CCDCOE
NATO COOPERATIVE
CYBER DEFENCE
CENTRE OF EXCELLENCE

Autonomous Cyber Capabilities under International Law

Rain Liivoja

Associate Professor, TC Beirne School of Law, University of Queensland

Maarja Naagel

Legal Researcher, NATO CCDCOE

Ann Väljataga

Legal Researcher, NATO CCDCOE

Tallinn 2019

CCDCOE

The NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) is a NATO-accredited cyber defence hub focusing on research, training and exercises. It represents a community of 25 nations providing a 360-degree look at cyber defence, with expertise in the areas of technology, strategy, operations and law. The heart of the Centre is a diverse group of international experts from military, government, academia and industry backgrounds.

The CCDCOE is home to the *Tallinn Manual 2.0*, the most comprehensive guide on how International Law applies to cyber operations. The Centre organises the world's largest and most complex international live-fire cyber defence exercise Locked Shields. Every spring the Centre hosts in Tallinn the International Conference on Cyber Conflict, CyCon, a unique event joining key experts and decision-makers of the global cyber defence community. As of January 2018 CCDCOE is responsible for identifying and coordinating education and training solutions in the field of cyber defence operations for all NATO bodies across the Alliance.

The Centre is staffed and financed by its member nations – to this date Austria, Belgium, Bulgaria, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Spain, Sweden, Turkey, the United Kingdom and the United States. NATO-accredited centres of excellence are not part of the NATO Command Structure.

www.ccdcoe.org
publications@ccdcoe.org

Disclaimer

This publication is a product of the NATO Cooperative Cyber Defence Centre of Excellence (the Centre). It does not necessarily reflect the policy or the opinion of the Centre or NATO. The Centre may not be held responsible for any loss or harm arising from the use of information contained in this publication and is not responsible for the content of the external sources, including external websites referenced in this publication.

Digital or hard copies of this publication may be produced for internal use within NATO and for personal or educational use when for non-profit and non-commercial purpose, provided that copies bear a full citation.

Note

The authors would like to thank Mr Mihkel Pikkat for his valuable research assistance.

Table of Contents

List of Abbreviations

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
AWS	autonomous weapon system
C2	command and control
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
GGE LAWS	Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems
CIA	US Central Intelligence Agency
C-RAM	counter rocket, artillery and mortar system
CIWS	close-in weapon system
DARPA	US Department of Defence's Advanced Research Projects Agency
DDoS	distributed denial-of-service attack
DoD	US Department of Defence
GPS	Global Positioning System
ICC	International Criminal Court
ICD	industrial control device
ICL	international criminal law
ICRC	International Committee of the Red Cross
ICTR	International Criminal Tribunal for Rwanda
IGE	International Group of Experts
IHL	international humanitarian law
ILC	International Law Commission
IP	internet protocol
LAWS	lethal autonomous weapons systems
UN GGE	United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security
USB	Universal Serial Bus

Introduction

The application of international law to cyber operations and to the use of autonomous military technology continues to be the subject of intensive debate. Yet the discussions regarding the international law implications of cyber and autonomy have largely taken place separately from each other and somewhat out of sync. The debate concerning cyber operations is much further along. In 2004, the United Nations Secretary-General established a Group of Governmental Experts (UN GGE) to consider 'developments in the field of information and telecommunications in the context of international security'. From 2009 to 2015, successive UN GGEs produced a series of substantive reports, which touched upon the application of international law to cyberspace.[1] In 2009, the NATO Cooperative Cyber Defence Centre of Excellence convened an International Group of Experts (IGE), which released *Tallinn Manual 1.0* in 2013 and *Tallinn Manual 2.0* in 2017.[2]

The international law implications of autonomous weapon systems (AWS)[3] did not attract major international attention until about 2012. Notable developments around that time were the publication of doctrine documents by the British and the US armed forces,[4] and reports by Human Rights Watch and UN Special Rapporteur Christof Heyns that expressed legal concerns.[5] To address these concerns, a Meeting of Experts was convened under the auspices of the Convention on Certain Conventional Weapons (CCW) which met annually from 2014 to 2016,[6] and a more formal Group of Governmental Experts (CCW GGE) has been meeting since 2017.[7] These processes have been accompanied by a significant amount of scholarly activity but there has not yet been a concerted academic effort to clarify the law along the lines of the *Tallinn Manuals.*

The discussions relating to AWS and international law have not achieved quite the same level of maturity as the discussion around cyber operations. In relation to the latter, a number of

[1] 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security' (30 July 2010) UN Doc A/65/201; '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; '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 (2015 UN GGE Report).

[2] *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Cambridge University Press 2013); *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017).

[3] AWS refers in this report exclusively to kinetic weapon systems that have autonomous functions (as explained in Part 1.2 below). For the cyber equivalent, we use the term 'autonomous cyber weapon', even though from a legal perspective these should be seen as a type of autonomous weapon system.

[4] UK Ministry of Defence, 'Joint Doctrine Note (JDN) 2/11: The UK Approach to Unmanned Aircraft Systems' (30 March 2011); US Department of Defence, 'DoD Directive No. 3000.09: Autonomy in Weapon Systems' (21 November 2012).

[5] Human Rights Watch, *Losing Humanity: The Case against Killer Robots* (2012); Christof Heyns, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions' (9 April 2013) UN Doc A/HRC/23/47.

[6] 'Report of the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)' (10 June 2014) UN Doc CCW/MSP/2014/3; 'Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)' (1 June 2015) UN Doc CCW/MSP/2015/3; 'Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)' (10 June 2016) UN Doc CCW/CONF.V/2.

[7] 'Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS)' (22 December 2017) UN Doc CCW/GGE.1/2017/3; '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 CCW GGE Report).

points of convergence between in the views of States and commentators have emerged, even though there are, as always, some outliers. With regard to AWS, however, some fairly fundamental legal questions remain unresolved, not least the question of what constitutes an AWS and whether the use of AWS would be inconsistent with existing law.

The interconnection between autonomy and cyber has been occasionally acknowledged in the ongoing discussions. For example, the CCW GGE agreed in 2017 that, when developing weapons systems with autonomous functionality, States must consider, *inter alia*, 'non-physical safeguards (including cyber-security against hacking or data spoofing)'.[8] That raises the question of whether States have an obligation under existing international law to put certain safeguards in place.

The legal implications of autonomous cyber capabilities have also received little attention. This is despite the view that the most dramatic advancements in terms of autonomous military capabilities have taken place in the cyber context.[9] Indeed, it has been argued that the first highly autonomous weapon used was Stuxnet, with Jason Healey, for example, writing that 'Stuxnet […] appears to be the first autonomous weapon with an algorithm, not a human hand, pulling the trigger'.[10]

This paper seeks to make a preliminary foray into the international law aspects of autonomous cyber capabilities. We emphasise that debatable legal issues remain in relation to regular cyber capabilities and wholly unresolved sets of issues in relation to AWS. These uncertainties become compounded when it comes to autonomous cyber capabilities. Accordingly, this paper does not purport to provide definitive answers, and by no means tries to be comprehensive in identifying legal and policy issues.

The paper proceeds in the following manner. Part 1 provides a general technological background. It explains the notion of autonomy in technological systems and gives examples of autonomous functionality in defensive and offensive cyber capabilities. Part 2 makes a few overarching observations about the relationship between autonomy and the law, and Part 3 considers breaches of sovereignty that may result from the use of autonomous functionality in cyber capabilities. Parts 4 and 5 discuss autonomous cyber capabilities in the context of *jus ad bellum* and *jus in bello* respectively, and Part 6 looks at responsibility under international law for uses of autonomous cyber capabilities.

[8] 2018 CCW GGE Report (n 7) para 21(e).

[9] Jeffrey S Thurnher, 'Feasible Precautions in Attack and Autonomous Weapons' in Wolff Heintschel von Heinegg, Robert Frau and Tassilo Singer (eds), *Dehumanization of Warfare* (Springer 2018) 104.

[10] Jason Healey, 'Stuxnet and the Dawn of Algorithmic Warfare' *Huffington Post* (16 April 2013) <www.huffingtonpost.com/jason-healey/stuxnet-cyberwarfare_b_3091274.html>.

1 Autonomy in cyber capabilities

1.1 Different approaches to autonomy

Many technological systems – civilian and military, existing and hypothetical – have been characterised as 'autonomous'. This term has been applied, for example, to: driverless trains and buses found at airports or in mass transit systems; the 'self-driving' cars that are taking to the streets in many parts of the world; unmanned aircraft that can take off, navigate or land on autopilot; point defence weapon systems such as CIWS or C-RAM; and loitering munitions such as the Harpy. Much ink has been spilt over whether any of these systems can be regarded as 'truly' autonomous.

Much of the difficulty arises from the notion of 'autonomy' being deployed quite differently in different disciplines.[11] In moral and political philosophy, and cognate fields such as law, 'autonomy' tends to be used with considerable precision; this is helped by the fact that, over the past three decades or so, there has been a significant amount of theorisation and debate about its precise meaning and implications.[12] In technical disciplines such as computer science and robotics, the situation seems radically different. While the terms 'autonomy', 'decisional autonomy', 'levels of autonomy', 'degrees of autonomy', 'autonomous', 'autonomous system', 'autonomous agent', 'autonomous control' and so on are frequently used, this use tends to be quite loose and often without any overt attempt to clarify what 'autonomy' means. Indeed, the use of the word 'autonomy' in artificial intelligence (AI) and robotics has been criticised as 'undisciplined', such that it has arguably 'robbed these fields of an important concept'.[13]

The second difficulty naturally arises from the first. There is considerable disagreement between disciplines as to what autonomy, should one choose to define it, precisely entails. In philosophy, the central idea is relatively easy to outline, although the details get quite complicated. In the philosophical literature, the use of the word autonomy closely tracks its Greek roots – *autós*, 'self' and *nómos*, 'law'. Thus, autonomy means self-regulation or self-governance – the ability of a system to establish its own rules of conduct and then to follow them. Thus, according to Tim Smithers:

> 'the underlying notion is one of self-law making, or self-governing, and it is closely related to the concepts of self-identity and self-determination: an autonomous agent is one whose behaviour is regulated by rules or laws generated by itself'.[14]

[11] Tim Smithers, 'Autonomy in Robots and Other Agents' (1997) 34 *Brain & Cognition* 88; Willem FG Haselager, 'Robotics, Philosophy and the Problems of Autonomy' (2005) 13 *Pragmatics & Cognition* 515.

[12] James Stacey Taylor, 'Autonomy' (*Oxford Bibliographies*, 19 May 2017) <www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0167.xml>.

[13] Smithers (n 11) 89.

[14] Ibid.

Similarly, Willem Haselager notes that:

> '[a]uto-nomos, being or setting a law to oneself, indicates the importance of self-regulation or self-government. Autonomy is deeply connected to the capacity to act on one's own behalf and make one's own *choices*, instead of following the goals set by other agents'.[15]

John Christman suggests that:

> 'to be autonomous is to be one's own person, to be directed by considerations, desires, conditions, and characteristics that are not simply imposed externally upon one, but are part of what can somehow be considered one's authentic self'.[16]

In technical fields, however, the meaning of autonomy remains much less clear but generally refers to something less elaborate than autonomy in its philosophical sense. Philip Brey and Johnny Hartz Søraker seem to be on the right tracks when they suggest that:

> '[a]t a minimum, "autonomous" carries *some* of its philosophical meaning in the sense that an autonomous agent should be able to make informed decisions (based on its knowledge base, rules and sensory input) and act accordingly'.[17]

Indeed, many definitions of autonomous systems (whether autonomous computational agents or autonomous robots that have physical sensors and actuators) note the ability of the system to sense its environment and to act in the environment in pursuit of certain goals. Here are a few examples:

- 'The notion of (artificial) agency is often used in computer science to refer to a computer program that is able to act on and interact with its environment'. [18]
- 'Autonomous agents are computational systems that inhabit some complex, dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed'.[19]
- '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'.[20]

An important implication of these definitions, though not always expressly articulated, is that, owing to the ability to sense and act, an autonomous system 'can operate, self-contained,

[15] Haselager (n 11) 519 (original italics).

[16] John Christman, 'Autonomy in Moral and Political Philosophy' in Edward N Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Metaphysics Research Lab, Stanford University 2018) <plato.stanford.edu/archives/spr2018/entries/autonomy-moral>.

[17] Philip Brey and Johnny Hartz Søraker, 'Philosophy of Computing and Information Technology' in Anthonie Meijers (ed), *Philosophy of Technology and Engineering Sciences* (Elsevier Science & Technology 2009) 1373 (original italics).

[18] Ibid 1372.

[19] Pattie Maes, 'Artificial Life Meets Entertainment: Lifelike Autonomous Agents' (1995) 38 *Communications of the ACM* 108, 108.

[20] 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) 5.

under all reasonable conditions without requiring recourse to a human operator'.[21] In other words, autonomy refers more properly to *autonomy from human operators*.

However, different additional requirements have been proposed to distinguish 'truly' autonomous systems or agents.[22] Some of these have been expressed in rather tautological terms: for example, Pattie Maes suggests that an autonomous system must sense and act 'autonomously'.[23] Others emphasise adaptability: according to Robin Murphy, '[a]utonomy means that a robot can adapt to changes in its environment or itself and continue to reach its goal'.[24] Others would require some human-level intelligence: for Pertti Saariluoma, '[a]utonomous systems are technologies with the capacity to perform tasks that previously required human operators to contribute the higher cognitive processes associated with human thinking'.[25]

One common approach is to use such criteria to distinguish between systems that are autonomous and those that are (merely) automated. Nikolaus Correll does so by reference to preset rules. Thus, '[r]obots are *autonomous* when they make decisions in response to their environment vs. simply following a pre-programmed set of motions'.[26] The difficulty with this line of reasoning is that even very complex systems that might be described as autonomous are, at the end of the day, based on code, i.e. pre-programmed rules.

Saariluoma attempts to distinguish between automated and autonomous systems somewhat differently:

> 'Traditional stimulus/response-type technical artefacts are not autonomous systems. For example, a door that opens when it registers human body temperature or movement is an automatic system and independent from users' continuous control, but it is not an autonomous system. If it had the capacity to decide for which people it should open, for instance depending on the gravity of their illness (e.g., Alzheimer's disease) or personality type, it could then be described as having autonomous capacity as it would be conducting a demanding (autonomous) categorization task. Thus, while there is an overlap between automatic and autonomous systems, in practical contexts the two types are different enough to be considered separately'.[27]

While this is interesting, it is not entirely convincing. Saariluoma effectively requires an autonomous system to have super-human qualities. A person guarding a door and opening it with the push of a button would not be able to tell whether an approaching person has Alzheimer's disease. Also, there would be grey areas: would a door-opening system that can engage in facial recognition and only open the door to persons whose photos can be found in a database qualify as an automated or an autonomous system?

[21] Robin R Murphy, *Introduction to AI Robotics* (MIT Press 2000) 4.

[22] See also Brey and Søraker (n 17) 1372: 'Different sets of requirements have been proposed for what it means to be an agent, which has resulted in a complex and often inconsistent set of terms for different kinds of agency'.

[23] Maes (n 19) 108.

[24] Murphy (n 21) 4.

[25] Pertti Saariluoma, 'Four Challenges in Structuring Human-Autonomous Systems Interaction Design Processes' in Andrew P Williams and Paul D Scharre (eds), *Autonomous Systems: Issues for Defence Policymakers* (Headquarters Supreme Allied Commander Transformation 2016) 226.

[26] Nikolaus Correll, *Introduction to Autonomous Robots* (v17, Magellan Scientific 2016) 15 <open.umn.edu/opentextbooks/BookDetail.aspx?bookId=316> (original italics).

[27] Saariluoma (n 25) 227.

1.2 A broad conception of autonomy

The question about the specific technical requirements of autonomy or what 'true' or 'full' autonomy entails cannot be resolved here. Nor does this seem necessary; we are concerned with the significance of autonomy to the interpretation, application and development of international law. Plainly, an autonomous system need not have human-level intelligence to raise regulatory issues. So, for present purposes, a broad definition of autonomy would be workable and indeed desirable to capture a broad range of systems.

With that in mind, 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.

This broad conception of autonomous systems has a few important implications.

First, autonomy might be, but need not be, facilitated by AI. Thus, we need not engage in a debate about what is or is not AI, and simply note that the development of different AI techniques (such as machine-learning) is likely to produce and improve autonomous capabilities. In other words, AI is an enabler for autonomy but neither synonymous with it nor a prerequisite for it.

Second, this 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. Autonomy has been described as a peculiar kind of control that a human exercises over a system, one that is qualitatively different from manual control.[28] That difference may have legal significance because, even if the human operator does not disappear altogether, their role will change.

Two further preliminary points about autonomy in the technological sense need to be made here. First, autonomy relates to specific functions of a system. Thus, a system may perform some of its functions quite autonomously, while requiring human input for other functions. To give a simple example, a robotic vacuum cleaner might be capable of vacuuming the floor and recharging itself without any input from a human operator, but the operator might need to periodically empty the dustbin. Thus, asking whether the vacuum cleaner, as a whole, is autonomous or not blends the different features of the device into one and thus overlooks the complexity of the system.

Second, autonomy comes in degrees. Different systems require different amounts and types of human input to accomplish some tasks. To return to the example of the robotic vacuum cleaner, some devices might be capable of mapping their areas of operation and avoiding obstacles, whereas other systems might require some input from the human. This fluid aspect of autonomy is sometimes reduced to simple categories. The most popular of these would involve distinguishing between functions with a person 'in the loop' (in constant manual control), 'on the loop' (in a supervisory position) or 'out of the loop' (without any real-time supervision). However, there is often no principled way of drawing the line between those categories.

[28] Tim McFarland, 'Autonomous Weapons and Human Control' (*Humanitarian Law & Policy*, 18 July 2018) <blogs.icrc.org/law-and-policy/2018/07/18/autonomous-weapons-and-human-control>.

It therefore seems more sensible to consider autonomy as being on a spectrum and having different degrees.

In short, the question as to whether some physical system or some cyber capability is autonomous or not is misguided. The better questions would be: in what respect and to what extent is a system capable of autonomous functioning? 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.

To put the discussion of autonomous cyber capabilities into more practical terms, it seems useful to provide some extant and potential examples. To make this more manageable, it also seems helpful to make a basic distinction between offensive and defensive capabilities, while noting that many capabilities have both offensive and defensive dimensions and thus defy easy categorisation.

1.3 Defensive autonomous cyber capabilities

Having adopted a technical rather than philosophical approach to autonomy, this section describes autonomous cyber defence capabilities, including passive and active defence measures. The line between active and passive cyber defence is just as vague as that between defensive and offensive activities. From a legal perspective, drawing a clear line is not necessary in either case as the applicable norms of international law do not depend on the categorisation of measures taken, but on the foreseeable consequences of each specific activity. The element of autonomy does not alter this. Rather, the distinction between defensive and offensive measures reflects the political sensitivities attached to certain cyber capabilities and, in this paper, serves the purpose of systematisation.

Using the analogy of active air and missile defence, active cyber defence can be described as 'direct defensive action taken to destroy, nullify or reduce the effectiveness of cyber threats against friendly forces and assets'.[29] Active cyber defence is generally associated with 'hacking back' or deploying measures outside one's own networks to counter malicious cyber activity against one's networks.[30] Passive cyber defence covers measures other than active cyber defence. It is usually focused on preventing intrusions by making one's network and systems more resilient.

Passive cyber defence measures include cryptography and steganography (analogous to the use of camouflage and stealth aircraft), security engineering and verification, configuration monitoring and management, vulnerability assessment and mitigation, risk assessment, backup and recovery of lost data, and education and training of users. They also include mechanisms to log and monitor network and host activity.[31]

[29] Dorothy E Denning, 'Framework and Principles for Active Cyber Defence' (2014) 40 *Computers & Security* 108, 109.

[30] See, eg, Robert S Dewar, 'The "Triptych of Cyber Security": A Classification of Active Cyber Defence', *2014 6th International Conference on Cyber Conflict (CyCon 2014)* (IEEE 2014).

[31] Denning (n 29) 109.

Measures like firewalls, intrusion detection and prevention systems, and honeypots can be considered borderline – they are active or passive depending on where one draws the line, on how much pro-activeness is required for something to be considered active, and on specific actions of the relevant defence systems.

Autonomy is widely used in measures limited to one's own networks. For example, a firewall that detects suspicious packets of network traffic and removes them from incoming traffic can be considered autonomous in these functions, as it operates without the intervention of a human operator. In fact, the system administrator may never find out what packets were removed unless they specifically look it up. Another example of such an internally functioning autonomous agent would be a program that detects instances of unauthorised access and deletes the data contained in a database on detecting a suspicious access pattern. Given the sheer amount of network traffic flowing through even a regular office network, human intervention or even supervision rarely happens.

A widely discussed example of autonomous active cyber defence is the Mayhem Cyber Reasoning System, the winning system in DARPA's 2016 Cyber Grand Challenge. Although it is a prototype designed to operate in a simplified operating system specifically developed for the Cyber Grand Challenge, it demonstrated remarkable advances in the use of autonomous passive and active cyber defence features.[32] Mayhem and other contestants had to perform three main tasks during the competition. First, they had to protect their software from adversaries by finding and patching vulnerabilities. Second, they needed to keep their software available, functional, and efficient. Third, they needed to exploit vulnerabilities in their adversaries' software, and all that autonomously without any ongoing human input.

1.4 Offensive autonomous cyber capabilities

The most prominent example of an autonomous offensive cyber capability to date is Stuxnet. The malware itself and the circumstances of its deployment have been widely discussed in cyber security and legal literature,[33] and so a brief overview will suffice to highlight its autonomous functionality. The W32.Stuxnet worm was discovered in 2010. It was found to have infected tens of thousands of Windows computers, predominantly in Iran. The target of the worm, however, was a particular setup of an industrial control system (ICS) manufactured by Siemens. The worm appears to have been specifically designed to infect the ICS used to control gas centrifuges at the Natanz nuclear enrichment facility. Stuxnet could manipulate the operation of the centrifuges while supplying innocuous information to the operators. This resulted in the damaging or destruction of around 1,000 centrifuges.

The novelty of Stuxnet was twofold. First, it caused physical damage. As Michael V Hayden, a former director of the US Central Intelligence Agency, noted: 'Previous cyberattacks had

[32] Thanassis Avgerinos and others, 'The Mayhem Cyber Reasoning System' (2018) 16 *IEEE Security & Privacy* 52.

[33] See, eg, Ralph Langner, 'Stuxnet: Dissecting a Cyberwarfare Weapon' (2011) 9 *IEEE Security & Privacy* 49; Nicolas Falliere, Liam O Murchu and Eric Chien, 'W32.Stuxnet Dossier' (Symantec February 2011) Version 1.4 <www.wired.com/images_blogs/threatlevel/2011/02/Symantec-Stuxnet-Update-Feb-2011.pdf>; PW Singer, 'Stuxnet and Its Hidden Lessons on the Ethics of Cyberweapons' (2015) 47 *Case Western Reserve Journal of International Law* 79.

effects limited to other computers […] This is the first attack of a major nature in which a cyberattack was used to effect physical destruction'.[34] Hence, Stuxnet has been called, not unreasonably, the first cyber weapon.

The second novel feature of Stuxnet, important for present purposes, was that controlling it remotely was an option but not a necessity. Stuxnet did not require access to the internet to propagate or to achieve its aim; it relied on local networks and removable media for distribution. It was able to identify, infect and control the target ICS without any further input from the attacker. According to Falliere Murchu and Chien:

> '[w]hile attackers could control Stuxnet with a command and control server […] the key computer was unlikely to have outbound Internet access. Thus, all the functionality required to sabotage a system was embedded directly in the Stuxnet executable'.[35]

Stuxnet appears to have contacted the command-and-control servers essentially to provide evidence of compromise.

Numerous commentators have underlined the autonomous functionality of the worm. Paul Scharre has noted that:

> 'Stuxnet had a tremendous amount of autonomy. It was designed to operate in "air-gapped" networks, which aren't connected to the internet for security reasons. In order to reach inside these protected networks, Stuxnet spread via removable USB flash drives. This also meant that once Stuxnet arrived at its target, it was on its own […] Unlike other malware, it wasn't enough for Stuxnet to give its designers access. Stuxnet had to perform the mission autonomously'.[36]

Stuxnet amounted to an autonomous cyber weapon, but autonomous cyberweapons have not featured with any prominence in the ongoing debate about AWS. Significantly, the US DoD Directive 3000.09 on autonomy in weapon systems does not apply to 'cyberspace systems for cyberspace operations'.[37] Scharre, who led the drafting of the directive, provides the following explanation:

> 'This wasn't because we thought autonomous cyberweapons were uninteresting or unimportant when we wrote the directive. It was because we knew bureaucratically it would be hard enough simply to create a new policy on autonomy. Adding cyber operations would have multiplied the complexity of the problem, making it very likely we would have accomplished nothing at all'.[38]

In other words, the exclusion of autonomous cyber weapons from the directive – and many subsequent discussions – does not appear to be the consequence of a perceived lack of significance. Rather, autonomous cyber weapons have been placed in the 'too-hard basket'.

[34] David E Sanger, 'Obama Ordered Wave of Cyberattacks against Iran' *The New York Times* (1 June 2012) <www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html>.

[35] Falliere, Murchu and Chien (n 33) 3.

[36] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (W W Norton & Company 2018) 214–5; see also Healey (n 10).

[37] US Department of Defence (n 4) para 2.b.

[38] Scharre (n 36) 227–8.

2 General observations regarding the legal issues

Before turning to specific issues that arise under international law in the context of autonomous cyber capabilities, a few general remarks are in order to draw attention to some overarching problems and concerns.

First, the legal issues of garden variety cyber operations and the uncertainties with respect to physical autonomous systems are both *prima facie* relevant to autonomous cyber capabilities. Thus, as a general matter, regulatory complexity increases as legal concerns become compounded. In some circumstances, however, the challenges created by autonomy may be alleviated by the cyber context. For example, with regard to AWS, questions about the distribution of responsibilities and accountability between the operator and the developer of a system have frequently been raised.[39] Cyber capabilities, particularly offensive cyber capabilities, are less likely to be acquired off the shelf. In many, if not most, instances, the developers of the capability are also responsible for deploying the capability, hence becoming the operators. Thus, the issue of allocation of responsibility is less acute with respect to autonomous cyber capabilities compared to physical autonomous systems.

The second issue, related to the first, is the matter of intent. When operating a system manually, the system normally gives effect to the direct intent of the operator. For example, where an operator launches a GPS-guided munition at a target identified by its coordinates, the operator plainly intends to damage or to destroy that target. With an autonomous system, however, the specificity of the operator's intent changes. The operator who deploys a loitering munition intends to destroy certain types of targets – for example, tanks – in a given area of operation. However, they do not necessarily intend to destroy a *specific tank*, but rather *any tank* within the area. This may have legal significance.

Third, one of the persistent worries about autonomous systems, particularly AWS, is that they may operate in an unpredictable fashion,[40] although this is not a problem unique to autonomous systems. Any system, even if strictly manually controlled, may perform in an unintended way. The design and testing processes of all systems are geared towards increasing their predictability, thus reducing the likelihood of mishaps.[41] Also, the idea that a carefully-designed and rigorously tested autonomous system will necessarily be less predictable than a human being who has been deployed to a complex and fast-paced battlespace does not seem to have a solid foundation. Furthermore, not all malfunctions of manually operated systems can necessarily be corrected by the operator, and thus real-time human control is not necessarily a guarantee against unintended consequences.

[39] See, eg, Tim McFarland and Tim McCormack, 'Mind the Gap: Can Developers of Autonomous Weapons Systems Be Liable for War Crimes?' (2014) 90 *International Law Studies* 361.

[40] See, eg, Human Rights Watch, *Losing Humanity* (n 5); Mary Ellen O'Connell, 'Banning Autonomous Killing' in Matthew Evangelista and Henry Shue (eds), *The American Way of Bombing: How Legal and Ethical Norms Change* (Cornell University Press 2013).

[41] Estonia and Finland, 'Categorizing Lethal Autonomous Weapons Systems: A Technical and Legal Perspective to Understanding LAWS' (24 August 2018) UN Doc CCW/GGE2/2018/WP2.

Having said that, predictability remains a potential source of concern. Also, it should be readily admitted that autonomy may reduce the opportunity to correct the misfunctioning of a system and to prevent undesirable consequences. For one thing, the system may operate at speeds that make it impossible for the operator to meaningfully intervene. Also, the operator may be tempted to over-rely on a system that appears to operate normally, not monitoring its operations as vigilantly as appropriate or disregarding information suggesting that the system is malfunctioning. Thus, autonomous functionality can create additional risks of a system not performing as intended. The question that arises here is whether the malfunctioning of a manually controlled system differs, in some legally significant way, from the unanticipated operation of an autonomous system.

3 Sovereignty

Numerous rules and principles of international law restrict the ability of States to undertake cyber operations that affect other States. Three of these norms have the most general application and the broadest effect: the principle of sovereignty, the principle of non-intervention, and the prohibition of the use of force. These three norms relate to each other in that they all derive from the same foundational notion of international law, namely that States enjoy sovereignty. Also, the breaches of these norms form a kind of gradation: an intervention is, by and large, an aggravated form of the breach of sovereignty, whereas the use of force is an aggravated form of intervention. An armed attack is generally seen as a particularly serious form of the use of force.

As we show in this paper, autonomous functionality in cyber capabilities has significance in the application of some of those rules, but not others. In this part, we address the principles of sovereignty and non-intervention. These have particular relevance where cyber operations fall below the threshold of the use of force, which would be the case with the vast majority of cyber operations. In Part 4, we turn to the prohibition of the use of force.

3.1 Breaches of sovereignty

Somewhat surprisingly, the greatest controversy to emerge in relation to international law applicable to cyber operations in the wake of the publication of *Tallinn Manual 2.0* concerns sovereignty. There are two main difficulties. The first of these concerns the nature of the principle itself. The IGE who drafted the *Tallinn Manual* plainly viewed the principle of sovereignty as both an overarching inspirational principle for more specific rules of international law and a primary rule in its own right. This understanding underpins much of Part 1 of the *Manual* and becomes most evident in Rule 4 which stipulates that '[a] State must not conduct cyber operations that violate the sovereignty of another State'.[42] It has been aptly noted that this rule's importance 'lies chiefly in its characterisation of violations of sovereignty as internationally wrongful conduct'.[43]

The alternative view, which did not emerge during the drafting of the *Manual* but subsequently, suggests that 'sovereignty serves as a principle of international law that guides state interactions, but is not itself a binding rule that dictates results under international law'.[44] In practical terms, this approach would mean that the principle of sovereignty 'does not establish an absolute bar against […] cyber operations that affect cyberinfrastructure within another state, provided that the effects do not rise to the level of an unlawful use of force or an unlawful

[42] *Tallinn Manual 2.0* (n 2) rule 4.

[43] Sean Watts and Theodore Richard, 'Baseline Territorial Sovereignty and Cyberspace' (2018) 22 *Lewis & Clark Law Review* 803, 856.

[44] Gary P Corn and Robert Taylor, 'Sovereignty in the Age of Cyber' (2017) 111 *AJIL Unbound* 207, 208.

intervention'.[45] While initially put forward by the legal advisers in the US Department of Defence,[46] it appears to have been publicly adopted by the UK Government.[47]

We acknowledge this view but do not adopt it, as this understanding of sovereignty does not appear to enjoy broad support.[48] Rather, we proceed on the assumption that the IGE was correct in identifying the principle of sovereignty is an 'operational' rule of international law that has practical normative contents and is capable of being violated separately from any other norm of international law, in particular, the principle of non-intervention and prohibition of the use of force.[49]

When adopting this approach, a second difficulty arises. This concerns the nature and degree of cyber interference in a State that would breach its sovereignty. The IGE agreed that if one State conducts cyber operations against another through an agent that is physically present in the territory of that second State, a violation of sovereignty will have taken place.[50] Thus, for example, if Stuxnet was introduced into the national infrastructure of Iran on a USB key brought into Iran by an agent of another State, as was alleged, then a breach of sovereignty would have occurred. The type of damage or interference caused by Stuxnet generally, or by its autonomous functionality, has no bearing on this conclusion.

Matters become more complicated where a State launches a cyber operation from its own territory but with effects in another State. According to the IGE, whether the principle of sovereignty is breached in such circumstances ought to be assessed on two alternative bases: '(1) the degree of infringement upon the target State's territorial integrity; and (2) whether there has been an interference with or usurpation of inherently governmental functions'.[51] These parameters, by broadly referring to a 'degree of infringement' and 'inherently governmental functions' create something of an interpretative 'grey zone'.[52]

With respect to the first prong of the test that they outlined, the IGE agreed that cyber operations that have physical consequences in another State would violate sovereignty.[53] They also agreed that cyber operations resulting in the loss of functionality of cyber infrastructure would in 'some circumstances' – which circumstances not being entirely clear – amount to a breach of sovereignty.[54]

[45] Ibid 208–9.

[46] Jennifer M O'Connor (General Counsel of the US Department of Defence), 'Memorandum: International Law Framework for Employing Cyber Capabilities in Military Operations' (19 January 2017).

[47] The Rt Hon Jeremy Wright MP, 'Cyber and International Law in the 21st Century' (Speech, Chatham House, London, 23 May 2018) <www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>: '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'.

[48] Cf Watts and Richard (n 43) 863: 'On balance, the *Manual* offers the more persuasive view'.

[49] See also Michael N Schmitt and Liis Vihul, 'Respect for Sovereignty in Cyberspace' (2017) 95 *Texas Law Review* 1639.

[50] *Tallinn Manual 2.0* (n 2) 19.

[51] Ibid 20.

[52] Michael N Schmitt, 'In Defence of Sovereignty in Cyberspace' (*Just Security*, 8 May 2018) <www.justsecurity.org/55876/defence-sovereignty-cyberspace>.

[53] *Tallinn Manual 2.0* (n 2) 20 para 11.

[54] Ibid 20–21 para 13.

There was no agreement, however, as to whether cyber operations falling below the loss of functionality threshold would violate sovereignty.[55] Cyber operations falling into this legally murky category would include those that consist of, for example, the monitoring, exfiltration or modification of data in a State's territory by another State. Doubts have been cast, however, on the IGE's reluctance to view such operations as breaches of sovereignty. In particular, the analogy with physical incursions into a State's territory, which do not necessarily cause any real harm but would nonetheless be seen as breaches of sovereignty, supports a fairly broad reading of the norm. Thus, it has been cogently argued that any cyber operation 'that *penetrates* computer networks and systems supported by cyber infrastructure situated within the territory of another state constitutes a violation of that state's territorial sovereignty, irrespective of whether that operation causes damage or harm'.[56]

As for the second prong of the *Manual*'s test, the IGE agreed that interference in the 'the delivery of social services, the conduct of elections, the collection of taxes, the effective conduct of diplomacy, and the performance of key national defence activities' would amount to an interference with inherently governmental functions, and thus violate sovereignty.[57]

To return to the example of Stuxnet, even if it had been launched against Iran remotely, it would have amounted to a breach of sovereignty under either prong of the test espoused by the IGE: it caused physical damage and it interfered with national defence activities. More interesting is the question about the propagation of Stuxnet beyond Iran and its discovery on computers in other States. Under the IGE's approach, this would not have been a violation of the sovereignty of the other States in the absence of the effects mentioned in the two-prong test. However, if one adopts a lower threshold for breaches of sovereignty, namely that of penetrating national infrastructure, this becomes a live question under the first prong.

The complication that autonomy adds here is that Stuxnet does not appear to have been intended by its authors to infect computers or to cause damage outside Iran. Indeed, the 'infectiveness' of the malware – that is to say, its capacity to propagate – was fairly low. For example, Stuxnet had a self-limitation mechanism, whereby it deleted itself from removable media after a certain number of computers had been infected,[58] but it remains the case that Stuxnet ended up infecting thousands of computers around the world. The question thus arises as to whether the presumptive intention of Stuxnet's authors to limit the effects of the malware to Iran has any significance given that the worm autonomously infected computers in other States.

A similar problem can be raised in relation to the second prong of IGE's test. A hypothetical scenario constructed based on the WannaCry ransomware attack helps to understand the issue. In 2017, WannaCry shut down computers in more than 80 UK National Health Service organisations, resulting in 20,000 cancelled appointments and five hospitals diverting ambulances.[59] If this cyber operation had been launched against the UK by another State, it would plainly have breached UK sovereignty under the second prong because the national delivery of healthcare, an important social service, was seriously disrupted. However, would it matter

[55] Ibid 21 para 14.

[56] Russell Buchan, *Cyber Espionage and International Law* (Hart 2018) 54 (emphasis added).

[57] *Tallinn Manual 2.0* (n 2) 22 para 16.

[58] Falliere, Murchu and Chien (n 33) 29.

[59] Alex Hern, 'WannaCry, Petya, NotPetya: How Ransomware Hit the Big Time in 2017' *The Guardian* (30 December 2017) <www.theguardian.com/technology/2017/dec/30/wannacry-petya-notpetya-ransomware>.

whether WannaCry's authors intended to disrupt the operation of the NHS or whether they were reckless as to that consequence by giving the malware an unlimited ability to propagate?

We submit that intrusion into another State inadvertently or due to unpredictable (or unpredicted) autonomous operation of cyber capabilities has no impact on the legal assessment of the breach of sovereignty. This is because the intention of the author of an action is irrelevant to establishing whether a breach of sovereignty has occurred. Drawing a parallel with physical breaches of sovereignty should help to illustrate the point. Flying a military aircraft of State A without permission into the airspace of State B will result in a breach of the territorial integrity of State B. Whether or not the aircraft was flown into the airspace of State B on the orders of State A, or whether it happened inadvertently due to a navigational error of the pilot or the malfunctioning of the aircraft, has no legal consequence. Certainly, State B may choose to react to the incident differently if it can be established that the breach of airspace was unintended, but this is a political choice about invoking State responsibility, not whether a primary rule of international law was breached.

Likewise, it does not matter whether the aircraft was manned and controlled by a pilot on board, or unmanned and controlled by an operator remotely. Taking this one step further, it does not matter whether the aircraft was manually flown by a pilot at the time of the aerial incursion, or whether it was on autopilot or some other autonomous mode of flight. Again, a State may choose to treat the entry of an unmanned surveillance aircraft into its airspace differently from intrusion by manned fighter jets, but this choice is not dictated by law.

Applying this to cyber operations, the degree of real-time human control exercised over the operation has no impact on whether the operation breaches the sovereignty of another State. That said, autonomous capability seems to be legally significant here in that it might increase the likelihood of sovereignty breaches. While this does not seem to generate novel legal questions, it may increase legal risk. Thus, when relying on cyber capabilities with autonomous functionality, prudent States would need to assess the likelihood and consequences of inadvertent breaches of sovereignty.

This in turn raises the question of whether the risk of breaches of sovereignty could be effectively mitigated by technological means. We have already noted that some of the features of Stuxnet, though not able to prevent the spread of the malware beyond Iran, nonetheless made infections fairly geographically concentrated. More sophisticated technological means could be used to further reduce the risk of sovereignty breaches. In particular, the operation of a cyber capability could be restricted to a particular geographical area by geolocating the computers affected. Admittedly, references to IP addresses, for example, do not provide entirely accurate results. There are several reasons for this, including gaps or errors in IP geolocation datasets, and attempts by users to hide the 'true' IP address of their computers by using VPNs, proxies and relays.[60] That said, at least attempting geolocation might be one way of mitigating legal risk.

Such strategies, however, bring us back to the question of the threshold for a breach of sovereignty. Assume that software used in a cyber operation needs to install itself on the target computer to establish that computer's location by reference to the IP address and other factors,

[60] David Belson, 'Finding Yourself: The Challenges of Accurate IP Geolocation' (*Oracle Dyn Blog*, 29 January 2018) <dyn.com/blog/finding-yourself-the-challenges-of-accurate-ip-geolocation>.

and that this software, having established that it has moved outside its intended area of operation, promptly deletes itself. In such an instance, it is unclear whether the mere replication of the software onto the target system and its actions to establish its location could be described as a penetration of the system, and thus potentially a breach of sovereignty.

3.2 Interventions in internal affairs

The principle of non-intervention is somewhat better defined in international law than the principle of sovereignty, making its application to cyber operations marginally easier.

The *Tallinn Manual* stipulates that '[a] State may not intervene, including by cyber means, in the internal or external affairs of another State'.[61] The meaning of intervention has been defined by the ICJ in the following manner:

> A prohibited intervention must […] be one bearing on matters in which each State is permitted, by the principle of State sovereignty, to decide freely […] Intervention is wrongful when it uses methods of coercion in regard to such choices, which must remain free ones. The element of coercion […] defines, and indeed forms the very essence of, prohibited intervention'.[62]

This articulation has been widely understood, including in the cyber context and by the IGE, as providing for a cumulative two-element test: (a) the conduct must impinge upon certain sovereign prerogatives of a State (the so-called *domaine réservé*); and (b) the conduct must involve coercion.[63] The precise scope of *domaine réservé* depends on the extent of a State's obligations under international law but would typically include choices relating to its political system and its organisation, the development of foreign policy, and so on.[64] Coercion, on the other hand, 'refers to [any] affirmative act designed to deprive another State of its freedom of choice, that is, to force that State to act in an involuntary manner or involuntarily refrain from acting in a particular way'. [65] For example, a cyber operation designed to alter the results of a referendum or an election, thereby resulting in the passage of legislation not supported by the public or the installation of an elected official who did not win the vote, would constitute intervention.

The majority of the IGE took the view that 'the coercive effort must be *designed* to influence outcomes in, or conduct with respect to, a matter reserved to the target State'.[66] At the same time, the fact that it 'fails to produce the desired outcome has no bearing on whether [the principle of non-intervention] has been breached'.[67] Thus, coercive intent plays a key role. Indeed, the IGE concurred that intent 'is a further constitutive element of a violation of the prohibition of intervention'. [68]

[61] *Tallinn Manual 2.0* (n 2) rule 66.

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

[63] *Tallinn Manual 2.0* (n 2) 314 et seq.

[64] Ibid 314–317.

[65] Ibid 317.

[66] Ibid 318 (emphasis added).

[67] Ibid 322.

[68] Ibid 321.

This has important implications for cyber capabilities with autonomous functionality. The use of such capabilities can result in a prohibited intervention only where the State acted with coercive intent, that is to say, it designed or deployed the capability so as to coerce another State. If, however, the autonomous functioning of a cyber capability inadvertently led to a situation where another State felt compelled to change its course of action with respect to something falling within the *domaine réservé*, there has been no violation of the non-intervention principle. Briefly put, it does not seem possible to intervene in another State's affairs by accident or through an unforeseen maloperation of an autonomous functionality.

The question does arise, however, how direct the coercive intent must be. What if a State deploys a cyber capability without intending to coerce another State, but knowing that the other State would feel coerced? What if there is only a likelihood that the other State would feel coerced? Current international law does not allow us to give definitive answers. However, assuming that autonomous functionality tends to increase the inadvertent but not entirely unforeseeable effects of cyber operations, answers to these questions need to be proposed.

4 Law on the use of force (*jus ad bellum*)

Autonomous cyber capabilities that go beyond passive defence measures may raise issues of *jus ad bellum* if the consequences of the measures reach the level of use of force or armed attack. These measures may include some active cyber defence and offensive cyber operations. As stated in the *Tallinn Manual*, the question of whether a cyber operation constitutes a use of force or an armed attack depends on the operation's scale and effects.[69] The consequences of the cyber operation in question should be comparable to those of a kinetic use of force or armed attack. This covers cyber operations that have physical consequences – injury or death to personnel or damage or destruction of property. The gravest forms of use of force constitute an armed attack, which triggers the victim State's right of self-defence.

These elements also apply if a cyber operation is carried out using autonomous cyber capabilities, but this may raise additional legal questions. This part of the paper attempts to map these questions.

4.1 Automatic hack-backs

Most discussed in this context are so-called automatic hack-back defensive measures. That would mean software that, without real-time human input, detects an intrusion, identifies its origin and acts outside its own system to stop it and possibly to harm the system where the intrusion originated from. From a legal perspective, it is irrelevant whether such software would be qualified as defensive or offensive cyber capability. What counts as a matter of law is the consequences of such an operation for the affected external systems, assuming they are in another State.

If we take the example of software like Mayhem described above, it is the part of it that finds and exploits vulnerabilities in the systems of others that may be of concern in this context. As usual, the devil is in the detail.

The legal concerns of automatic hack-back relate to its unforeseen consequences and to attribution. Software is becoming better at detecting abnormalities and vulnerabilities in the system that it is protecting. The sheer amount of data to be monitored for such detection is already far beyond human abilities, and the speed at which things happen limits what humans can do. Therefore, it is only to be expected that cybersecurity experts are looking for ways to automate this process and make it more efficient. For the same reasons, and since detection is only the first step in countering an intrusion, software developers are also looking for ways for computer programs to take measures against intrusions autonomously. Depending on the type of malware used, the attack vector and the vulnerabilities exploited, it may prove to be most efficient to take external action to stop the intrusion. A simple example would be taking down a command and control (C2) server of a botnet or the whole botnet to stop a distributed denial-of-service (DDoS) operation against one's systems. A more elaborate example would be destroying a server to which cyber espionage malware is sending data extracted from your system.

[69] *Tallinn Manual 2.0* (n 2) rules 69 and 71.

An even more critical example is a counterstrike against a system that seems to be the source of a cyber operation against critical infrastructure. In this case, the victim State may be entitled to take countermeasures against the State conducting this operation (assuming there is attribution) or act under the plea of necessity. However, even if we assume that there is a legal right to respond by taking destructive action against the assumed source of the incoming operation (which takes time to determine), there is a risk of exceeding the legal limitations of the response.

As the predictability of autonomous systems is one of the main challenges, any cyber capability autonomously executing such measures risks causing unforeseen effects. If those effects entail damage rising to the level of use of force, then the State operating the autonomous cyber capability has violated the prohibition on the use of force, and should these effects reach the threshold of an armed attack, then the other State has the right to respond with force in self-defence.

Unforeseen consequences may be due to malfunction of the system caused by a technical failure, the program disobeying the operator (most relevant in AI-based systems) or external manipulation. Finding the exact cause may be relevant for individual accountability under national law; but in terms of international law, the State operating the autonomous system is responsible for its actions.

Another legal concern is attribution. The autonomous cyber capability may well be able to establish immediate technical attribution of the incoming cyber operation, but due to using various techniques of obfuscation such as spoofing it may not be the real source of the operation. Should this happen in combination with the unforeseen consequences, the State operating the autonomous cyber capability risks using force against a third State that has been made to look like the originator of a malicious cyber operation.

These concerns can partly be mitigated by setting limits to what the autonomous cyber capability can do in the course of an automatic hack-back. It has been suggested, for example, that autonomous cyber capabilities should be limited to effects not likely to raise use of force concerns. According to Stuard and McGhee, this could be achieved by:

> ‘(1) blocking connections to our networks; (2) gathering information from an intruding machine and machines associated or laterally connected to the intrusion; (3) lacking the ability to influence the root level software or hardware of target machines so as not to threaten the overall operation of hardware or entire systems; (4) being reversible in nature in that they do not require complete reformatting of a system or replacement of hardware’.[70]

However, even if those rules worked and were followed in technical execution so that the consequences of an automatic hack-back remained below the threshold of use of force, one would still risk violating other norms of international law, which might trigger a response in the form of countermeasures, retorsions or measures under the plea of necessity.[71]

[70] Jarrod H Stuard and James McGhee, ‘Is Skynet the Answer? Rules for Autonomous Cyber Response Capabilities’ in Misty Blowers (ed), *Evolution of Cyber Technologies and Operations to 2035* (Springer 2015) 159.

[71] ‘Articles on Responsibility of States for Internationally Wrongful Acts’ (2001) UN Doc A/56/10 (ARSIWA) ch II and art 25.

Should the automatic hack-back happen in response to a cyber operation that itself constitutes a use of force or an armed attack, the legal questions relate to the assessment of the initial incoming cyber operation and the limitations on the response. If the incoming cyber operation reaches the threshold of an armed attack,[72] the State may exercise its inherent right of self-defence. Assessing whether the scale and effects of the cyber operation are grave enough to consider it an armed attack is a political decision taken in the framework of international law. The decision is not made based only on technical information, but also after assessing the strategic context and the effect of the cyber operation beyond cyberspace. Assessing whether physical damage to property or injury to people has been caused and whether those consequences are grave enough to go beyond the mere use of force and amount to an armed attack seems too complex for current technology. Therefore, for the foreseeable future, it remains a human decision. However, this political decision could be implemented through technical criteria, the fulfilment of which would under any circumstances constitute an armed attack. That would mean pre-defining a red line in technical terms which would leave no doubt as to the gravity of the impact, and inserting this pre-definition in the autonomous cyber capability. These criteria could relate to the damage caused to the cyber infrastructure that an autonomous cyber capability is designed to protect. Whether such pre-definition of a red line on such a sensitive issue is reasonable remains for political leaders to decide. In any case, it does not deprive them of the possibility of making the assessment in any given situation by themselves.

Ultimately it is a matter of policy to decide how much risk is acceptable for a State when weighing the gains and potential losses of employing such capabilities. It is important for States to make these decisions in an informed way, being aware of the benefits and potential consequences.

Having made such a choice, the next step is to make sure that the response in the form of automatic hack-back stays within the boundaries of what is permitted when exercising the right to self-defence. A relatively safe option would be to limit the automatic hack-back to intelligence gathering only and not the use of force. Should the State choose to have the automatic use of force response option available, it must make sure that the force used is necessary attack and proportionate, using only the amount of force required to defeat the armed attack.[73] Both are context-dependent and do not have to be limited only to cyberspace. Therefore, the autonomous cyber capability should be able to determine whether passive cyber defence measures or external measures below the use of force are sufficient to terminate the cyber armed attack, or whether human input should sufficiently clearly predetermine which are the situations where non-forceful measures would clearly be insufficient. In any case, human supervision should always be exercised.

When developing an autonomous cyber capability that is able to conduct automatic hack-back in response to an armed attack, the State should also make sure that the effects of the force used in the course of the hack-back are limited to the systems of the culprit State and are able to comply with the requirements of the law of armed conflict (see Part 5 of this paper). Any effects on a third State would entail State responsibility *vis-à-vis* those States.

[72] *Tallinn Manual 2.0* (n 2) rule 71.

[73] Ibid 348–9 rule 72 and commentary.

4.2 Autonomous offensive cyber capabilities

Legal issues regarding *jus ad bellum* may also rise in relation to outright offensive autonomous cyber capabilities. If such a capability qualifies as a means or method of warfare, its use in an armed conflict is governed by international humanitarian law (IHL) (see Part 5 of this paper) but outside armed conflict States may be tempted to use autonomous cyber capabilities that not only function as a means of intelligence gathering but also have broader effects on the target system. If these do not rise to the level of use of force or armed attack, they may still trigger State responsibility under other rules of international law (see Part 3). If they do so even unintentionally, the State risks violating the prohibition of the use of force or even conducting an armed attack.

Another aspect of using autonomous offensive cyber capabilities is in the situation where a State decides to act in response to an armed attack, cyber or otherwise, and to use offensive cyber means for exercising its right to self-defence, either alone or along with conventional force. In this case, the State would have to make sure that its autonomous cyber capability acts within the boundaries of self-defence described above – necessity, proportionality, attribution and containment of effects. Since the decision to use the capability, the more immediate concerns relate to proportionality and containing the effects of the operation to the culprit of the armed attack. That capability would have to be programmed and configured in such a way that the consequences do not exceed the force necessary to terminate the armed attack, and do not go beyond the systems of the State that conducted the armed attack.

To be able to conduct such a response a State may, in addition to gathering cyber intelligence on the potential target systems, want to insert dormant malware into the target systems. This may raise the question of when the response operation actually began. Also, if the target State discovers this dormant malware in its systems and establishes its origin, it may be tempted to consider itself the target of an imminent armed attack and exercise anticipatory self-defence. This, of course, depends on what the malware is able to do and how much of it is understandable for the target State once it discovers the malware's ability to cause real damage. Should this happen during a period of heightened tensions between the States in question, it may well lead to the most serious questions regarding national security.

Many have argued that the use of AWS would make it easier for States to resort to the use of force.[74] This concern applies also to cyberspace and the use of autonomous cyber capabilities, aggravated by the fact that, especially in cyberspace, escalation happens at machine speeds and even if there are human operators supervising the systems they may not be able to react fast enough to avoid unwanted consequences. Although it is a valid policy, technical, operational, ethical and philosophical concerns remain to be addressed at the political level. In a purely legal sense, if States comply with their legal obligations in the development and use of autonomous cyber capabilities, their conduct is lawful unless they agree on new specific rules in this regard.

[74] Human Rights Watch, *Losing Humanity* (n 5) 4, 39; Heyns (n 5) para 58; Philip Alston, 'Interim Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions' (23 August 2010) UN Doc A/65/32 para 44.

5 International humanitarian law (*jus in bello*)

5.1 Applicability of the law

The applicability of IHL appears to be one of the few issues on which there is more agreement in relation to AWS than cyber capabilities. According to the mainstream position, IHL governs cyber activities undertaken during an armed conflict, a position reflected in the *Tallinn Manual*[75] and confirmed by the opinion of many States.[76] However, a handful of States either reject this view or are at least ambivalent as to whether they accept it.[77] As for AWS, however, States expressly agree that IHL applies 'fully' to their development and use.[78] This is unsurprising, as AWS are plainly a category of weapon systems, the use of which in armed conflict falls squarely within the scope of IHL.

Any uncertainties about the application of IHL to cyber capabilities also become relevant to autonomous cyber capabilities; in particular, it is not entirely clear when a cyber operation amounts to an attack. The *Tallinn Manual* treats as an attack any cyber operation 'that is reasonably expected to cause injury or death to persons or damage or destruction to objects'.[79] In line with this premise, the IGE considered cyber capabilities that 'are used, designed, or intended to be used' to bring about such consequences to amount to means of warfare.[80] This approach remains contentious insofar as it fails to designate as an attack cyber operations that, while capable of causing potentially large-scale adverse consequences, cannot be reasonably expected to cause physical damage or loss of function.[81] We do not intend to enter this debate here; we merely note that under a more liberal interpretation of the notion of attack, the principles of law applicable to attacks would have broader application to cyber activities. Furthermore, if one accepts the broad view that the destruction of data may constitute an attack, the notion of means or method of warfare needs to be adjusted and taken into account when considering autonomous cyber capabilities.

In the context of kinetic AWS, the most challenging questions relate to autonomy in what the ICRC has referred to as the 'critical functions' – the capacity of the system to select and engage targets. Numerous concerns have been raised regarding accountability for the use of such systems, the risk of proliferation and fundamental ethical principles. With regard to the substantive rules of IHL, the question has been raised in particular of whether autonomy in critical functions can be implemented and used consistently with the core principles of distinction and proportionality, and the obligation to take precautionary measures during the attack. In this Part of the paper, we will focus on those principles.

[75] *Tallinn Manual 2.0* (n 2) rule 80.

[76] 2015 UN GGE Report (n 1) para 28(d), where the UN GGE 'notes the established international legal principles, including, where applicable, the principles of humanity, necessity, proportionality and distinction'.

[77] See, eg, Michael N Schmitt and Liis Vihul, 'International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms' (*Just Security*, 30 June 2017) <www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms>.

[78] 2018 CCW GGE Report (n 7) para 21(a).

[79] *Tallinn Manual 2.0* (n 2) rule 92.

[80] Ibid 452 para 2.

[81] Ibid 418 para 13.

The legal issues of cyber and autonomy plainly interact and multiply in this context, as something could constitute both a cyber means of warfare and an AWS depending on the perspective. Indeed, on a certain level of abstraction, the difference between AWS and autonomous cyber capabilities disappears completely. For example, a counter rocket, artillery and mortar system (C-RAM) will, once installed, configured and activated, detect incoming threats. It will respond to those threats autonomously by controlling a mechanical actuator – a quick-firing gun. An autonomous cyber capability that causes physical harms generally turns some industrial device (centrifuge, dam, power plant, etc.) into an *ad hoc* actuator. Though the design of autonomous cyber capabilities and AWS is different, the relationship between the software and the hardware can be quite similar.

Earlier in this paper we have used Stuxnet as an example of an offensive autonomous cyber capability. We can also use it here, particularly as it provides a paradigmatic example of a cyber capability that was designed and used to bring about physical damage. Thus, if Stuxnet were to be deployed in a pre-existing armed conflict or triggering an armed conflict, its use would have to be assessed under IHL on the conceptualisation of attacks found in the *Tallinn Manual*. Note also that while the propagation of Stuxnet raised questions in relation to the principle of sovereignty, even under the most liberal reading of the concept of attack such propagation would be unlikely to be governed by IHL rules relating to attacks, as no destruction or corruption of data occurred in the infected systems. At the same time, such propagation would have been subject to the geographical limitations imposed by IHL,[82] raising issues similar to those considered in Part 3.

5.2 Principle of distinction

The principle of distinction requires that parties to a conflict distinguish at all times between civilians and combatants, and between civilian objects and military objectives.[83] The *Tallinn Manual* confirms that the principle of distinction also applies to cyber attacks.[84] The principle of distinction has two main implications. First, it is prohibited to employ means and methods of warfare that are indiscriminate by nature.[85] Second, it is prohibited to make the civilian population, individual civilians or civilian objects the object of a cyber attack.[86] Again, the *Tallinn Manual* confirms that these rules apply to cyber means of warfare and to cyber attacks.[87]

Some have taken the view that AWS are either inherently indiscriminate or at the very least have serious difficulty being deployed in a discriminate manner. There are two central planks to this argument. One suggests that technology remains inadequate and unreliable for positively identifying lawful objectives, such that civilians and civilian objects can be inadvertently targeted. The second concern is that determining what constitutes a lawful military objective requires subjective evaluative judgments which AWS are fundamentally incapable of making. In particular, the

[82] See *Tallinn Manual 2.0* (n 2) rule 81.

[83] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (8 June 1977, in force 7 December 1978) 1125 UNTS 3 (AP I) art 48.

[84] *Tallinn Manual 2.0* (n 2) rule 93.

[85] AP I art 51(4)(b)–(c).

[86] AP I art 51(2) and 52.

[87] *Tallinn Manual 2.0* (n 2) rules 105, 94 and 99.

argument goes, an AWS would not be able to assess whether a civilian is taking a direct part in hostilities[88] or whether an object satisfies the two-prong test of military objectives.[89]

These arguments presume, however, that States and commanders would want to delegate to an AWS the full discretion to the use of force that a combatant would have as a matter of law. In such a scenario, the AWS would be permitted to target anything that qualifies as a military objective but would be unable to determine whether someone or something is a military objective, especially in a borderline case. This assumption overlooks the fact that the ability of combatants to conduct attacks is generally further regulated and restricted by rules of engagement. Rules of engagement might only permit the attacking of particular types or categories of persons or objects. A similar approach could be taken in relation to AWS so that an AWS is authorised to attack only certain kinds of military objective, especially those qualifying as military objects by their nature (for example, weapons), which the AWS can reliably identify. Put in broader terms, commanders will likely only entrust to AWS a role that they determine the AWS is capable of carrying out consistent with their intent, the requirements of law and national policy.[90]

To ensure compliance with the law, the designers of an AWS or an autonomous cyber capability need not make the system capable of understanding or applying the law. Rather, they must ensure that the features of the system in the particular operational environment will only produce lawful outcomes. The operation of Stuxnet provides an excellent example of that. Stuxnet had no ability to distinguish between military and civilian objectives and to operate only in relation to the former. Rather, it was programmed to target a particular type of system which had been predetermined by human operators to be a lawful target. The same design principle could be applicable more broadly: autonomous cyber capabilities could be designed to neutralise air defence systems or command and control capabilities. The ability of the autonomous cyber capability to identify such systems in conjunction with the operator's determination that these types of systems constitute military objectives seems likely to satisfy the requirements of the principle of distinction.[91]

From a technical perspective, identifying targets by means of technical criteria poses slightly different problems in the cyber context. One the one hand, the cyber context mitigates some of the difficulties likely to be experienced with AWS. For example, an AWS configured to attack tanks would probably need to rely on electro-optical sensors and image recognition software for target identification. The ability of an AWS to complete this task would depend, *inter alia*, on the capabilities of the sensor (sensitivity, spatial and temporal resolution etc.) and a myriad of environmental condition (such as light levels, precipitation, and presence of clouds or fog). This problem would not arise in relation to cyber capabilities. Thus, reliance on purely technical

[88] Ibid rule 97.

[89] Ibid rule 100: 'Military objectives are those objects which by their nature, location, purpose, or use, make an effective contribution to military action and whose total or partial destruction, capture or neutralisation, in the circumstances ruling at the time, offers a definite military advantage'.

[90] Cf Michael N Schmitt and Jeffrey S Thurnher, '"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict' (2013) 4 *Harvard National Security Journal* 231, 241.

[91] Cf 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 Weapons Reviews' (2012) 94 *International Review of the Red Cross* 483, 492; Marco Sassòli, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified' (2014) 90 *International Law Studies* 308, 327.

means of target identification may be more workable in cyberspace. At the same time, various obfuscation techniques, such as spoofing, may cause difficulties in this regard.

5.3 Principle of proportionality

The principle of proportionality prohibits the launching of an attack that may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated.[92] According to the *Tallinn Manual*, this principle also applies in relation to cyber attacks.[93] Compliance with the principle of proportionality has been a major source of concern with both cyber capabilities and AWS, but for slightly different reasons. With cyber attacks, the collateral damage that they cause, especially any reverberating effect, may be difficult to predict, quantify and evaluate.[94] An AWS, by contrast, may be able to predict and quantify the extent of collateral damage – perhaps even more accurately than a combatant since combatants already use software tools for collateral damage estimates.[95] However, an AWS may be incapable of assessing the concrete and direct military advantage contemplated and balancing that against the amount of collateral damage; this appears to be a decision requiring human judgment.[96] Any benefits that would arise from a computerised system's ability to assess the extent of the damage might, therefore, be negated by the difficulty of determining the scope of collateral damage in case of cyber operations.

One way around this problem would be to use autonomous targeting functionality only in circumstances where the presence of civilians or civilian objects, or an adverse effect on them, can be excluded ahead of time.[97] Another option would be for the designers or operators to establish levels of acceptable collateral damage in advance and introduce them into the system as technically describable restrictions.[98] The adequacy of these approaches depends heavily on the nature of the cyber capability in question. The problem seems to be more easily solvable in cyber capabilities which are designed with a specific operation in mind – again, Stuxnet provides an example of this, as collateral damage was likely to negligible. The issue becomes more acute in circumstances where a cyber capability is intended to operate for an extended period and in complex environments, and thus will be confronted with a range of different operational circumstances.

[92] AP I art 57(2)(iii).

[93] *Tallinn Manual 2.0* (n 2) rule 113.

[94] Cf ibid 475 para 13.

[95] Schmitt and Thurnher (n 90) 254–255.

[96] Sassòli (n 91) 331–2 (considering this the most serious legal issue in relation to AWS).

[97] Ian S Henderson, Patrick Keane and Josh Liddy, 'Remote and Autonomous Warfare Systems: Precautions in Attack and Individual Accountability' in Jens Ohlin (ed), *Research Handbook on Remote Warfare* (Edward Elgar 2017) 351–352.

[98] Schmitt and Thurnher (n 90) 256.

5.4 Precautionary measures

Those who plan or decide on attacks must take feasible precautionary measures to ensure that the intended targets are lawful, that collateral damage is minimised through the choice of means and methods, and that the anticipated collateral damage is not excessive.[99] Those who plan, decide upon or execute attacks must cancel or suspend them if it becomes apparent that the attack would not comply with the principle of distinction or proportionality.[100] The *Tallinn Manual* confirms the application of these rules to cyber attacks.[101]

It has occasionally been suggested that if AWS have the ability to select and engage targets without human intervention, then the obligation to take precautionary measures would fall on the system. This claim is based on a misunderstanding of the obligation, and indeed the nature and objective of legal regulation generally. Precautionary measures must be taken by people who plan, authorise or execute attacks; they are not functions that can be delegated to a weapon system or autonomous cyber capability. Such systems may have safeguards in place that prevent those systems from operating inconsistently with the requirements of IHL, especially in relation to distinction and proportionality. Assessing the effectiveness and sufficiency of such safeguards forms part of the duty to take precautionary measures, but the activation of such safeguard by the system should not be confused with the underlying duty itself.

Taking precautionary measures is more straightforward when autonomous cyber capabilities have been devised to carry out a specific attack, and a human operator has a degree of control over the time and the circumstances of the attack. Matters become more complicated when, for example, a cyber capability can, without real-time human intervention, react to offensive cyber operations by the adversary by initiating actions that cause physical harm. In those instances, the operator clearly cannot take precautionary measures with respect to each possible action. Thus, the deployment of the capability would, from a legal perspective, be seen as a decision to launch the attack. This situation is comparable, for example, to laying a mine. Precautionary measures would have to be taken when the capability is deployed and address all possible actions. We submit that there are good reasons to be cautious here: the taking of precautionary measures with regard to autonomous capabilities remains a debatable issue and technically a complex one.

[99] AP I art 57(2)(a).

[100] AP I art 57(2)(b).

[101] *Tallinn Manual 2.0* (n 2) rules 115–117 and 119.

5.5 Martens clause

No discussion of IHL is complete without a mention of the Martens clause. Introduced in the preamble of the 1899 Hague Convention II and repeated in many subsequent IHL instruments, the contemporary formulation of the clause provides as follows:

> 'In cases not covered by this Protocol [i.e. AP I to the Geneva Conventions] 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 the principles of humanity and from the dictates of public conscience'.[102]

The argument has been made that this clause prohibits AWS.[103] In particular, it is argued that AWS 'face significant obstacles' in complying with the principles of humanity '[d]ue to their lack of emotion and legal and ethical judgment'; also, AWS would be inconsistent with dictates of public conscience as '[m]any individuals, experts, and governments have objected strongly to the development of fully autonomous weapons'.[104]

We do not go into the merits of these arguments, other than to note that they are based on a reading of the Martens clause that is not universally shared. Indeed, there is no generally shared understanding of the clause.[105] For example, the *Tallinn Manual* simply acknowledges that the Martens clause functions to ensure cyber activities that occur in the context of an armed conflict 'are not conducted in a legal vacuum'.[106] We merely note that the arguments are likely to be less persuasive in relation to autonomous cyber capabilities, which are essentially anti-materiel capabilities – it is not possible to directly harm a human being by cyber means. That said, we acknowledge that cyber capabilities can be deployed to cause harmful effects on humans (for example, through the manipulation of pacemakers or other medical devices) or may have incidental harmful effects on civilians. Thus, the Martens clause might be more relevant to methods of warfare that rely on autonomous cyber capabilities, rather than the autonomous characteristics of the capabilities themselves.

[102] AP I art 1(2).

[103] Human Rights Watch, *Losing Humanity* (n 5) 30; Human Rights Watch, *Heed the Call: A Moral and Legal Imperative to Ban Killer Robots* (2018).

[104] Human Rights Watch, *Heed the Call* (n 103) 2.

[105] See, eg, Rupert Ticehurst, 'The Martens Clause and the Laws of Armed Conflict' (1997) 37 *International Review of the Red Cross* 125.

[106] *Tallinn Manual 2.0* (n 2) 378 para 12; cf Erki Kodar, 'Applying the Law of Armed Conflict to Cyber Attacks: From the Martens Clause to Additional Protocol I' in Rain Liivoja and Andres Saumets (eds), *The Law of Armed Conflict: Historical and Contemporary Perspectives* (Tartu University Press 2012) 109–111.

6 Responsibility and liability under international law

When it comes to autonomous systems generally and AWS in particular, responsibility and liability are among the most sensationalised and complex issues. It has sometimes been suggested that autonomous systems should be granted some form of legal personality.[107] In international law, however, there is neither State practice nor evidence of *opinio juris* to draw such conclusions and treat technological systems (whether or not using AI) as legal subjects of any kind. Furthermore, in the current and near-future technological backdrop, it is still possible to pinpoint the moment in the chain of causality when a human actor activated an autonomous system, whether or not aware of the accompanying risks.[108] Moreover, States have agreed in relation to AWS that '[h]uman responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines.'[109] They have also agreed that, '[a]ccountability for developing, deploying and using any emerging weapons system […] must be ensured in accordance with applicable international law'.[110] These principles seem to be fully applicable to autonomous cyber capabilities. For these reasons, the following discussion leaves aside any theory that rests on the premise that software or a device could constitute a person or some other type of entity capable of bearing responsibility. Accordingly, this part of the paper attempts to apply the doctrine of State and individual responsibility to the use of autonomous cyber capabilities.

6.1 State responsibility

Every internationally wrongful act of a State entails the international responsibility of that State.[111] Whether in cyber or any other domain and regardless of whether autonomous capabilities are involved, an internationally wrongful act occurs when an action or omission is attributable to a State and constitutes a breach of an international obligation of that State.[112]

Notably, these two criteria do not bind State responsibility to the consequences of an act or omission unless a special secondary norm requires a specific consequence for responsibility to be incurred. In other words, conduct violating an international obligation normally suffices to give rise to the responsibility of a State. Moreover, the doctrine of State responsibility does not require a specific form of intent in order for an act or omission to constitute an internationally wrongful act. Rather, State responsibility defers to the rule giving rise to the primary international obligation. For example, as discussed earlier in Section 3.2, coercive intent is an element

[107] Marshal S Willick, 'Constitutional Law and Artificial Intelligence: The Potential Legal Recognition of Computers as "Persons"', *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (1985); Gunther Teubner, 'Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law' (2006) 33 *Journal of Law & Society* 497.

[108] Schmitt and Thurnher (n 90).

[109] 2018 CCW GGE Report (n 7) para 21(b).

[110] Ibid para 21(c).

[111] ARSIWA (n 71) art 1.

[112] Ibid art 2.

of intervention; accordingly, such intent would need to be proven to invoke State responsibility for an act of intervention.

The second element of an internationally wrongful act – attribution to a State – has been long viewed as one the most difficult riddles or even the 'final frontier' of international law as applied to cyber operations. The increasing use of autonomous capabilities could exacerbate the problem. But this is more likely to be due to procedural, technical and political issues rather than doubts about the applicability of the customary international law of State responsibility to cyber or AI-powered operations. Therefore, rather than ask whether State responsibility applies, it would be important to investigate how the employment of autonomous cyber capabilities may lead to an internationally wrongful act for which a State can be held responsible. An issue even more worthy of research is how to prevent States from committing such internationally wrongful acts. As with cyber means in general, autonomous cyber defence systems can serve as tools for violating another State's sovereignty, interfering coercively in its internal affairs, using military force or conducting an armed attack (see Parts 3 and 4).

States are responsible for the conduct of their organs, which includes any person or entity that has that status according to the State's internal law.[113] State organs in the autonomous cyber capabilities context might include the officials responsible for procuring or reviewing cyber capabilities, engineers, designers and programmers creating the capabilities, and commanders making the decision to use a specific capability in a specific situation. The list is far from exhaustive and, depending on the scenario and administrative structure of the State in question, could extend from a law enforcement officer to the President. Importantly, attribution only occurs when the entity in question acts in the official capacity, and this includes *ultra vires* conduct.[114]

State responsibility also obtains in cases where an essentially governmental function has been delegated to a non-State entity. [115] The *Tallinn Manual* gives an example where a State lacks the capability to engage in sufficiently robust cyber defences of its governmental infrastructure and authorises a private company to defend State networks by employing passive defence measures. During an incident involving malicious cyber operations against the networks, the company engages in active cyber defence by hacking back.[116] The hack-back would be attributable to the State.

Consider a variant of this scenario. The company, contracted by State A, decides to exercise its delegated authority by using an autonomous machine-learning-based software agent capable of detecting various types of threats and identifying the origin of the potential immediate attack. The software is designed to allow for self-learning and autonomous decision-making with respect to the gravity of the threats. It is also capable of assessing the necessity for response and picking the most suitable form of counter-operation. As an unforeseeable result,

[113] ARSIWA (n 71) art 4(1): 'The conduct of any State organ shall be considered an act of that State under international law, whether the organ exercises legislative, executive, judicial or any other functions, whatever position it holds in the organization of the State, and whatever its character as an organ of the central Government or of a territorial unit of the State'.

[114] Ibid art 7.

[115] Ibid art 5.

[116] *Tallinn Manual 2.0* (n 2) 90.

the software ends up harming the functionality of a hijacked intruder system, which is connected to the civilian communications network of State B, rendering regular means of communication temporarily inaccessible for the civilian population of State B.

The crossing of the line from defence to offence, though occurring autonomously, is still attributable to State A. If the company had received orders from State A to engage exclusively in passive defence or it had been given instructions which had precluded the use of autonomous or unpredictable technological solutions, it would have acted *ultra vires*. But *ultra vires* conduct of the company, as an entity empowered to exercise elements of the governmental authority, is still attributable to State A.[117] Any other rule would facilitate misconduct due to the difficulty faced by State B in proving what orders the company actually received. A more complex situation would develop if the company used the same or a similar autonomous agent to pursue its own goals unrelated to the national security or public functions of State A. The complexity here is again more related to forensics and less to whether or not the delegated capacity has been exceeded or to the degree of autonomy.

A different problem arises when the wrongful conduct is carried out by non-State actors that have not been formally empowered to exercise governmental authority. The position of such non-State actors in cyber conflict has gained attention in media and scholarly literature, particularly since operating via non-State proxies seems to be the default *modus operandi* of States with more aggressive cyber strategies.[118] As a general matter, the conduct of a non-State actor becomes attributable to a State where the non-State actor is in fact acting on the instructions of, or under the direction or control of that State in carrying out the conduct.[119] Accordingly, as explained in the *Tallinn Manual,* 'cyber operations of a non-State actor are attributable to a State if the State factually exercises "effective control" over that specific conduct of the non-State actor'.[120] In practical terms, a State is in 'effective control' of a particular cyber operation by a non-State actor when it is the State that determines the execution and course of the operation, and the cyber activity engaged in by the non-State actor is an 'integral part of that operation'. [121] The *Tallinn Manual* provides a hypothetical case where

> 'a State plans and oversees an operation to use software updates to implant new vulnerabilities in software widely used by another State in its governmental computers. The former State concludes a confidential contract to embed the exploits with the company that produces the software and then directs the process of doing so. Such being the case, the company's behaviour is attributable to the controlling State'. [122]

[117] ARSIWA (n 71) art 7.

[118] Tim Maurer, *Cyber Mercenaries: The State, Hackers, and Power* (Cambridge University Press 2018).

[119] ARSIWA (n 71) art 8.

[120] *Tallinn Manual 2.0* (n 2) 96.

[121] See ARSIWA (n 71) commentary to art 8, para 3.

[122] *Tallinn Manual 2.0* (n 2) 97.

It is easy to insert an element of autonomy by imagining a different scenario where a State instructs a company, hacktivist group or individual to create a capability that finds and exploits vulnerabilities in a specific software widely in use in another State. The company, group or individual chooses to deploy an autonomous cyber capability for this task.

Adding some autonomy to the mix will not change the outcome of the legal analysis. Questions arise, however, in relation to the limits of the *de facto* control a State is able and obliged to exercise over a system in which the exact ways of operating might not always be foreseeable. Logically one might assume that while operating a highly complex system, the traditional notions of control tend to waver, especially for the actors not directly involved in the development of the system. However, hiding behind such notions would ultimately lead to a never-ending loop of irresponsibility, where proxies will have proxies that have proxies. With great caution, the IGE concluded that *ultra vires* acts of non-State actors are generally not attributable to the State.[123] It elaborated that the application of this general principle can prove highly complex and each case must be assessed on its own merits.[124] The *Tallinn Manual* is rather restrictive in mapping the scope of the aforementioned general principle. Could the employment of an autonomous agent, as described in the foregoing scenario, count as acting *ultra vires*? First, we should identify the exact mission which the *ultra vires* acts would have to be incidental to in order to invoke State responsibility. Secondly, it should be asked if the creation of the autonomous system serves the mission. In the case given above, it is apparent that the solution served the State's purpose and is thus, while *ultra vires*, incidental to the mission and consequently still attributable to the State. Even when the non-State group had been specifically instructed against the use of autonomous features, the operation would still be attributed to the State.

In sum, while the doctrine of State responsibility faces challenges when internationally wrongful acts are conducted using an autonomous cyber capabilities, these challenges do not derive from an ontological incompatibility between the existing law of State responsibility and the use of autonomous cyber capabilities. Emerging *opinio juris* lends some plausibility to this conclusion. In 2017, during the third CCW GGE, the United States submitted a working paper in which it agreed that States are responsible for the uses of weapons with autonomous features by members of their armed forces and other acts that may be attributed to the State.[125] The misuse or malfunctioning of an autonomous system results therefore in State responsibility.[126] If a system were to perform on a level of autonomy where instructions from an operator were not required, the State would remain responsible for the activity of its organ or empowered entity.

[123] *Tallinn Manual 2.0* (n 2) 97

[124] United States of America, 'Autonomy in Weapon Systems' (10 November 2017) UN Doc CCW/GGE.1/2017/WP.6 paras 24–31.

[125] Ibid para 26.

[126] Ibid.

6.2 International criminal responsibility

The 2017 US working paper also noted that '[t]he responsibilities of any particular individual belonging to a State or a party to the conflict may depend on that person's role in the organisation or military operations'; usually this means that the commander who makes the decision to use the system, whether or not informed or aware of the actual autonomous features and the technical functioning thereof, is held responsible.[127] The debates around AWS and, for the purposes of this part, especially the consideration of issues relating to responsibility, accountability and attribution under international criminal law (ICL) are as relevant in the context of autonomous intelligent agents operating in cyberspace as they are for autonomous kinetic weapons systems.

With regard to AWS, questions about the distribution of accountability between the operator and the developer(s) of a system have been raised.[128] With some primitive AWS (such as a landmine), as with manually operated systems (such as a rifle), the assumption is that the operator, not the designer or manufacturer of the system, has full responsibility to ensure that the use of the system complies with applicable international law. With more sophisticated AWS (such as a loitering or a sensor-fused munition, where important aspects of the functioning of the system are beyond the control of the operator), it can be asked whether responsibility becomes reallocated by operation of existing law, or whether the existing legal framework should be adjusted to a new technological reality, such that the designer or manufacturer bears a more significant portion of the responsibility.[129] The same problem emerges with respect to 'off-the-shelf' cyber capabilities, for example commercially available cyber defence systems, where the developer provides a product to the end user who can essentially use the system without further input from the manufacturer. However, other cyber capabilities, particularly offensive cyber capabilities, are less likely to be acquired off the shelf. In many, if not most, instances, the developers of the capability are also responsible for deploying the capability, hence becoming the operators. Thus, the issue of allocation of responsibility is less acute with respect to autonomous cyber capabilities compared to physical autonomous systems. Regardless of the latter there are legal problems related to the subjective element of an international crime.

In armed conflict autonomous agents are often thought of as more likely than humans to operate inconsistently with the principles of distinction and proportionality and thus facilitate the commission of war crimes.[130] Under ICL, whether an act is carried out by means of a kinetic weapons system or a cyber operation is less relevant than the intent of the human actor and the consequences of the act itself. As the *Tallinn Manual* states: '[c]yber operations may amount to war crimes and thus give rise to individual criminal responsibility under international law'.[131] What is essential, however, is the relationship between the behaviour of by autonomous intelligent agents during cyber operations and the knowledge and intent (*mens rea* –

[127] Ibid para 28.

[128] See, eg, McFarland and McCormack (n 39).

[129] See, eg, 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) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2679669>.

[130] Human Rights Watch, *Heed the Call* (n 103); O'Connell (n 40); Noel E Sharkey, 'The Evitability of Autonomous Robot Warfare' (2012) 94(886) *International Review of the Red Cross* 787.

[131] *Tallinn Manual 2.0* (n 2) 391 rule 84.

culpable frame of mind) of the developers and operators involved. It then follows that debates around accountability in the context of AWS provide a good analogy and basis for discussing accountability with respect to autonomous cyber capabilities since both AWS and autonomous cyber capabilities face similar challenges in identifying intent or negligence and defining the reasonable extent of knowledge that could be expected from an operator.

War crimes can be defined as serious violations of customary or treaty-based IHL that have grave consequences for the victim and which entail individual criminal responsibility for the perpetrator.[132] It follows that not all violations of IHL are serious enough to constitute a war crime. To distinguish between unlawful acts that constitute war crimes and those that do not, it may be helpful to look at which unlawful acts national and international courts have deemed to reach the level of war crimes in the past. Another route would be to consider what conduct has been explicitly listed as war crimes in the Statutes of international criminal tribunals. In this context, the Statutes of the International Criminal Tribunal for the Former Yugoslavia (ICTY), the International Criminal Tribunal for Rwanda (ICTR) and the International Criminal Court (ICC) are useful in identifying certain unlawful acts as crimes of war. In the context of AWS, advocates of a ban have argued that an artificial system, no matter how advanced, would not be able to weigh abstract matters such as distinction and proportionality. These are among the basic principles of IHL, the violation of which could constitute a war crime.[133]

Majority of the legal complications arise when defining and establishing intent and knowledge. Article 30 of the Rome Statute states that, unless otherwise provided, a person shall be criminally responsible and liable for punishment for a crime only if the material elements are committed with intent and knowledge.[134] The Statute therefore sets the threshold of *mens rea* to intentionality. Pursuant to the Statute, a person has intent where: (a) in relation to conduct that person means to engage in the conduct; and (b) in relation to a consequence, that person means to cause that consequence or is aware that it will occur in the ordinary course of events. Knowledge, in the sense of Article 30, means awareness that a circumstance exists or a consequence will occur in the ordinary course of events.

Prior knowledge of *exactly* how an autonomous agent will operate in any given environment is often an unrealistic expectation.[135] When deploying a self-learning autonomous system capable of reacting to environments that differ vastly from laboratory conditions, a commander might be unaware whether an automated hack-back operation taking place in an armed conflict ends up targeting a civilian network. There is a high probability that not every operator is equipped with the required technical know-how and even if they are, the crucial decisions would be made in the black box.

In *Bemba* and *Lubanga*, the ICC stated that the mental state of the perpetrator who does not intend to cause the forbidden result, but foresees its occurrence as a necessary, certain or

[132] *Prosecutor v Tadić* (Decision on the Defence Motion for Interlocutary Appeal on Jurisdiction) IT-94-1-AR72 (2 October 1995) para 94.

[133] See, eg, Rome Statute of the International Criminal Court (17 July 1998, in force 1 July 2002) 2187 UNTS 90 especially art 8(2)(b)(i)–(v) and (e)(i)–(iii).

[134] Rome Statute art 30.

[135] For a discussion on how the very capability to exceed human capabilities in speed and reasoning forms the basis of why autonomous features are often preferred in trading and cyber defence, see, eg, Scharre (n 36) 230.

highly probable consequence of the achievement of his main purpose and nevertheless engages in the conduct would satisfy the test of of Article 30. The court defined the standard as 'virtual certainty'.[136] When it comes to an autonomous system, however, the operator might recognise that a consequence may occur but nonetheless choose to use the system. Whether or not this is covered by the concept of being 'aware that a consequence will occur in the ordinary course of events' is a question yet to be answered by the ICC or legal scholars.[137]

The grave breaches regime envisioned in AP I of the Geneva Convention sets out a somewhat different framework for culpability since it requires a certain consequence to arise from a breach. Grave breaches of the principle of distinction occur when the civilian population or individuals have been made the objects of an attack: (a) wilfully; (b) in violation of the relevant provisions of the Protocol; and (c) it has caused serious injury to body or health.[138] Commentary on Article 85 AP I explicates that conduct has been 'wilful' when the accused has acted consciously and with intent: i.e. with their mind on the act and its consequences and willing them – 'criminal intent' or 'malice aforethought'. This is interpreted to encompass the concepts of wrongful intent or recklessness, viz., the attitude of an agent who, without being certain of a particular result, accepts the possibility of it happening. Wilfulness has been interpreted to include direct intent, indirect intent and recklessness.[139] Prosecuting war crimes committed via an autonomous system might therefore be easier in States where criminal law closely follows the grave breaches system of the Geneva Conventions and AP I, or otherwise lowers the mental element of recklessness or *dolus eventualis*.

[136] *Prosecutor v Lubanga* (Judgment on the Appeal against Conviction) ICC-01/04-01/06 A 5 (1 December 2014) para 447; also *Prosecutor v Bemba* (Decision on Confirmation of Charges) ICC-01/05-01/08 (15 June 2009) para 362.

[137] Marta Bo, 'The Human-Weapon Relationship in the Age of Autonomous Weapons and the Attribution of Criminal Responsibility for War Crimes', We Robot Conference (University of Miami School of Law, 11–13 April 2019).

[138] AP I art 85(3).

[139] Bo (n 7).

6.3 International liability for high-risk actions

The assumed unpredictability and capacity to surpass human oversight[140] raise questions about the reasonable level of control that a State should exercise while deploying an autonomous cyber capability or outsourcing its defence or intelligence functions in a way that will result in the use of an autonomous agent. Due diligence obligations typically require States not to allow knowingly their territory to be used for acts contrary to the rights of other States, and failing to do so is deemed an internationally wrongful act.[141] States, however, face the challenge of contextualising these open notions on the novel and unexplored technological landscape of autonomous cyber capabilities. While the definitions are yet to crystallise, a situation may arise where victims have no recourse to compensation for injury. The aim of strict liability is to prevent such situations.[142] Despite the numerous initiatives and a high degree of interest, there is currently no general regime of liability for injurious consequence arising out of activities not prohibited by international law[143]. There are, however, treaty-based regimes custom-tailored for a specific field of activity, whereas State liability is established only for activities in outer space. To date, however, there is no specialised treaty addressing liability in cyberspace. The risks involved in the broad use of autonomous features may well initiate a debate over the necessity of such a treaty.

The development and use of autonomous features in cyber defence seems to be inevitable since successful attack vectors increasingly involve a machine-learning element.[144] Unlike the debate on AWS, the questions associated with autonomous cyber capabilities seldom find a solution in a rigid ban. This leads us to the question of liability for lawful actions and the preferable liability scheme for damage caused by autonomous cyber capabilities. The traditional approach of State responsibility always requires wrongful conduct of a State as the hypothetical examples above illustrate. While the accountability gap is sometimes exaggerated, the requirement for a breach of an obligation to occur persists. Major difficulties arise when trying to establish just when and how a State has erred. Not unexpectedly, this has led to much discussion about whether direct or strict liability should apply, i.e. whether States should be liable for any damage caused by certain activities under their control regardless of negligence or fault.[145] The ICRC commentaries on AP I at least contemplate the possibility of strict liability.[146]

[140] Cf Scharre (n 36) 347.

[141] See, eg, International Law Association, Study Group on Due Diligence in International Law, *First Report* (2014) 2–4.

[142] James Crawford, Alain Pellet, and Simon Olleson. The law of international responsibility. (Oxford University Press, 2010), 504-505.

[143] Ibid, 511.

[144] Rebecca Crootof, 'Autonomous Weapon Systems and the Limits of Analogy' (2018) 9 *Harvard National Security Journal* 51, 63.

[145] Robin Geiss, *The International-Law Dimension of Autonomous Weapons Systems* (Friedrich-Ebert-Stiftung, 2015) 22.

[146] Jean de Preux, 'Protocol I – Article 91' 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) 1058 para 3661.

Strict liability regimes are well-suited for high-risk activities that are unforeseeable and based on novel technologies and have been established for activities in outer space and in transboundary environmental damage.[147] In the latter realm, however, liability treaties have faced great barriers to adoption and even greater to enforcement.[148] High-risk activities have been described as those that involve novel technologies or which could bring about unforeseeable consequences.[149] Autonomous cyber capabilities seem to fit these criteria. Perhaps more than the technological revolutions of the past, AI will necessitate the re-evaluation of the concept of control over activities that are difficult to monitor and capable of operating without direct supervision.

International liability has relevance to autonomous cyber capabilities, although liability schemes might prove insufficient for the victim State since they can only at best lead to compensation. Strict liability is particularly relevant to activities in the cyber domain since it is meant for cases of potential transboundary harm which result from otherwise legal activities. As it is generally the State that provides the general framework for dangerous activities, conducts oversight and enforces the law, the State should also be prepared to bear the consequences of such activities going wrong, regardless of whether due care has been taken. Currently, however, very few States have legislation that specifically regulates the conduct of autonomous agents.[150] Many more have expressed interest in developing autonomous cyber capabilities.

The ILC Draft Articles on the 'Prevention of Transboundary Harm Arising from Hazardous Activities' limits the types of harm for which strict liability applies by prescribing that: 'The present articles apply to activities not prohibited by international law which involve a risk of causing significant transboundary harm through their physical consequences'.[151] The restriction does not work well with the notion of autonomous systems, where the very hazard lies in the fact that it is difficult to predict just what kind of consequences they might bring about – functional, physical, indiscriminate, self-replicating, limited or indeterminate – in their temporal and territorial scope. Therefore, any instrument that limits liability to a certain physical consequence would not be in line with the concerns about autonomous cyber capabilities.

We find perhaps the strictest international liability scheme from a historical situation which bears close resemblance to the current developments in AWS. In the wake of the Cold War, space technology was thought too complicated and advanced to be safe and predictable and had the potential to trigger an arms race. Furthermore, the risks arising from the application of space technologies were perceived as too complex to be mitigated by meeting the usual requirements of due diligence. Therefore, concerns were raised over a potential legal gap emerging where the risk-creating state had caused grave damages to another, yet not committed an internationally wrongful act, since it had applied all the adequate safety measures. The novelty

[147] Alexandre Kiss and Dinah Shelton, 'Strict Liability in International Environmental Law' in Tafsir Malik Ndiaye and Rüdiger Wolfrum (eds), *Law of the Sea, Environmental Law and Settlement of Disputes: Liber Amicorum Judge Thomas A. Mensah* (Nijhoff 2007).

[148] Ibid.

[149] John M Kelson, 'State Responsibility and the Abnormally Dangerous Activity' (1972) 13 *Harvard International Law Journal*197.

[150] A Atabekov and O Yastrebov, 'Legal Status of Artificial Intelligence Across Countries: Legislation on the Move' (2018) 21(4), *European Research Studies Journal* 773.

[151] Draft Articles on Prevention of Transboundary Harm from Hazardous Activities (2001) in UN Doc A/56/10 art 1.

of space technology and the uncertainty surrounding it made it difficult to substantiate the principle of due diligence. Making the breach of the latter a pre-requisite for any claim would, considering the magnitude of potential harm, put the victim State in an unfavourable position.[152]

The Outer Space Treaty of 1967[153] reinforces the principle of responsibility for all damage which the launched objects cause on Earth, in air space or in outer space. Article VII prescribes a strict liability scheme according to which States are internationally liable for damage to other States, their property or persons, caused by their space objects – which for our purpose may be considered as being equal to damage resulting from an activity in space.[154] The Liability Convention of 1972 develops this scheme further.[155] Space liability, therefore, is not dependent on any other objective (breach of an international obligation) or subjective (intent or negligence) element, only two elements matter: damage has occurred and the damage was caused by a space object.

Robin Geiss argues that taking space liability as a blueprint for a strict liability scheme applicable to the use of autonomous weapons systems might prove feasible since the technological revolutions share a number of features.[156] First, they concern technologies that once implemented may be difficult to control and the effect might be hard to trace, and the long-term effect is difficult to foresee. What distinguishes autonomous systems and makes the prospect of a similar strict liability scheme unattractive to States is that they are easier to copy, distribute and launch. A State is likely to end up being liable for acts that it has no power over. Second, the material scope of the treaty is limited to space objects – a term that, while not an entirely static target, is less difficult to capture than ‘autonomy’. Therefore, while a liability scheme might form a part of an attempt to regulate the use of autonomous features in both kinetic and cyber warfare, analogies only go so far, at least as long as there is no consensus over what level of skill, caution and knowledge that operators, commanders, regulators and engineers should be expected to possess.

[152] René Lefeber, *Transboundary Environmental Interference and the Origin of State Liability*. Vol. 24. (Martinus Nijhoff Publishers, 1996), 145

[153] Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies, (27 January 2967, in force 10 October 1967)18 UST 2410, 610 UNTS 205. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (8 June 1977, in force 7 December 1978) 1125 UNTS 3 (AP I) art 48

[154] Ibid art VII.

[155] Convention on International Liability for Damage Caused by Space Objects (29 March 1972, in force 1 September 1972) 961 UNTS 187, art II.

[156] Geiss (n 145).

Conclusions

This study examined the interplay of autonomy and cyber capabilities through the lens of international law, focusing on sovereignty, law on the use of force, international humanitarian law, international criminal law, and the law of state responsibility and liability. The intersection of cyber and autonomy can easily be seen as creating an endless array of complex legal issues. In our view, increased autonomy does not always have a significant impact on the application of the law in the cyber context. Many rules of international law continue to apply irrespective of any autonomous functionality of a kinetic or cyber capability. This being said, autonomy does add complexity to some of the legal and technical problems. In particular, certain rules of international law, the breach of which requires a particular mental element, may be difficult to apply with respect to conduct involving the autonomous features of a technological system.

Autonomous cyber capabilities can be deployed in ways that breach sovereignty or amount to unlawful intervention. Incorporating autonomous functionality into cyber capabilities has no bearing on whether or not a particular cyber operation breaches the principle of sovereignty. A State using a cyber capability retains responsibility for breaching the sovereignty of other States even where such breaches arose due to the autonomous functionality of that capability. At the same time, whether or not a cyber operation amounts to an intervention can depend on the autonomous functionality of the capability. Where there is an interference in the *domaine réservé* of a State resulting from the autonomous functionality of a State, but without the intent of that State to coerce another, an intervention has not taken place.

The use of autonomous systems, including cyber capabilities, entails challenges for compliance with rules that require the assessment of facts in light of abstract concepts, or the balancing of considerations that cannot be easily quantified. In particular, determinations as to whether a use of force amounts to an armed attack, whether self-defence measures are necessary and proportionate, and whether collateral damage resulting from an attack is excessive, fall within this category.

Technology may well assist in some parts of the assessment, such as identifying particular types of targets or assessing the extent of (collateral) damage caused by attacks. Identifying lawful targets by technical means might be less difficult in the cyber domain, since less depends on sensors and environmental conditions. On the other hand, all systems, irrespective of their level of autonomy, could be led astray by various obfuscation techniques such as spoofing.

At the same time, the overall evaluation of the facts as required by law might be beyond the capabilities of current or immediately foreseeable technology. Yet from a legal and operational perspective this problem is not insurmountable. Scenarios where States and commanders would want to delegate to an AWS or autonomous cyber capability the full discretion available to a combatant as matter of law, remain scarce. Commanders will likely only entrust to an autonomous system a role that they determine the system to be capable of carrying out consistent with their intent, the requirements of law and national policy. On an operational level this would imply that, just as the actions of combatants are in addition to IHL restricted by rules of engagement, the functionalities of an autonomous cyber capability would be limited to finding and engaging legitimate targets that are easily recognisable, and where collateral damage is unlikely to be an issue.

Another area, where autonomous systems are thought to underperform and create an accountability gap is the obligation to take feasible precautionary measures. Precautionary measures are not functions that can be delegated to an AWS or autonomous cyber capability, but they must be taken by people who plan, authorise or execute attacks. The deployment of the capability would therefore, from a legal perspective, be seen as a decision to launch an attack. Taking precautionary measures may, however, become complicated in circumstances where a cyber capability is intended to operate for an extended period and in complex environments, and thus be confronted with a range of different operational circumstances. Therefore, the higher the degree of autonomy and the graver the potential consequences, the greater the risk to err in the taking of precautionary measures.

The popular idea that the use of autonomous capabilities would create an unbridgeable accountability gap overstates the problem. Existing regimes of State and individual responsibility remain relevant for the use of autonomous capabilities. A number of uncertainties do emerge, however.

With regard to State responsibility, autonomous capabilities will likely create new practical problems of attribution, as those capabilities are not continuously ‘tethered’ to a human operator. From a legal perspective, this is more an evidentiary than a conceptual problem. Difficulties arise, however, where the breach of a rule of international law (such as the prohibition of intervention, or of genocide) requires a particular mental element. In such instances, if the harm that the law seeks to avoid is brought about inadvertently because of the unpredictable performance of an autonomous capability, direct State responsibility does not obtain. It may therefore become necessary to rely on a broad principle due diligence. Here, the missing piece of the puzzle is a clear understanding of the due care standards that a State should abide by when developing, deploying or approving autonomous cyber capabilities. Achieving an international consensus on States’ obligations in developing, acquiring or overseeing autonomous cyber capabilities is a distinctly complicated task that combines the unsolved questions of the law that applies to cyber operations and the governance of AWS.

Another way of addressing the accountability of States would be by means of strict liability, since this is less concerned with the subjective element of State action. Strict liability regimes are generally based on treaties. Creating such a regime for autonomous cyber capabilities, even if it might be desirable legally, does not seem realistic politically, since the use and development of such technologies is much harder to control than, for example, the activities outer space, which are prominently subject to a strict liability regime.

When it comes to individual accountability, intent plays a critical role. Therefore, holding individuals accountable under ICL for consequences resulting from the use of autonomous capabilities if fraught with difficulty. If an autonomous capability causes harm to protected persons in an armed conflict because of the negligence of programmers, it is uncertain whether anyone could be held responsible for a war crime, the commission of which requires intent (or, at the very least, recklessness or *dolus eventualis*). This, in turn, highlights the importance of national law in disciplining individuals for wrongdoings that fall below the threshold of international crimes.

This paper sought to shed light on the most debated and relevant international law aspects of autonomous cyber capabilities. Rather than giving definite answers it opens doors more detailed insights into a subject that has in comparison to, for instance, AWS gained relatively little political and scholarly attention. This is despite the view that the most dramatic advancements in terms of autonomous military capabilities have taken place in the cyber context. While on a closer look many of the disputes could be in fact reduced to practical, procedural or technical matters, some vital legal questions remain, among them (not exhaustively): autonomous cyber capabilities and the element of intent in prohibited intervention, an autonomous system's capability to assess the severity of an incoming attack, autonomous cyber capability and the duty to take feasible precautionary measures, autonomous cyber capabilities and *mens rea* and international liability schemes for damages caused by the use of an autonomous cyber capability.

References

Treaties

International Convention for the Unification of Certain Rules relating to Damage caused by Aircraft to Third Parties on the Surface (29 May 1933, in force 13 February 1942) 192 LNTS 191.

Convention on Damage Caused by Foreign Aircraft to Third Parties on the Surface (7 October 1952, in force 4 February 1958) 310 UNTS 182.

Vienna Convention on Civil Liability for Nuclear Damage (21 May 1963, in force 12 November 1977) 1063 UNTS 265.

Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies (27 January 1967, in force 10 October 1967) 610 UNTS 205.

Convention on International Liability for Damage Caused by Space Objects (29 March 1972, in force on 1 September 1972) 961 UNTS 187.

International Convention on Civil Liability for Oil Pollution Damage (29 November 1969, in force 19 June 1975) 973 UNTS 3.

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

Convention on Civil Liability for Damage Resulting from Activities Dangerous to the Environment (21 June 1993) ETS no 150.

Rome Statute of the International Criminal Court (17 July 1998, in force 1 July 2002) 2187 UNTS 90.

Cases

Military and Paramilitary Activities in and against Nicaragua (Nicaragua v US) [1986] ICJ Rep 14.

Prosecutor v Bemba (Decision on Confirmation of Charges) ICC-01/05-01/08 (15 June 2009).

Prosecutor v Lubanga (Judgment on the Appeal against Conviction) ICC-01/04-01/06 A 5 (1 December 2014).

Prosecutor v Tadić (Decision on the Defence Motion for Interlocutary Appeal on Jurisdiction) IT-94-1-AR72 (2 October 1995).

Documents

Alston, Philip, ‘Interim Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions’ (23 August 2010) UN Doc A/65/32.

Estonia and Finland, ‘Categorizing Lethal Autonomous Weapons Systems: A Technical and Legal Perspective to Understanding LAWS’ (24 August 2018) UN Doc CCW/GGE2/2018/WP2.

Heyns, Christof, ‘Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions’ (9 April 2013) UN Doc A/HRC/23/47.

International Law Commission, ‘Articles on Responsibility of States for Internationally Wrongful Acts’ (2001) UN Doc A/56/10.

International Law Commission, ‘Draft Articles on Prevention of Transboundary Harm from Hazardous Activities’ (2001) in UN Doc A/56/10.

‘Report of the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)’ (10 June 2014) UN Doc CCW/MSP/2014/3.

‘Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)’ (1 June 2015) UN Doc CCW/MSP/2015/3

‘Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)’ (10 June 2016) UN Doc CCW/CONF.V/2.

‘Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS)’ (22 December 2017) UN Doc CCW/GGE.1/2017/3.

‘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.

‘Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security’ (30 July 2010) UN Doc A/65/201.

‘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.

‘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.

United Kingdom Ministry of Defence, ‘Joint Doctrine Note (JDN) 2/11: The UK Approach to Unmanned Aircraft Systems’ (30 March 2011).

United States, ‘Autonomy in Weapon Systems’ (10 November 2017) UN Doc CCW/GGE.1/2017/WP.6.

United States Department of Defence, ‘DoD Directive No. 3000.09: Autonomy in Weapon Systems’ (21 November 2012).

Other Sources

Atabekov, A, and Yastrebov, O, ‘Legal Status of Artificial Intelligence Across Countries: Legislation on the Move’ (2018) 21(4) *European Research Studies Journal* 773.

Avgerinos, Thanassis, and others, ‘The Mayhem Cyber Reasoning System’ (2018) 16 *IEEE Security & Privacy* 52.

Backstrom, Alan, and Henderson, Ian, ‘New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews’ (2012) 94 *International Review of the Red Cross* 483.

Belson, David, ‘Finding Yourself: The Challenges of Accurate IP Geolocation’ (*Oracle Dyn Blog*, 29 January 2018) <dyn.com/blog/finding-yourself-the-challenges-of-accurate-ip-geolocation>.

Brey, Philip, and Søraker, Johnny Hartz, ‘Philosophy of Computing and Information Technology’ in Anthonie Meijers (ed), *Philosophy of Technology and Engineering Sciences* (Elsevier Science & Technology 2009).

Bo, Marta, ‘The Human-Weapon Relationship in the Age of Autonomous Weapons and the Attribution of Criminal Responsibility for War Crimes’, We Robot Conference (University of Miami School of Law, 11–13 April 2019).

Buchan, Russell, *Cyber Espionage and International Law* (Hart 2018).

Christman, John, ‘Autonomy in Moral and Political Philosophy’ in Edward N Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Metaphysics Research Lab, Stanford University 2018) <plato.stanford.edu/archives/spr2018/entries/autonomy-moral>.

Corn, Gary P, and Taylor, Robert, ‘Sovereignty in the Age of Cyber’ (2017) 111 *AJIL Unbound* 207.

Correll, Nikolaus, ‘Introduction to Autonomous Robots’ (v17, Magellan Scientific 2016) 15 <open.umn.edu/opentextbooks/BookDetail.aspx?bookId=316>.

Crootof, Rebecca, ‘Autonomous Weapon Systems and the Limits of Analogy’ (2018) 9 *Harvard National Security Journal* 51.

de Preux, Jean, ‘Protocol I – Article 91’ 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).

Denning, Dorothy E, ‘Framework and Principles for Active Cyber Defence’ (2014) 40 *Computers & Security* 108.

Dewar, Robert S, ‘The ‘Triptych of Cyber Security’: A Classification of Active Cyber Defence’, *2014 6th International Conference on Cyber Conflict (CyCon 2014)* (IEEE 2014).

Falliere, Nicolas, Murchu, Liam O, and Chien, Eric, ‘W32.Stuxnet Dossier’ (version 1.4, Symantec February 2011) <www.wired.com/images_blogs/threatlevel/2011/02/ Symantec-Stuxnet-Update-Feb-2011.pdf>.

Franklin, Stan, and Graesser, Art, ‘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).

Geiss, Robin, ‘The International-Law Dimension of Autonomous Weapons Systems’ (Friedrich-Ebert-Stiftung, October 2015).

Haselager, Willem FG, ‘Robotics, Philosophy and the Problems of Autonomy’ (2005) 13 *Pragmatics & Cognition* 515.

Healey, Jason, ‘Stuxnet and the Dawn of Algorithmic Warfare’ *Huffington Post* (16 April 2013) <www.huffingtonpost.com/jason-healey/stuxnet-cyberwarfare_b_3091274.html>.

Henderson, Ian S, Keane, Patrick, and Liddy, Josh, ‘Remote and Autonomous Warfare Systems: Precautions in Attack and Individual Accountability’ in Jens Ohlin (ed), Research Handbook on Remote Warfare (Edward Elgar 2017).

Hern, Alex, ‘WannaCry, Petya, NotPetya: How Ransomware Hit the Big Time in 2017’ The Guardian (30 December 2017) <www.theguardian.com/technology/2017/dec/30/wannacry-petya-notpetya-ransomware>.

Human Rights Watch, *Losing Humanity: The Case against Killer Robots* (2012).

Human Rights Watch, *Heed the Call: A Moral and Legal Imperative to Ban Killer Robots* (2018).

Kelson, John M, ‘State Responsibility and the Abnormally Dangerous Activity’ (1972) 13 *Harvard International Law Journal* 197.

Kiss, Alex, and Shelton, Dinah L, ‘Strict Liability in International Environmental Law’ in Tafsir Malik Ndiaye and Rüdiger Wolfrum (eds), *Law of the Sea, Environmental Law and Settlement of Disputes: Liber Amicorum Judge Thomas A. Mensah* (Nijhoff 2007).

Kodar, Erki, ‘Applying the Law of Armed Conflict to Cyber Attacks: From the Martens Clause to Additional Protocol I’ in Rain Liivoja and Andres Saumets (eds), *The Law of Armed Conflict: Historical and Contemporary Perspectives* (Tartu University Press 2012).

Langner, Ralph, ‘Stuxnet: Dissecting a Cyberwarfare Weapon’ (2011) 9 *IEEE Security & Privacy Magazine* 49.

Maes, Pattie, ‘Artificial Life Meets Entertainment: Lifelike Autonomous Agents’ (1995) 38 *Communications of the ACM* 108.

Margulies, Peter, ‘Sovereignty and Cyber Attacks: Technology’s Challenge to the Law of State Responsibility’ (2013) 14 *Melbourne Journal of International Law* 496.

Maurer, Tim, *Cyber Mercenaries: The State, Hackers, and Power* (Cambridge University Press 2018).

McFarland, Tim, and McCormack, Tim, 'Mind the Gap: Can Developers of Autonomous Weapons Systems Be Liable for War Crimes?' (2014) 90 *International Law Studies* 361.

McFarland, Tim, 'Autonomous Weapons and Human Control' (*Humanitarian Law & Policy*, 18 July 2018) <blogs.icrc.org/law-and-policy/2018/07/18/autonomous-weapons-and-human-control>.

Murphy, Robin R, *Introduction to AI Robotics* (MIT Press 2000).

O'Connell, Mary E, 'Banning Autonomous Killing' in Matthew Evangelista and Henry Shue (eds), *The American Way of Bombing: How Legal and Ethical Norms Change* (Cornell University Press 2013).

O'Connor, Jennifer M, 'Memorandum: International Law Framework for Employing Cyber Capabilities in Military Operations' (19 January 2017).

Saariluoma, Pertti, 'Four Challenges in Structuring Human-Autonomous Systems Interaction Design Processes' in Andrew P Williams and Paul D Scharre (eds), *Autonomous Systems: Issues for Defence Policymakers* (Headquarters Supreme Allied Commander Transformation 2016).

Sanger, David E, 'Obama Ordered Wave of Cyberattacks against Iran' *The New York Times* (1 June 2012) <www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html>.

Sassòli, Marco, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified' (2014) 90 *International Law Studies* 308.

Scharre, Paul, *Army of None: Autonomous Weapons and the Future of War* (W W Norton & Company 2018).

Schmitt, Michael N, and Vihul, Liis, 'International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms' (*Just Security*, 30 June 2017) <www.justsecurity.org/42768/ international-cyber-law-politicized-gges-failure-advance-cyber-norms>.

Schmitt, Michael N, and Vihul, Liis, 'Respect for Sovereignty in Cyberspace' (2017) 95 *Texas Law Review* 1639.

Schmitt, Michael N, and Thurnher, Jeffrey S, ''Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict' (2013) 4 *Harvard National Security Journal* 231.

Schmitt, Michael N, 'In Defence of Sovereignty in Cyberspace' (*Just Security*, 8 May 2018) <www.justsecurity.org/55876/defence-sovereignty-cyberspace>.

Sharkey, Noel E, 'The Evitability of Autonomous Robot Warfare' (2012) 94(886) *International Review of the Red Cross* 787.

Singer, Peter W, 'Stuxnet and Its Hidden Lessons on the Ethics of Cyberweapons' (2015) 47 *Case Western Reserve Journal of International Law* 79.

Smithers, Tim, 'Autonomy in Robots and Other Agents' (1997) 34 *Brain & Cognition* 88.

Stuard, Jarrod H, and McGhee, James, 'Is Skynet the Answer? Rules for Autonomous Cyber Response Capabilities' in Misty Blowers (ed), *Evolution of Cyber Technologies and Operations to 2035* (Springer 2015).

Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Cambridge University Press 2017).

Tallinn Manual on the International Law Applicable to Cyber Warfare (Cambridge University Press 2013)

Taylor, James S, 'Autonomy' (*Oxford Bibliographies*, 19 May 2017) available at: <www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0167.xml>.

Teubner, Gunther, 'Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law' (2006) 33 *Journal of Law & Society* 497.

Thurnher, Jeffrey S, 'Feasible Precautions in Attack and Autonomous Weapons' in Wolff Heintschel von Heinegg, Robert Frau and Tassilo Singer (eds), *Dehumanization of Warfare* (Springer 2018).

Ticehurst, Rupert, 'The Martens Clause and the Laws of Armed Conflict' (1997) 37 *International Review of the Red Cross* 125.

Watts, Sean, and Richard, Theodore, ‘Baseline Territorial Sovereignty and Cyberspace’ (2018) 22 *Lewis & Clark Law Review* 803.

Willick, Marshal S, ‘Constitutional Law and Artificial Intelligence: The Potential Legal Recognition of Computers as “Persons”’, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (1985).

Wright, Rt Hon Jeremy, MP, ‘Cyber and International Law in the 21st Century’ (Speech, Chatham House, London, 23 May 2018) <www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>.