the aforementioned actions are not appliable.

At a minimum, where an ACW commits a crime as a result of a design flaw, it is incumbent upon the commander to take steps to patch the defect which may mean referring it to the programmer or to the military cyber command authorities for re-assessment and re-programming. If that is not possible or the defect cannot be corrected immediately, it would be necessary for the commander to pull the weapon from the field until it can be safely re-deployed.

An interesting question is whether a commander is under a duty to repress when an ACW has committed a crime while under the command and control of a previous commander. This scenario may arise in the case of ACWs not only because of the usual change of commands but also because, given the virtual and interconnected nature of cyberspace, ACWs can produce reverberating effects and any crimes they commit may not become known until some time after the attack has been completed and a full technical assessment has been conducted.

68 *Hadžihasanović* (n 27) [125]; *Bemba (Judgment Pursuant to Article 74)* (n 20) [206].
69 ibid [205].

The above raises the spectre of successor command responsibility. The Nuremberg Military Tribunal,[70] *ad hoc* tribunals[71] and ICC[72] have rejected the possibility of successor responsibility because they require the commission of crimes by subordinates to coincide with the commander's exercise of command and control at the time the crimes were committed. It seems that existing jurisprudence takes a formal and time-limited approach to command and control. The causal link between the crimes and the commander's failure to exercise proper command required by Article 28 is another reason advocating against successor responsibility.[73]

In our opinion, the acceptance of successor command responsibility rests on the nature and purpose of command responsibility.[74] If command responsibility entails responsibility for the crimes of subordinates, a successor commander cannot be held responsible for failing to repress crimes that were committed on another commander's watch. But as we argued in section II, the better approach is to see command responsibility as responsibility for dereliction of duty and, if the purpose of command responsibility is to ensure compliance with the law, a successor commander has a duty to repress past crimes. Thus, if a successor commander is aware that crimes have been committed by an ACW, s/he must reassess, re-programme or decommission it or request others to do so. Otherwise, time-limiting the commander's duty undermines IHL and the aims of command responsibility. If the commander fails to do so and uses the same cyber weapon, this would amount to a dereliction of the duty to prevent further crimes[75] and, if war crimes are indeed committed, s/he can also be charged as perpetrator or accomplice since, as we said in section II, command responsibility and individual criminal responsibility can run in parallel.

In relation to the fact pattern mentioned above where an ACW is deployed while under the effective command and control of a former commander but its violent effects are not felt until a later point in time because its activation is time-delayed (as would be the case with a logic

70 *Wilhelm* (n 21) 1230.

71 *Prosecutor v Enver Hadžihasanović and Amir Kubura (Decision on Interlocutory Appeal Challenging Jurisdiction in Relation to Command Responsibility)* ICTY-01-47-AR72 (16 July 2003) [51]; *Prosecutor v Orić (Appeals Chamber Judgement)* ICTY-03-68-A (3 July 2008) [167]. See Christopher Greenwood, 'Command Responsibility and the *Hadžihasanović* Decision' (2004) 2 Journal of International Criminal Justice 598.

72 *Bemba (Decision Pursuant to Article 61(7)(a) and (b))* (n 33) [419] ('[T]he Chamber is of the view that according to article 28(a) of the Statute, the suspect must have had effective control *at least* when the crimes were about to be committed').

73 Ambos (Treatise) (n 18) 219–20. For our views on causality, see below, section VI.

74 Barrie Sander, 'Unraveling the Confusion Concerning Successor Superior Responsibility in the ICTY Jurisprudence' (2010) 23 Leiden Journal of International Law 105.

75 *Hadžihasanović* (n 41) [30]–[31]; *Prosecutor v Orić (Judgement)* ICTY-03-68-T (30 June 2006) [326].

bomb, for example), if a successor commander becomes aware of the weapon's deployment and knows that, once activated, it will result in the commission of a crime, s/he should take all necessary and reasonable steps to prevent its activation. If the commander only becomes aware of the weapon after it has been activated and crimes have occurred, the duty to repress would be triggered as discussed in the preceding lines.

VI
CAUSALITY

According to Article 28 of the ICC Statute, crimes must be committed 'as a result' of the superior's failure to exercise proper control. Article 28 therefore seems to introduce causality into the doctrine of command responsibility in contrast to the *ad hoc* tribunals which have rejected causality.[76]

Even if Article 28 requires causality, it does not provide much clarity as to its scope. For example, commanders cannot be said to 'cause' subordinates to commit crimes where they fail to repress[77] or report the matter to a competent authority and this points to treating command responsibility as responsibility for dereliction of duty rather than as responsibility for participation in the crimes of subordinates. Causality may however be relevant in relation to the duty to prevent. But how can a commander's failure cause a crime of intent, such as genocide, where the applicable *mens rea* is the 'should have known' standard?

Another difficulty concerns the required standard of causation. In *Bemba*, the Trial Chamber said that the 'but for' test is such a threshold but not the only one.[78] Other judges established a causal link on the basis of the 'high probability' that crimes would not have been committed had the commander discharged his or her duties to prevent,[79] whereas other judges dismissed the need for a causal link altogether.[80]

76 *Delalić* (n 18) [398]–[400]; *Blaskić* (n 14) [77]; *Halilović* (n 17) [78]; *Hadžihasanović* (n 41) [40]; *Orić (Judgement)* (n 75) [338]. Also, no causality is required by the Law on the Establishment of the Extraordinary Chambers (n 33).

77 *Bemba (Decision Pursuant to Article 61(7)(a) and (b))* (n 33) [424]; *Delalić* (n 18) [400]; *Orić (Judgement)* (n 75) [338].

78 *Bemba (Judgment Pursuant to Article 74)* (n 20) [213].

79 *Bemba (Judgment Pursuant to Article 74)* (n 20) (Separate Opinion of Judge Steiner) [24]; *Bemba (Judgment on the Appeal of Mr Jean-Pierre Bemba Gombo)* (n 58) (Dissenting Opinion of Judge Monageng and Judge Hofmanski) [339].

80 ibid (Separate Opinion of Judge Van Den Wyngaert and Judge Morrison) [55]–[56].

In our view, if causation is an element of command responsibility it is ingrained in the notion of effective command and control. This is because, in order to establish the commander's failure to exercise proper control, it needs to be established first that the commander had effective command and control and, as we said previously, effective command and control also includes the ability to prevent or repress. Consequently, the commander's failure to fulfil his/her duty to prevent or repress when the material ability existed indicates a lack of proper control and is what links the commander to the underlying crime. Put differently, it is a case of objective causality proved by the commander's failure to prevent or repress without needing to also prove why the failure to exercise proper control could cause the crimes. This approach is closer to the approach taken by the *ad hoc* tribunals[81] and also comports with what we said in the introduction that macro and micro command are interrelated, interdependent and integrated and cannot therefore be separated.[82] It also comports with our approach to command responsibility as responsibility for dereliction of duty. What this means in the case at hand is that if a commander who has effective command and control detects, for example, a code malfunction but fails to correct it, a causal link with any committed crimes is established because s/he did not exercise his/her command properly by preventing the crimes.

VII
CRIMES COMMITTED

Under Article 28 of the ICC Statute, commanders are held criminally responsible under command responsibility for failing to prevent or repress the crimes committed by their subordinates. It is therefore important to

81 The French text seems to comport with this interpretation: 'Un chef militaire ou une personne faisant effectivement fonction de chef militaire est pénalement responsable des crimes relevant de la compétence de la Cour commis par des forces placées sous son commandement et son contrôle effectifs, ou sous son autorité et son contrôle effectifs, selon le cas, *lorsqu'il ou elle n'a pas exercé le contrôle qui convenait sur ces forces …*' (italics added).

82 This also means that, as we have argued, the failure to exercise control properly and the failure to prevent, repress or submit the matter are interrelated and do not constitute two separate elements that need to be established separately. Cf Separate Opinion of Judge Sylvia Steiner and Separate Opinion of Judge Kuniko Ozaki in *Bemba (Judgment Pursuant to Article 74)* (n 20) Annex 1 and II respectively. According to another approach, there is no causality but 'as a result' refers to the responsibility of the commander for his/her omission; 'Amnesty International Amicus Curiae Observations on Superior Responsibility Submitted Pursuant to Rule 103 of the Rules of Procedure and Evidence' ICC-01/05-01/08-406 (20 April 2009) [39]–[40].

explain what the term 'crimes committed' means[83] in the context of Article 28 and whether ACWs operating under a system of effective command and control can commit crimes. These questions interrelate and will be considered in tandem. It is important however to stress that this question is different from the question of whether subordinates can be held criminally responsible. Command responsibility is triggered when crimes are being committed or have been committed and not when subordinates are held criminally responsible for these crimes. This is a crucial distinction to make and it has implications for ACWs to the extent that they cannot be held criminally responsible because they are not moral agents.

One approach is to say that a crime is committed when both its *actus reus* and *mens rea* are present.[84] An ACW can of course commit the *actus reus* of a crime; for instance, by directly targeting a civilian network. The immediate question is whether they can have the requisite *mens rea* which, as we have noted, comprises intent or knowledge. According to Article 30(3) of the ICC Statute, knowledge is defined as an 'awareness that a circumstance exists or a consequence will occur in the ordinary course of events'. An ACW can have awareness of a circumstance where it is sensed or recognised. Because ACWs are aware of their capabilities, they can also be aware of the consequences of their actions; for example, that if an order is executed, the target will be destroyed. Moreover, ACWs possessing self-learning capabilities are able to learn from experience or 'trial and error' and use this information to enrich and adjust their knowledge and actions. It can thus be said that ACWs can have ingrained as well as acquired knowledge.

ACWs can also fulfil the *mens rea* of intent in relation to consequences which, according to Article 30(2)(b) of the ICC Statute, is defined as awareness that consequences 'will occur in the ordinary course of events'.[85] Evidently, there is overlap between the *mens rea* of intent and the *mens rea* of knowledge in relation to consequences. Regarding the required level of certainty, the ICC requires virtual certainty.[86] Virtual certainty is not absolute certainty which does not exist even in human reasoning. Virtual certainty means that any uncertainty that lingers is non-consequential. It follows from this that, depending on how they are programmed to reduce uncertainty, ACWs can have virtual certainty. For

83 Triffterer and Ambos (n 18) 1088–9.
84 *Williamson v Norris* [1899] 1 QB 7, 14.
85 ICC Statute (n 7) art 30(2)(b).
86 'Thus, this form of criminal intent presupposes that the person knows that his or her actions will necessarily bring about the consequence in question, barring an unforeseen or unexpected intervention or event to prevent its occurrence. In other words, it is nigh on impossible for him or her to envisage that the consequence will not occur'; *Prosecutor v Katanga (Judgment Pursuant to Article 74 of the Statute)* ICC-01/04-01/07 (7 March 2014) [777].

instance, when they attack a particular target in order to destroy it, they are certain of the consequences because they have been pre-determined in their coding whereas if doubt has been coded, they can abstain from acting because they are able to recognise the resulting consequences.

According to another approach, to commit a crime means to commit a proscribed act.[87] A crime as a legally wrongful act[88] can be decoupled from the notion of culpability (*mens rea*) which is about the attribution of culpability to a moral agent due to his or her personal stance towards it. What makes the act wrongful and legally proscribed is its harmful *actus* (with harm not necessarily being physical) rather than its attribution to a moral agent. It follows from this that the underlying crime for command responsibility purposes is an objective circumstance that is established where the *actus reus* or aspects of the *actus reus* that condition said crime are present. According to this view, ACWs can fulfil the objective elements of a crime for example where they directly target civilian networks, a wrongful act in itself.

In our opinion, this is a better approach and is supported by a number of other considerations. First, the commander does not need to know the specificities of the crimes or the identities of the perpetrators other than in general terms as being his/her subordinates.[89] Moreover, the obligation to prevent requires action before the commission of a crime (that is, while it is unfolding), in which case not all of the elements of the crime are present. Even when a commander suspects that a crime is about to be committed, s/he must intervene.[90] Also, the obligation to repress includes among others an obligation to investigate or institute criminal proceedings which will eventually establish the facts and/or culpability. All this means that command responsibility can be activated even if all the elements of a crime have not been fulfilled; and of course, before *mens rea* is established. What activates the commander's duty to act in these circumstances is instead the existence of acts that condition the commission of a crime.

Second, in order to enhance the effectiveness of command responsibility, international jurisprudence has not only applied this doctrine to all modes of perpetration or participation included in Article 25(3)(a)–(f) ICC Statute [91] but has also applied it to inchoate crimes that is, to incomplete crimes.

87 *Orić (Judgement)* (n 75) [296] (referring to the Prosecution brief).
88 Glanville Williams, 'The Definition of a Crime' (1955) 8 Current Legal Problems 107.
89 *Orić (Appeals Chamber Judgement)* (n 71) [35]; *Bemba (Judgment Pursuant to Article 74)* (n 20) [194].
90 *Hadžihasanović* (n 27) [852].
91 *Orić (Judgement)* (n 75) [294], [295]–[306] and [328].

Third, command responsibility still attaches in situations involving exculpatory circumstances. For example, where a subordinate engages in a wrongful act (i.e. killing civilians) but does not have the requisite *mens rea* due to an exculpatory circumstance such as mental impairment, command responsibility will be attached if the commander failed to prevent or repress. This can be contrasted with the case where the wrongful character of the act is removed because of a justification — for example, where a civilian is killed in self-defence. In this case, there is no wrongful act and a commander cannot incur responsibility for his/her failure to prevent or repress what is an inherently justifiable act.

Fourth, and most importantly, this approach accords with the nature of command responsibility as responsibility for dereliction of duty rather than responsibility for the crimes of subordinates, and comports with its rationale, which is to ensure that violations of IHL are prevented and repressed.

VIII
CONCLUSION

ACWs are likely to become a central feature of contemporary armed conflict. No technology is fail-safe and the question that arises is who can be held responsible where an ACW commits a war crime. This chapter has addressed this question from the perspective of the doctrine of command responsibility by placing ACWs within a system of command and control. It then explained how its constituent elements as they have been interpreted in international jurisprudence apply to ACWs. A number of questions have been raised in this context. To what extent does the autonomous nature of cyber weapons preclude commanders from exercising effective command and control over them? What is the role of intermediaries such as programmers and do they interfere with the effectiveness of a commander's control over ACWs? Do successor commanders have command responsibility?

As we move along the autonomy continuum, when can it be said that a reasonable commander knew or should have known that an ACW was about to commit or had committed a crime? What necessary and reasonable measures must commanders take to discharge their duty to prevent or repress? Can ACWs commit a crime for the purpose of command responsibility?

In addressing these questions, it became clear that the law of command responsibility can apply to ACWs with the necessary adjustments and interpretative refinement even if it faces serious challenges as one moves towards the upper echelons of autonomy. It is thus important to ensure that it remains compatible with legal principles for the purposes of legality and accountability. We believe that being human-made technology, it can be developed responsibly and aligned with the principles and aims of humanitarian law and international criminal law and that the doctrine of command responsibility represents an important and powerful tool for ensuring that international law remains in the loop when ACWs are used.

Autonomous cyber capabilities are comparable to kinetic autonomous weapons systems in the opportunities and risks they harbour. Yet, related legal and political debates so actively resonating with regard to kinetic systems have been largely led along parallel but not convergent tracks in respect of cyber means. In order to take a step towards a meaningful dialogue between and, where needed, convergence of the two discourses, the edited collection at hand combines the insights of acknowledged experts in law, ethics and technology. It serves as a valuable reference piece for scholars and students of international law as applied to autonomous capabilities and/or cyber operations as well as a stimulating read for legal advisors to States and international organisations, technologists or a broader audience interested in the future of cyber warfare.

9 789916 956526